



**POLITECNICO**  
MILANO 1863

## Personas Creation

**PyLab #03 (Oct 25<sup>th</sup> 2024)**

**Emanuele Tauro: [emanuele.tauro@polimi.it](mailto:emanuele.tauro@polimi.it)**

# Schedule

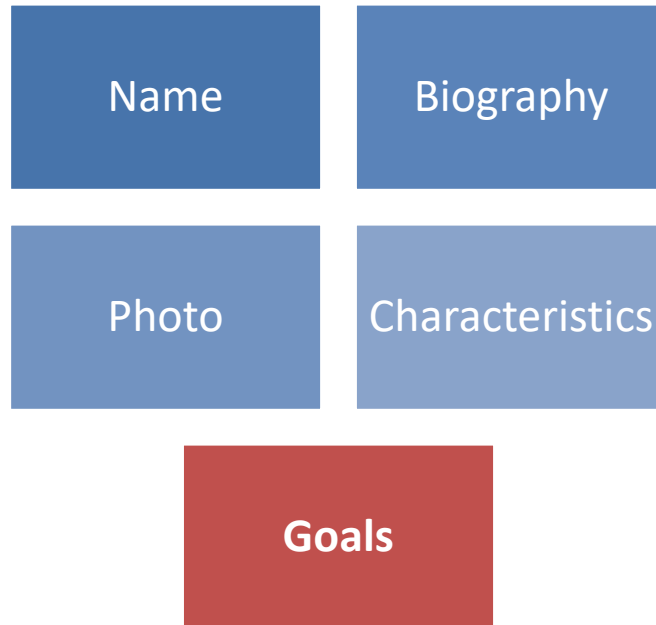
- **Personas Creation Workflow**
  - Data preparation
  - Preprocessing
  - Clustering
  - **How many Personas?**
  - Statistical Analysis
  - Personification



# Personas

*“Personas are not real people, but they represent them throughout the design process. Personas are defined by their **goals**.”*

- Alan Cooper



The image shows a detailed 'User Persona Type' template. It includes the following sections:

- User Persona Type** (Header)
- Goals** (List of goals to be completed, reached, or felt)
- Frustrations** (List of challenges, obstacles, and problems)
- Bio** (Short paragraph describing the user journey)
- Personality** (Introvert/Extrovert, Thinking/Feeling, Sensing/Intuition, Judging/Perceiving)
- Motivation** (Incentive, Fear, Growth, Power, Social)
- Brands & Influencers** (List of brands and influencers)
- Preferred Channels** (Traditional Ads, Online & Social Media, Referral, Guerrilla Efforts & PR)

Additional fields include: Name, Biography, Photo, Characteristics, Age, Work, Family, Location, Character, and a section for 'A quotation that captures this user's personality.'

# Personas Creation Workflow

- Goal Definition
- Survey creation
- Survey dissemination

**Data  
Collection**

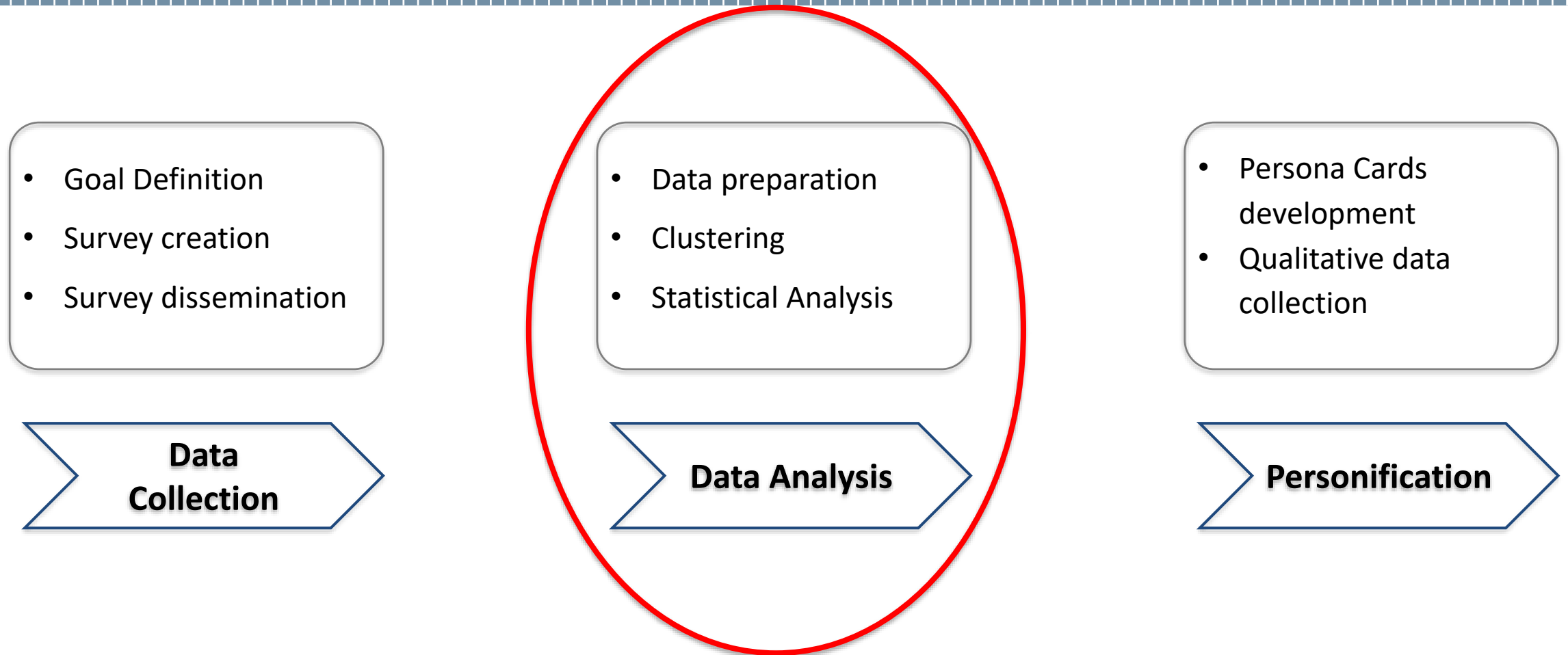
- Data preparation
- Clustering
- Statistical Analysis

**Data Analysis**

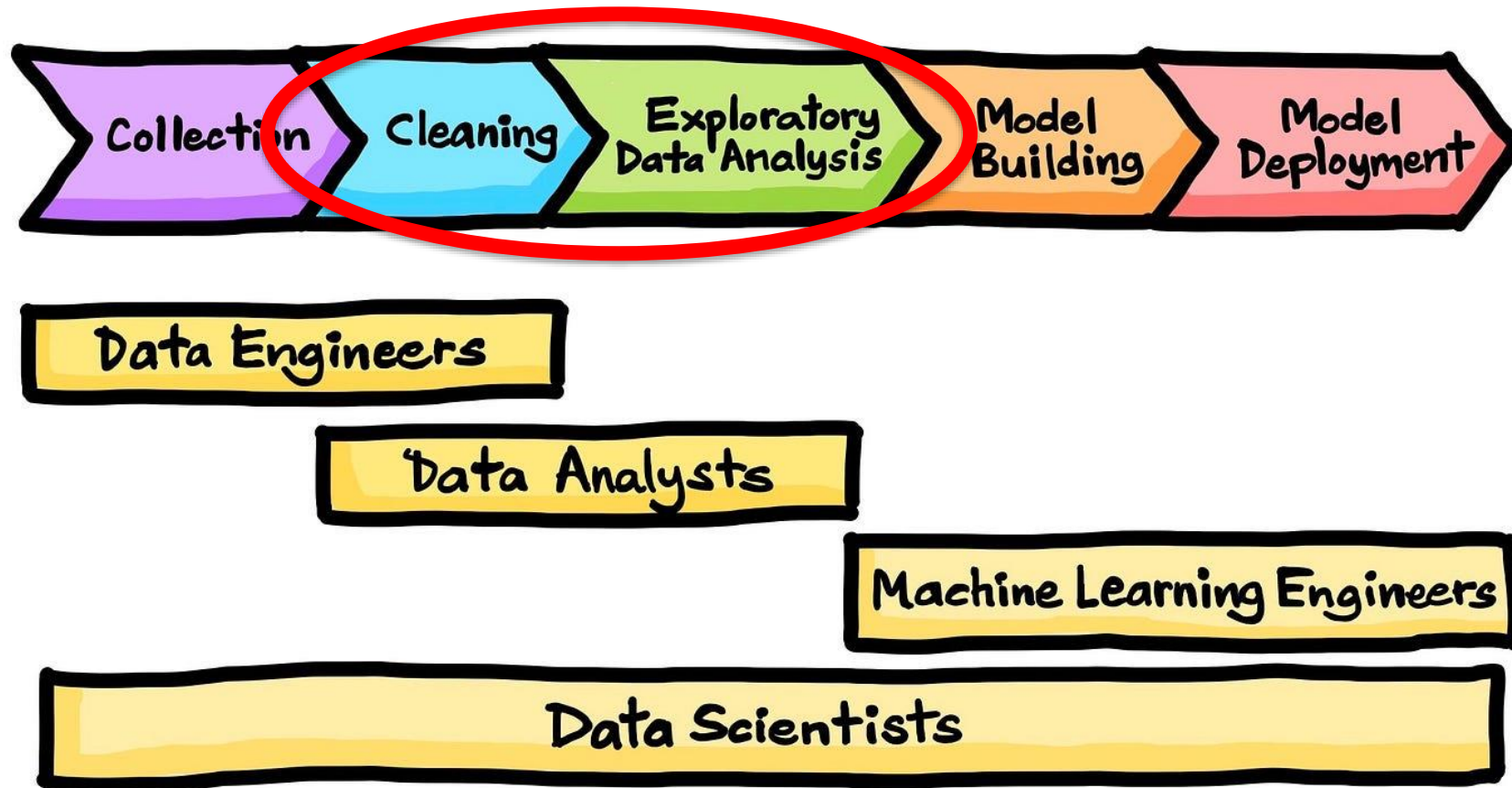
- Persona Cards development
- Qualitative data collection

**Personification**

# Personas Creation Workflow

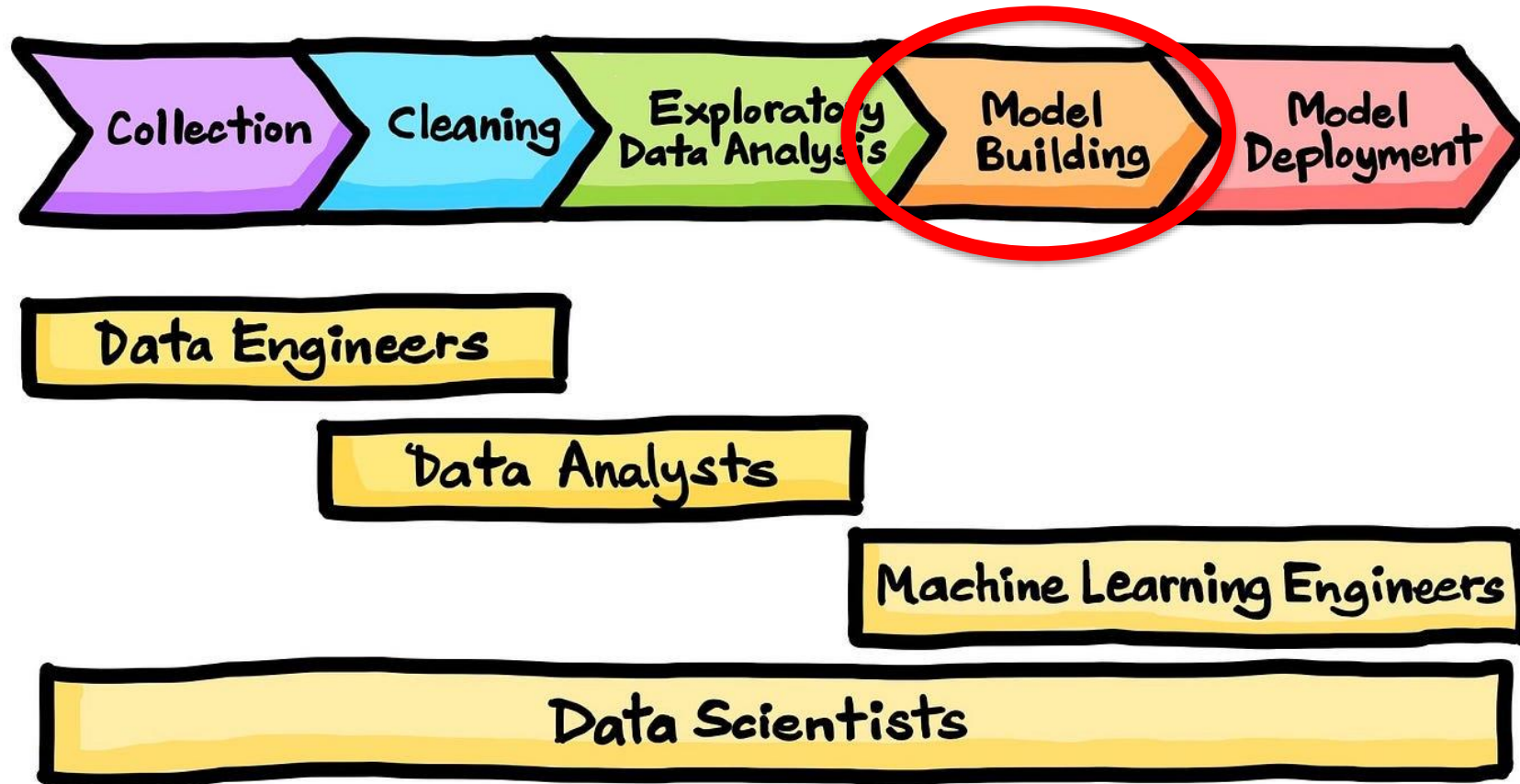


# Data Analysis: Data preparation



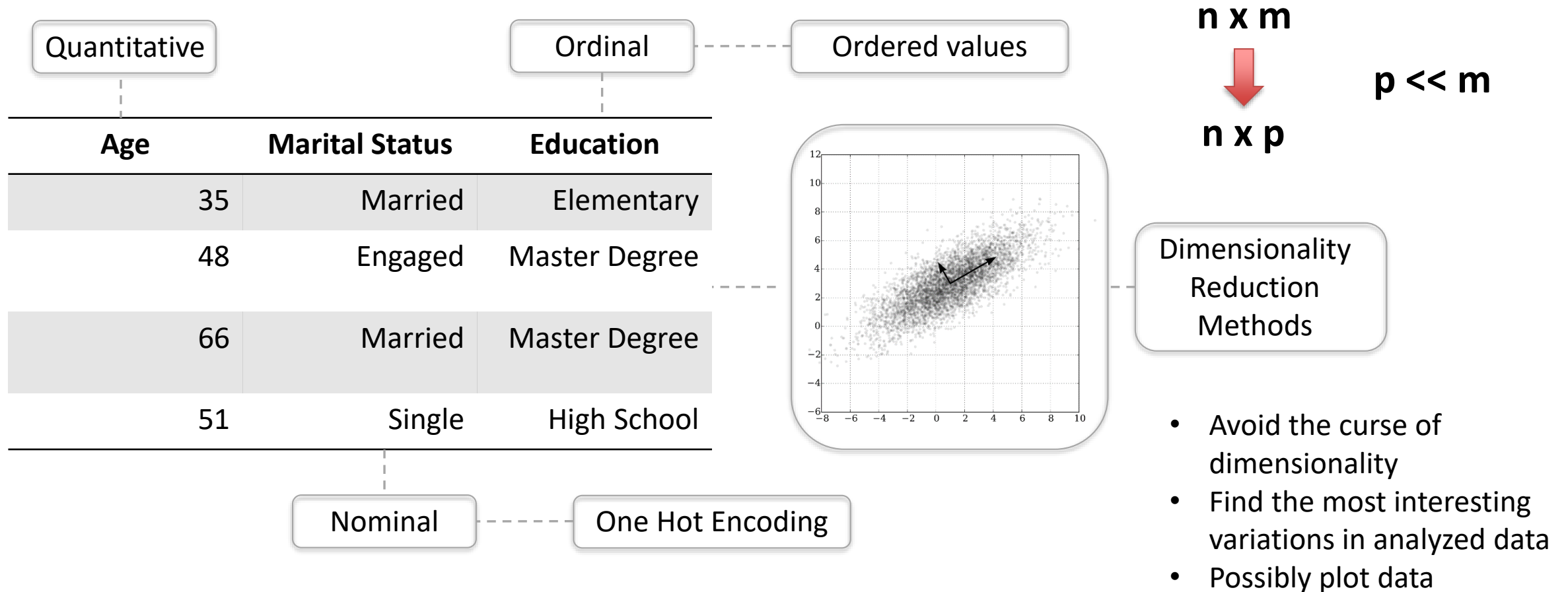
<https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b>

# Data Analysis: Data preparation



<https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b>

# Data Analysis: Data Preparation





# Data Analysis: Data Preparation

- **Principal Component Analysis (PCA):** most commonly used data reduction method. Works on numerical data and binary data only. Finds the principal components (linear combinations of variables) that explain most of the variance within the dataset. Implemented in **sklearn**.
- **Multiple Correspondence Analysis (MCA):** identical to PCA, but only for categorical variables. Implemented in **mca**.
- **Principal Component Analysis of Mixed Data (PCAMIX):** a combination of PCA and MCA. It performs PCA on quantitative data and MCA on categorical data. Can assign weights to each variable. Implementation is exclusively available in R.
- **Factor Analysis of Mixed Data (FAMD):** identical to PCAMIX. Has an implementation in python with the **prince** library. In R, it allows to divide the columns of the dataset into groups, differentiating it from PCAMIX. No weight is allowed.
- **Distance metrics:** when the data presents fewer records than features, it is possible to reduce its dimensionality by calculating a distance matrix. The distance metric used is of fundamental importance. For mixed datasets, with numerical and categorical data, one of the most used distance metrics is **Gower (GW)**. Implemented in **gower**.

# Data Analysis: Data Preparation

*How many components do we have to use when performing PCA, PCAMIX, FAMD or MCA on our dataset?  
It depends on the dataset you are analyzing.*

Each principal component obtained through these methods explains a certain **percentage** of variance within the dataset. Most times the focus is not on the number of principal components, but on the amount of variance explained by them. Reduced datasets should explain a percentage of variance from **70% to 90%** of the original dataset.

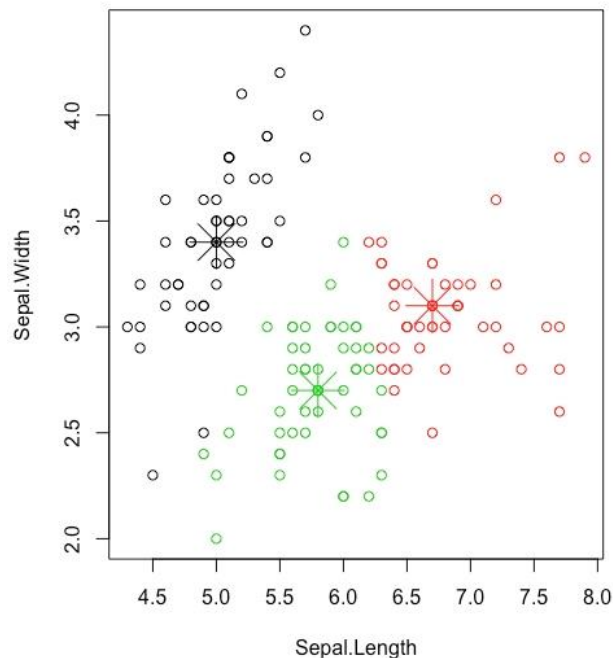
There is, however, no golden standard available. It is a **Trial and Error process**.

In other situations, you might be prompted to favour the number of components with respect to the percentage of variance. This might happen if the number of components that explain around 70% of the dataset variance is 4 or 5. With 3 components it is possible to plot data, gaining more visual insights, but at the same time information, in the form of variance explained, is lost.

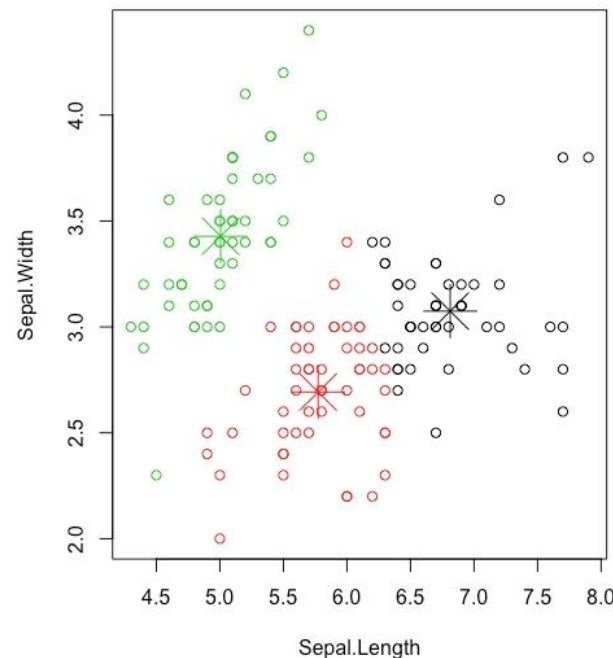
# Data Analysis: Clustering

## Partition Clustering

K-medoids Clustering



K-means Clustering



### K-Medoids (Partitioning Around Medoids)

#### Build

Assign each cluster a potential medoid.

#### Swap

Test each point as a medoid.

Assign each point to its closest medoid.



- The number of clusters  $k$  must be known a priori!
- Prone to find local minimum.
- Randomness counts.

# Data Analysis: Clustering

## Hierarchical Clustering



## Agglomerative Clustering

### Calculate

Measure the distance between clusters.

### Merge

Join the two closest clusters into one.



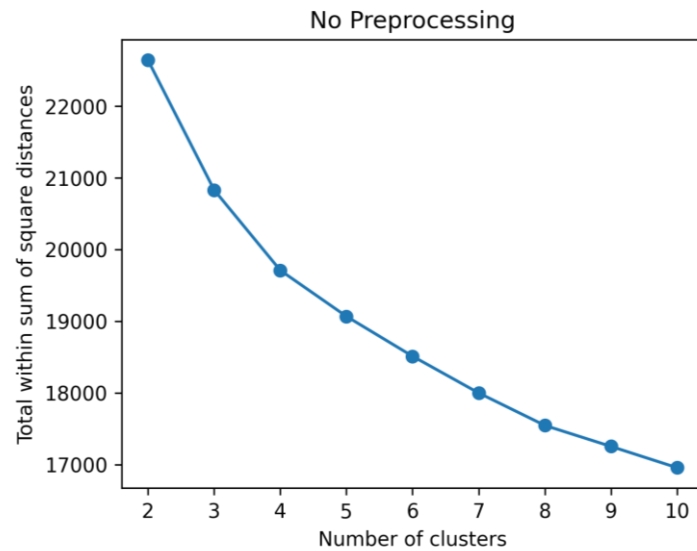
- The **linkage** (method to calculate distance) must be decided a priori.
- Does not work well on mixed data types.
- Must decide where to cut the dendrogram.



# Defining the optimal number of clusters

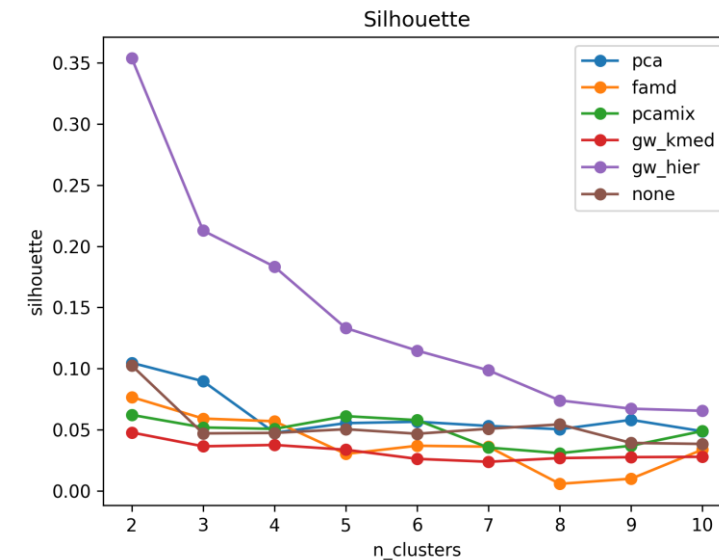
## Graphical Method

Calculating the total within sum of square distances (inertia) for a varying number of clusters, it is possible to graphically identify the **elbow** in the figure. The number of clusters where the elbow is found is the optimal number of clusters to be used.



## Analytical Method

The **average silhouette** is a measure of how similar an object is to its cluster, compared to other clusters in the same partition. The higher the value, the better. However, this information has to be supported by other information (number of samples in each cluster, statistical difference between clusters).



# Data Analysis: Clustering

Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	FALSO
Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	FALSO
Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	FALSO
VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	FALSO
Charmander	Fire		309	39	52	43	60	50	65	1	FALSO
Charmeleon	Fire		405	58	64	58	80	65	80	1	FALSO
Charizard	Fire	Flying	534	78	84	78	109	85	100	1	FALSO
CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	1	FALSO
CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	1	FALSO
Squirtle	Water		314	44	48	65	50	64	43	1	FALSO
Wartortle	Water		405	59	63	80	65	80	58	1	FALSO
Blastoise	Water		530	79	83	100	85	105	78	1	FALSO
BlastoiseMega Blastoise	Water		630	79	103	120	135	115	78	1	FALSO
Caterpie	Bug		195	45	30	35	20	20	45	1	FALSO
Metapod	Bug		205	50	20	55	25	25	30	1	FALSO
Butterfree	Bug	Flying	395	60	45	50	90	80	70	1	FALSO
Weedle	Bug	Poison	195	40	35	30	20	20	50	1	FALSO
Kakuna	Bug	Poison	205	45	25	50	25	25	35	1	FALSO
Beedrill	Bug	Poison	395	65	90	40	45	80	75	1	FALSO
BeedrillMega Beedrill	Bug	Poison	495	65	150	40	15	80	145	1	FALSO
Pidgey	Normal	Flying	251	40	45	40	35	35	56	1	FALSO
Pidgeotto	Normal	Flying	349	63	60	55	50	50	71	1	FALSO
Pidgeot	Normal	Flying	479	83	80	75	70	70	101	1	FALSO
PidgeotMega Pidgeot	Normal	Flying	579	83	80	80	135	80	121	1	FALSO
Rattata	Normal		253	30	56	35	25	35	72	1	FALSO

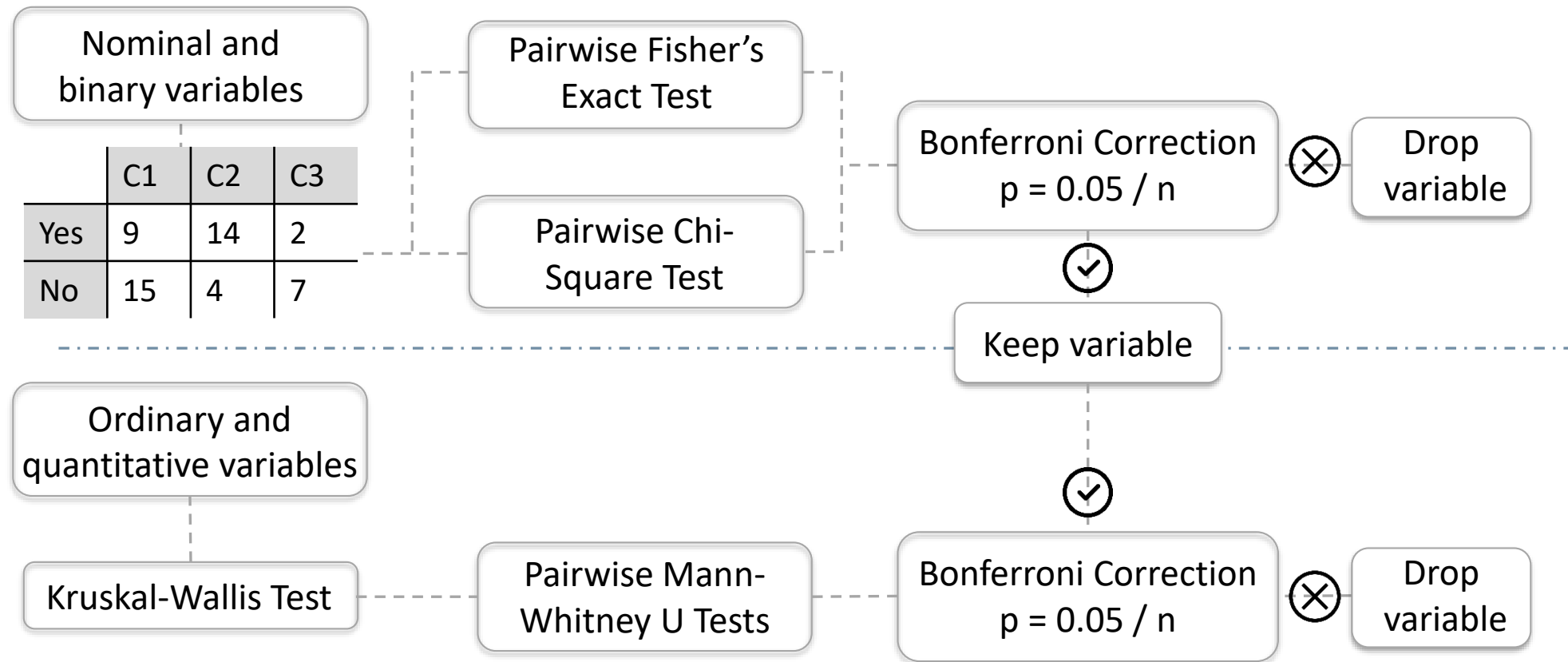
LABEL
1
0
1
1
0
0
1
1
0
1
1
1
1
0
0
1
0
1
0
1
1
0
1
1

# Data Analysis: Statistical Analysis

Statistical analysis is performed separately for each attribute in the dataset.

Depending on the type of the data, a different technique is applied.

In this case, the data in example is supposed to have a non-normal distribution.

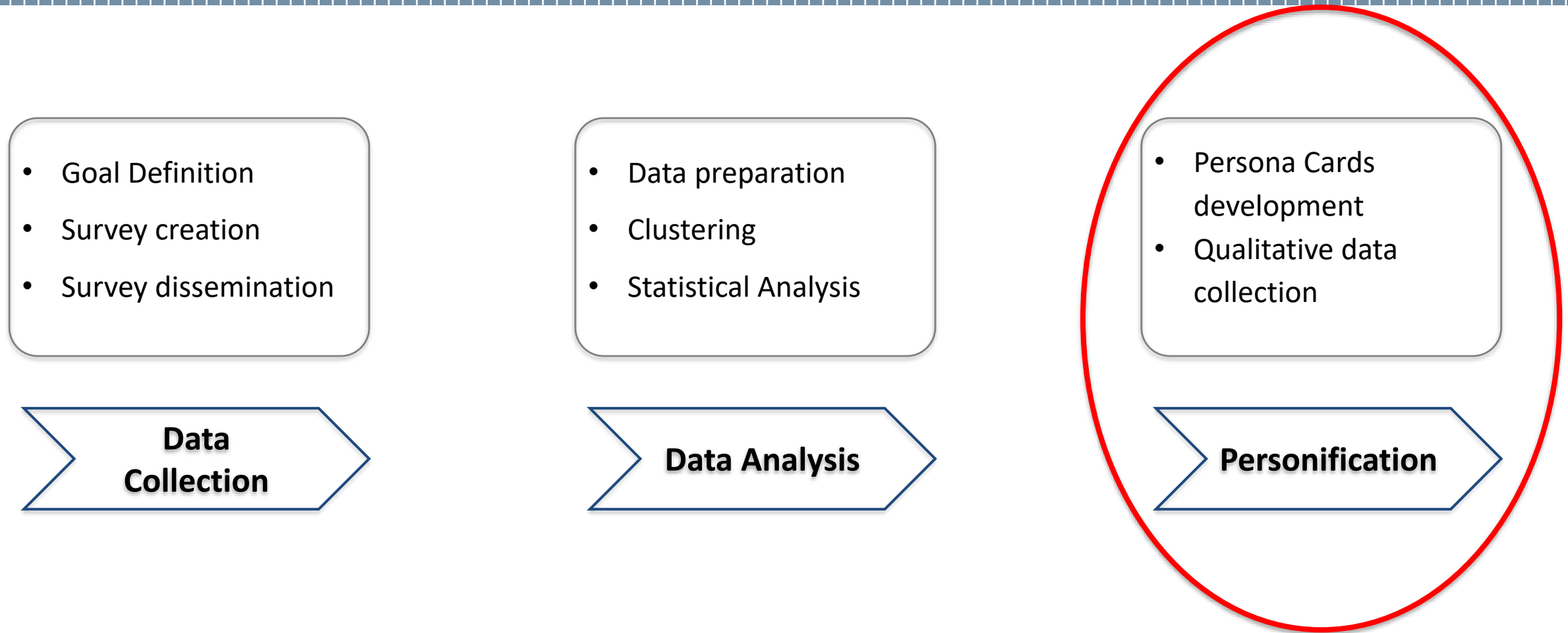


# Data Analysis: Statistical Analysis

	Cluster 1 (n = 67)	Cluster 2 (n = 38)	Cluster 3 (n = 70)	P value
Children	yes, two (42%)	no (76%) *	yes, two (50%) #	< 0.001
Marital status	married (58%)	engaged (39%) *	married (50%) #	< 0.001
Lives with	spouse + children (31%)	alone (47%) *	spouse (56%) #	< 0.001
Protective measures at home	PPE (52%)	no need (55%) *	PPE (27%) #	< 0.001
COVID-19 fear for family	83.5 (70; 95)	0 (0; 0) *	55 (46; 71.5)* #	< 0.001
COVID-19 fear for self	80 (70.38; 90.5)	64 (50; 76) *	52.5 (40; 70.5)*	< 0.001
COVID-19 cases in ward	yes (76%)	yes (76%)	yes (60%) * #	0.015
Workload impact	77 (70; 82)	69 (58; 77) *	51 (43; 58) * #	< 0.001
Stress	80 (65; 95)	70 (60; 85)	60 (52.5; 75) * #	< 0.001
MBI	16 (12; 20)	12 (8; 18) *	8 (5; 13) * #	< 0.001
IES	38 (27; 50)	25 (15;44) *	20 (12; 31) *	< 0.001
PHQ-4	5 (4; 9)	4 (2; 6) *	3 (2; 5) *	< 0.001



# Personas Creation Workflow



# Personification: Persona cards development

	Cluster 1 (n = 67)	Cluster 2 (n = 38)	Cluster 3 (n = 70)	P value
Children	yes, two (42%)	no (76%) *	yes, two (50%) #	< 0.001
Marital status	married (58%)	engaged (39%) *	married (50%) #	< 0.001
Lives with	spouse + children (31%)	alone (47%) *	spouse (56%) #	< 0.001
Protective measures at home	PPE (52%)	no need (55%) *	PPE (27%) #	< 0.001
COVID-19 fear for family	83.5 (70; 95)	0 (0; 0) *	55 (46; 71.5)* #	< 0.001
COVID-19 fear for self	80 (70.38; 90.5)	64 (50; 76) *	52.5 (40; 70.5)*	< 0.001
COVID-19 cases in ward	yes (76%)	yes (76%)	yes (60%) * #	0.015
Workload impact	77 (70; 82)	69 (58; 77) *	51 (43; 58) * #	< 0.001
Stress	80 (65; 95)	70 (60; 85)	60 (52.5; 75) * #	< 0.001
MBI	16 (12; 20)	12 (8; 18) *	8 (5; 13) * #	< 0.001
IES	38 (27; 50)	25 (15;44) *	20 (12; 31) *	< 0.001
PHQ-4	5 (4; 9)	4 (2; 6) *	3 (2; 5) *	< 0.001



Giovanni

**BIOGRAPHY**  

Giovanni/Anita lives alone in his/her apartment, thus having no need to use protective measures at home, even though he/she is scared of contracting the virus. He/She has been suffering from hypertension for 3 years until now.

His/Her workload has been impacted on by the pandemic, and he/she does not work in shifts.

He/She is at low risk of developing burnout, anxiety, depression and PTSD.

Age: 53  
Profession: Physician

**INDEXES**  

Stress: 50  
Workload Impact: 50  
Maslach Burnout Inventory: 30  
Impact of Event Scale: 1.5  
Patient Health Questionnaire: 12

Lorenzo

**BIOGRAPHY**  

Lorenzo/Valeria is married and lives with his wife/her husband and children. To protect his/her family, he/she uses Personal Protective Equipment at home.

His/Her workload has been highly impacted on by the pandemic, and he/she does work in shifts.

He/She is at medium risk of developing burnout, and borderline low risk of anxiety, depression and PTSD.

Age: 48  
Profession: Physician

**INDEXES**  

Stress: 62.5  
Workload Impact: 62  
Maslach Burnout Inventory: 30  
Impact of Event Scale: 2  
Patient Health Questionnaire: 12

Anita

Valeria

# Personification: Persona cards development

With the obtained information from the attributes that differentiate between clusters it is possible to create Persona Cards.

Information in a Persona card can be of three types.

- **Inferred information:** usually name and photo. Even though there are no attributes directly related to these attributes they can be inferred from sex and age.
- **Quantitative information:** obtained by the data analysis. Goals are part of the quantitative information and must be expressed clearly in the Persona card. Usage of color-coding bars and indicators is highly encouraged.
- **Qualitative information:** obtained by the data analysis and other sources.

**User Persona Type**

Trait 1 Trait 2 Trait 3 Trait 4

**Goals**

- A task that needs to be completed.
- A life goal to be reached.
- Or an experience to be felt.

**Frustrations**

- The challenges this user would like to avoid.
- An obstacle that prevents this user from achieving their goals.
- Problems with the available solutions.

**Bio**

The bio should be a short paragraph to describe the user journey. It should include some of their history leading up to a current use case. It may be helpful to incorporate information listed across the template and add pertinent details that may have been left out. Highlight factors of the user's personal and of professional life that make this user an ideal customer of your product.

*Remember - you may modify this template, remove any of the modules or add new ones for your own purpose.*

**Motivation**

Incentive

Fear

Growth

Power

Social

**Brands & Influencers**

**Preferred Channels**

Traditional Ads

Online & Social Media

Referral

Guerrilla Efforts & PR

**Personality**

Age: 1-100  
Work: Job Title  
Family: Married, kids, etc.  
Location: City, state  
Character: Archetype

Introvert Extrovert

Thinking Feeling

Sensing Intuition

Judging Perceiving

*"A quotation that captures this user's personality"*

You can start working!

