Reimagining Personas: Unsupervised User Segmentation

Victor Laputsky

Lead Data Analyst / Analytics Team Lead

https://www.linkedin.com/in/victor-laputsky-44199bab/

Arnold Böcklin – The Elysian Fields

This time we’ll speak about combination of Google Analytics data with common unsupervised learning ML algorithms: PCA and K-means.

This story is not universal, but I hope it will inspire you to do something new

Angelica Kauffman – The Family of the Earl of Gower

Persona 1: Daniel
Curious Browser

I do not know my educational goals so I…

1. Discover which area of study piques my interest
2. Understand what education paths I can take in areas of study
3. Understand potential career path options
4. Have a guide walk me through the process for furthering my education
5. Narrow down on what sorts of programs or products will help me achieve my goals

Persona 2: Jane
Beginning Learner

I have an idea of what area of study I want to pursue so I…

1. Learn more about what types of programs are available in my particular interest area
2. Understand what types of programs help me achieve my education goals
3. Compare products that aligns with my interests
4. Speak with an expert who can answer my questions on specific products and their pros/cons

Persona 3: Bob
Determined Learner

I know what program I'm interested in so I want to…

1. Understand what institutions offer my program of interest
2. Compare the different programs that I'm interested in
3. Complete a quiz to determine which program best fits
4. Understand what career options might be available after completing my program
5. Access resources that will prepare me to apply for my programs

Persona 4: Alex
Advanced Learner

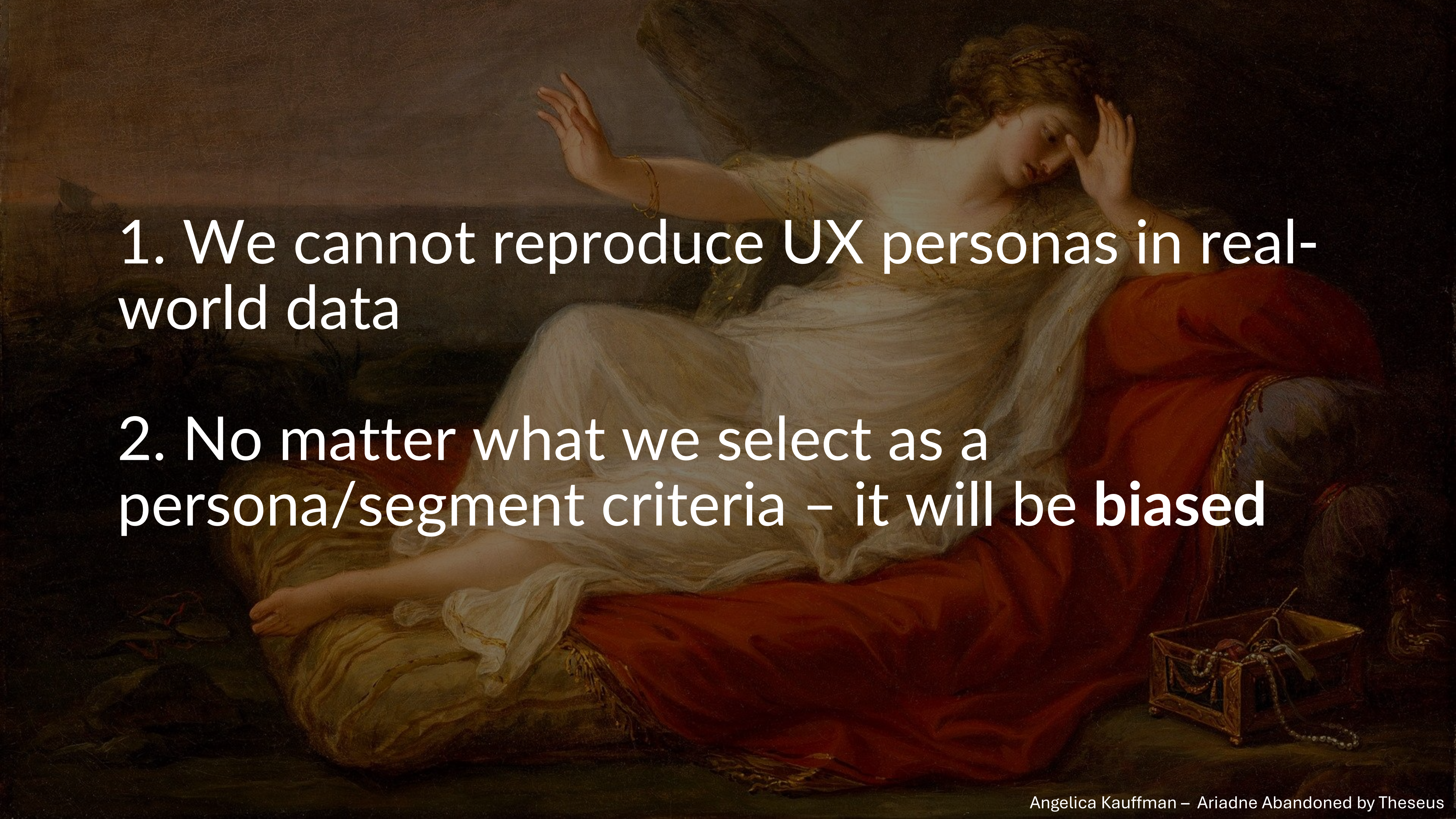
I know the institution and program I want so I…

1. View details about a specific degree offered by my institution of choice
2. Apply to programs that align to my goals within my institution of choice
3. Understand what career options might be available after completing my program
4. Access to resources that will prepare me to apply for my programs

Persona 5: Beth
Brand Learner

I know the Institution I want to apply to, but I'm not sure about what I want to study so I want to…

1. View further details on my chosen Institution to verify that it is a good fit
2. Understand the programs that are available at my chosen Institution
3. Utilize resources to determine what my academic goals are
4. Narrow down programs at my chosen Institution that will help me achieve my academic goals
5. Compare programs to help ultimately determine the best fit for me
6. Determine which program I want to apply to



1. We cannot reproduce UX personas in real-world data

2. No matter what we select as a persona/segment criteria – it will be **biased**

Angelica Kauffman – Ariadne Abandoned by Theseus

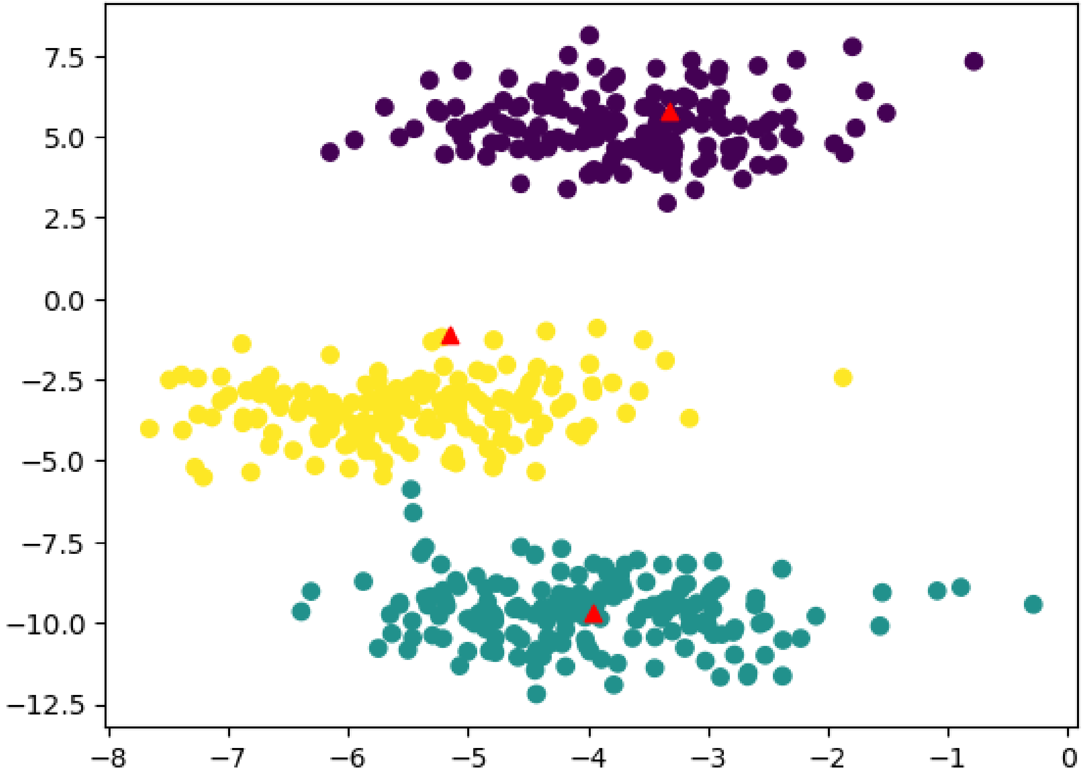
How can we make unbiased user segments?

What if to make it unsupervised?

Jean-Étienne Liotard– View of the Mont Blanc with a self-portrait

Player number one – **k-means!**

- Basic unsupervised classification method
- Aimed on grouping of items by clusters based on their similarity
- Works good with quantitative continuous metrics, a bit less good – with rank-based metrics



Player number two – **PCA!**

- Principal component analysis – common dimensionality reduction method
- Transforms selected metrics to abstract dimensions/indexes

user_id	metrics1	metrics2	metrics3	...	metrics100
0	100	27	45	...	50
1	86	17	45	...	4
2	94	21	40	...	56
3	83	29	44	...	62
4	87	18	43	...	48
5	82	11	47	...	7
6	97	23	47	...	61

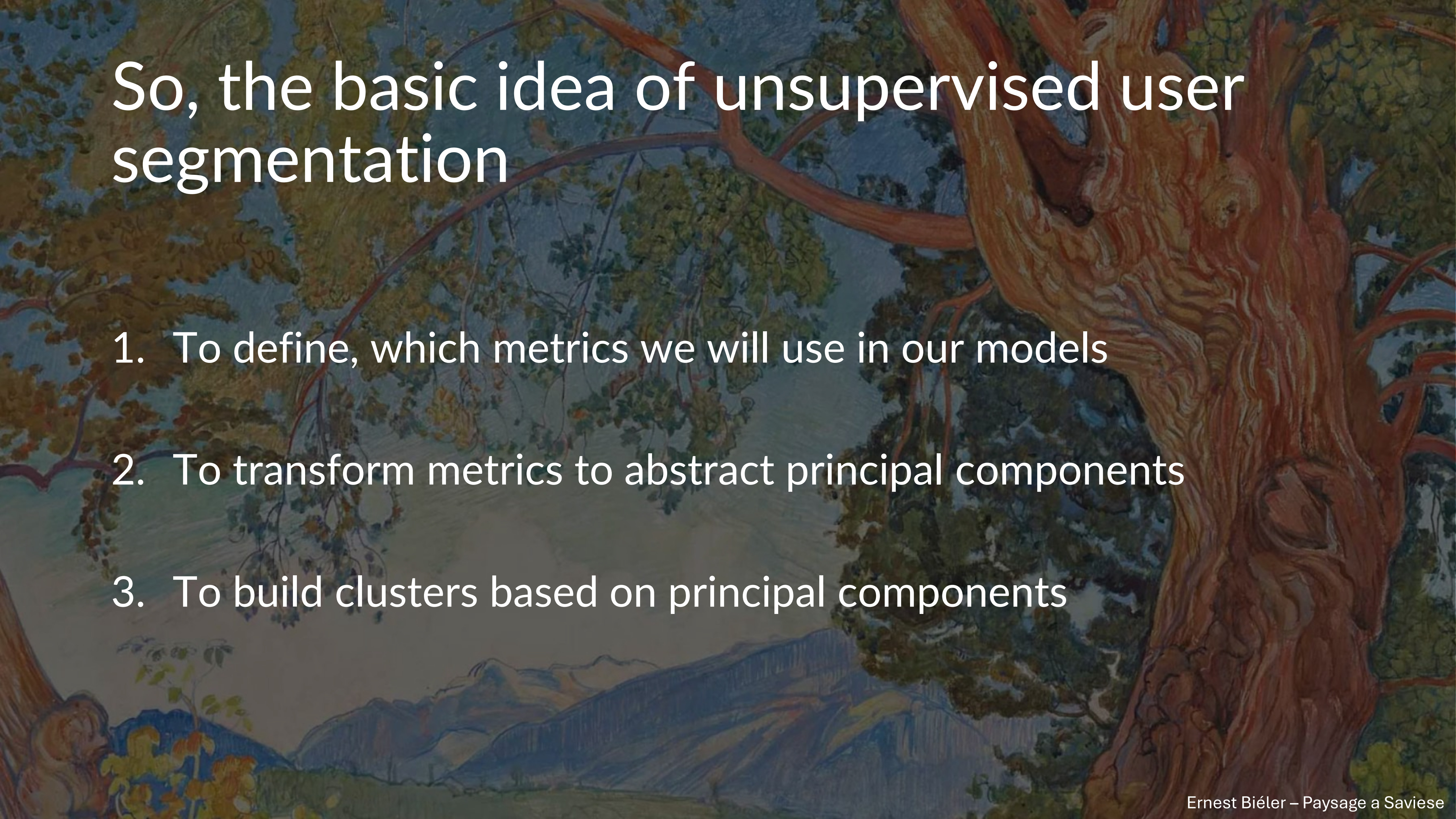
user_id	pc1	pc2
0	-0.13	-1.14
1	-0.55	0.03
2	-0.15	-1.22
3	1.66	-0.29
4	0.74	-0.68
5	-0.64	1.32
6	0.51	0.12

So, the basic idea of unsupervised user segmentation

1. To define, which metrics we will use in our models

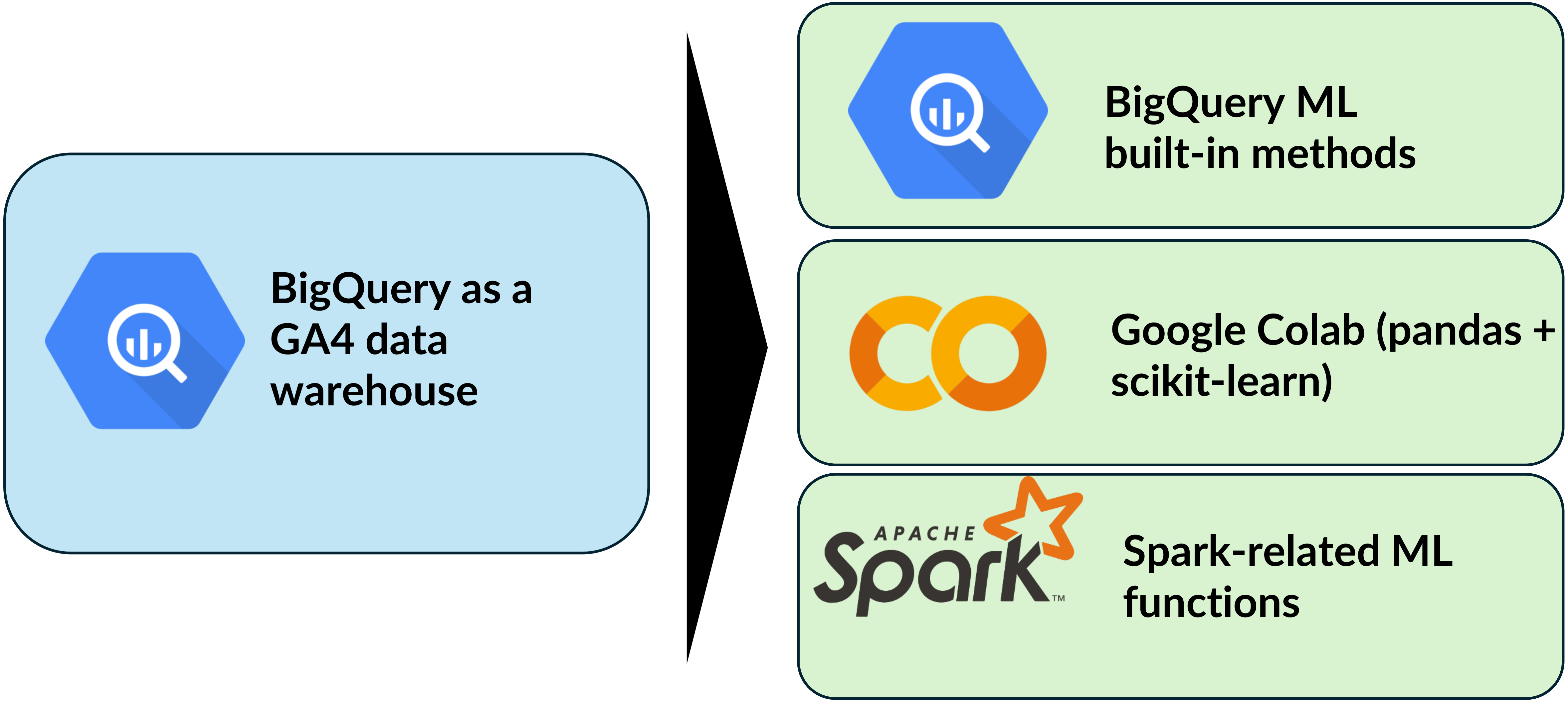
2. To transform metrics to abstract principal components

3. To build clusters based on principal components



Ernest Biéler – Paysage a Saviese

How it can be done (in case we still speak about GA4 data)



BigQuery as a GA4 data warehouse

BigQuery ML built-in methods

Google Colab (pandas + scikit-learn)

Spark-related ML functions

1. Data Collection

User behavior tracking handled by GTM + GA4

2. Data Preparation

Metrics tables preparation + feature engineering tables in BigQuery

3. Data cleaning & standardize

Outliers exclusion & transformation of metrics to the same scale

4. PCA Model Initialize

Definition of number of principal components

5. k-means model initialize

Usage of principal components as a criteria for clusters

6. Results evaluation

Clusters volume + descriptive statistics

7. Results interpretation

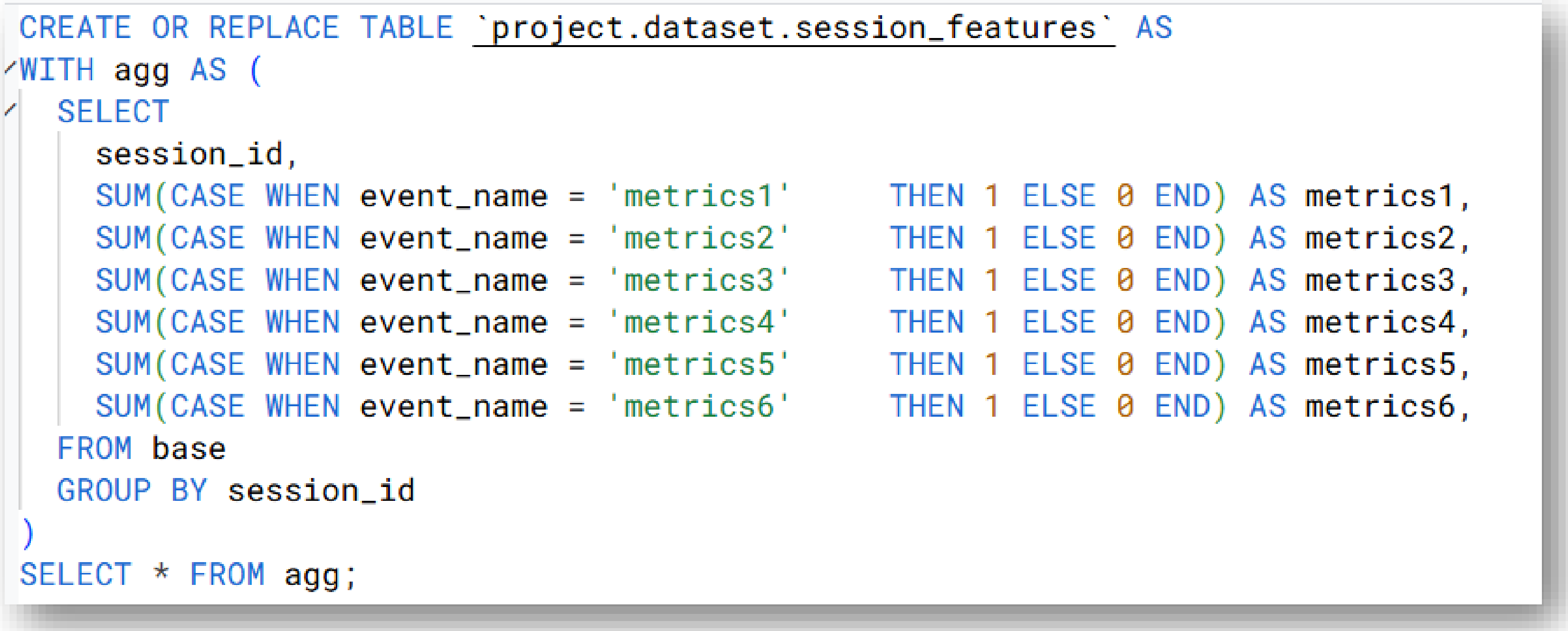
Identifying personas!

Pieter Bruegel the Elder – The Hunters in the Snow

2. Data Preparation

2.1. Making of session-based or user-based tables with target metrics as columns

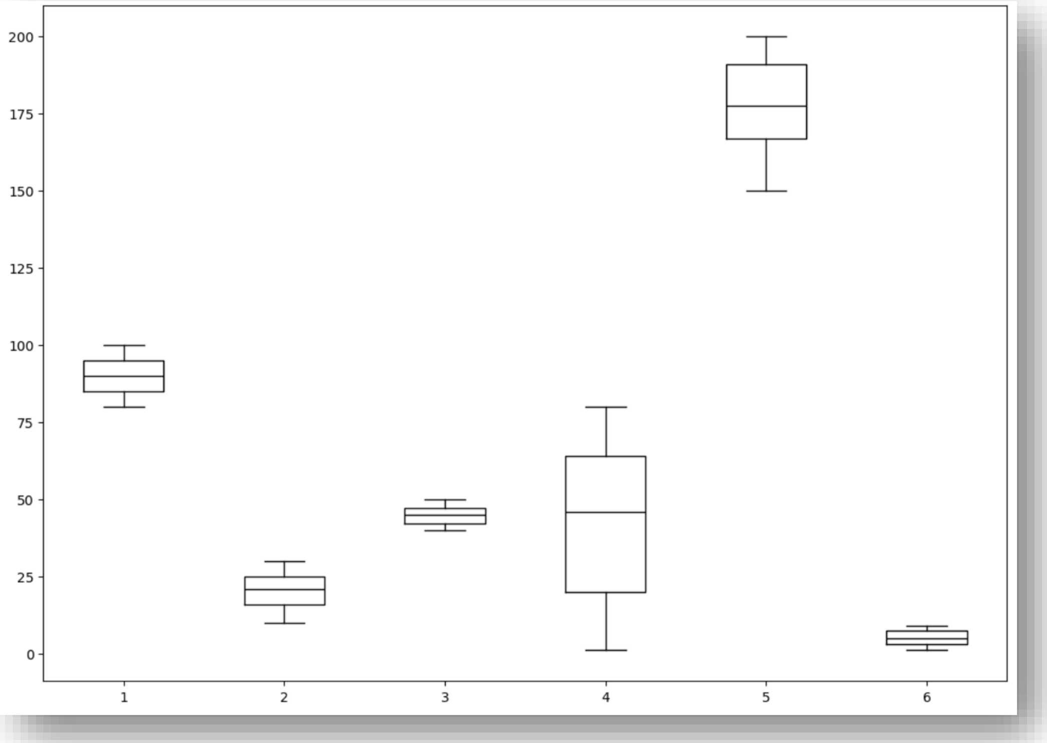
2.2. Optional – Preparation of custom metrics e.g. feature engineering

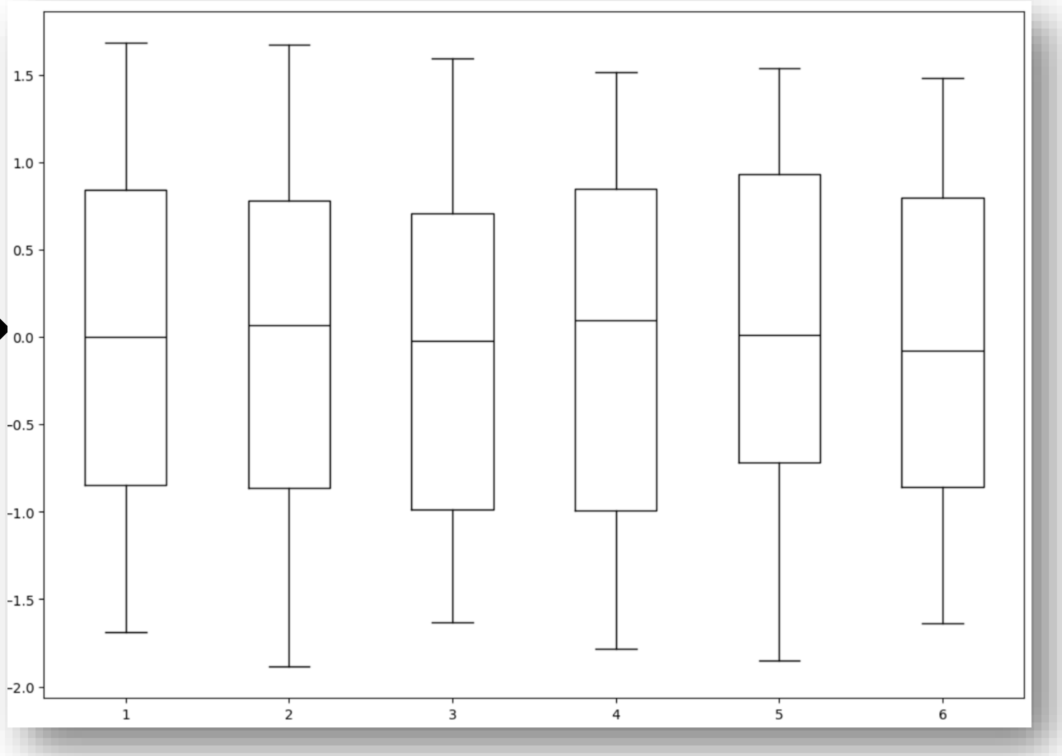


```
CREATE OR REPLACE TABLE `project.dataset.session_features` AS
WITH agg AS (
  SELECT
    session_id,
    SUM(CASE WHEN event_name = 'metrics1'     THEN 1 ELSE 0 END) AS metrics1,
    SUM(CASE WHEN event_name = 'metrics2'     THEN 1 ELSE 0 END) AS metrics2,
    SUM(CASE WHEN event_name = 'metrics3'     THEN 1 ELSE 0 END) AS metrics3,
    SUM(CASE WHEN event_name = 'metrics4'     THEN 1 ELSE 0 END) AS metrics4,
    SUM(CASE WHEN event_name = 'metrics5'     THEN 1 ELSE 0 END) AS metrics5,
    SUM(CASE WHEN event_name = 'metrics6'     THEN 1 ELSE 0 END) AS metrics6,
  FROM base
  GROUP BY session_id
)
SELECT * FROM agg;
```

Limitation! *PCA is sensitive to non-linear metrics dynamics, in that case more advanced methods can be used*

3. Data cleaning & standardize





Automatic options:

- Built-in BigQuery ML/ Scikit-learn/PySpark options (during model initiation)

Manual options:

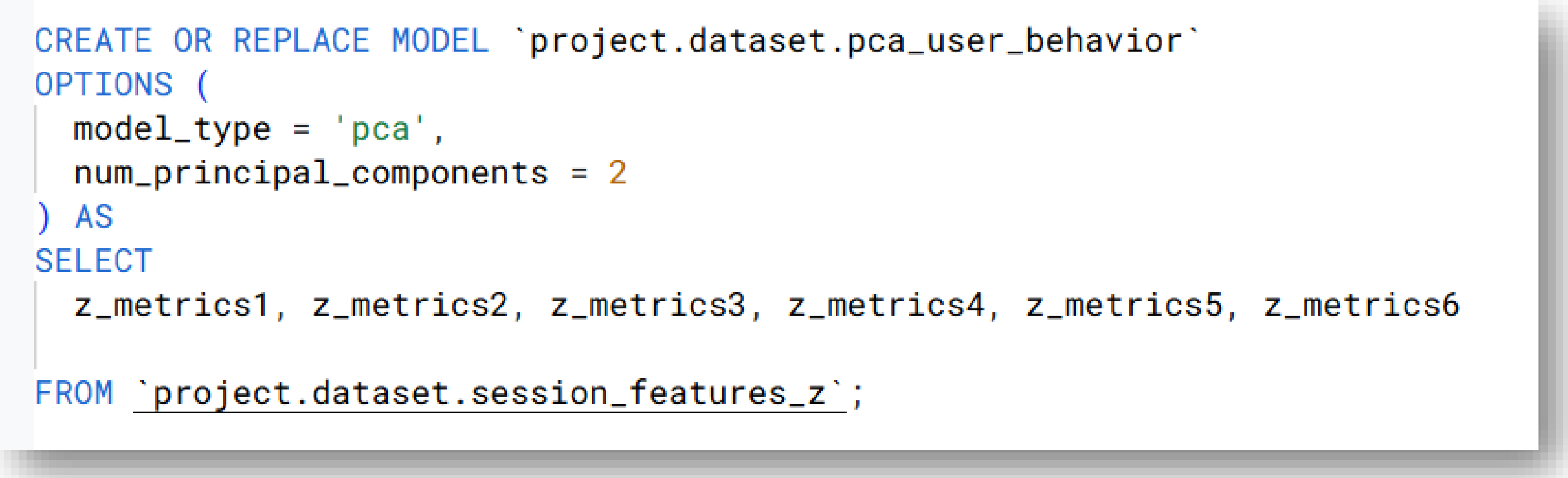
- Z-score normalization
- Exclusion of outliers (99th or 95th percentile)
- Log1p approach

4. PCA Model Initialization

On this step, we can either select number of components or define target explained variance.

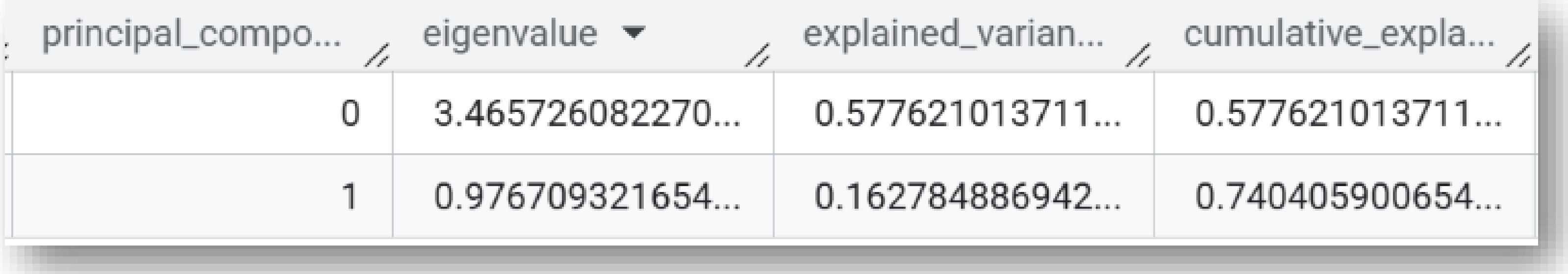
After created model, we also should check created model metrics:
- explained variance,
- pivot between principal components and metrics

If results are good, we should merge its results with the main table for further cluster analysis.



```
CREATE OR REPLACE MODEL `project.dataset.pca_user_behavior`
OPTIONS (
  model_type = 'pca',
  num_principal_components = 2
) AS
SELECT
  z_metrics1, z_metrics2, z_metrics3, z_metrics4, z_metrics5, z_metrics6

FROM `project.dataset.session_features_z`;
```

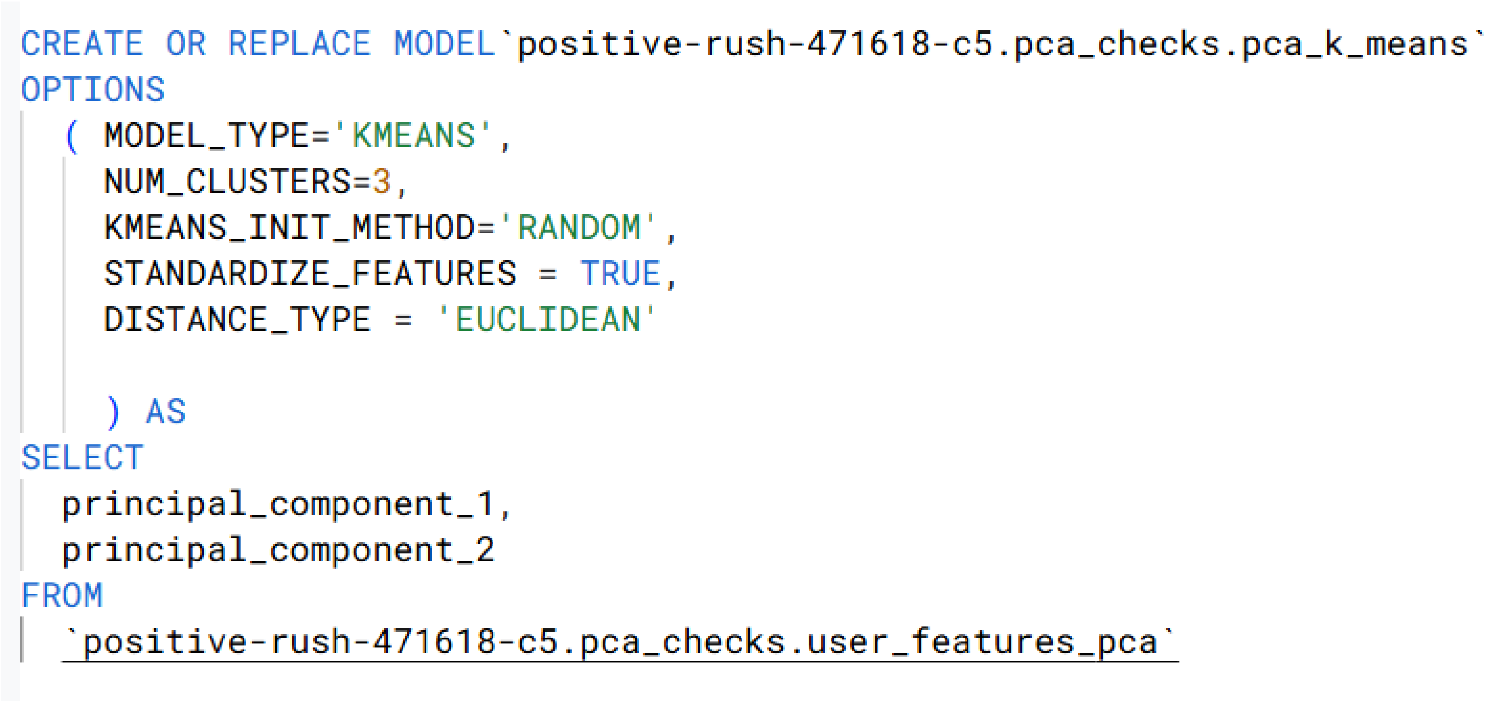


principal_compo...	eigenvalue ▼	explained_varian...	cumulative_expla...
0	3.465726082270...	0.577621013711...	0.577621013711...
1	0.976709321654...	0.162784886942...	0.740405900654...

5. k-means Model Initialization

On this step, we can define:

- Number of clusters
- Cluster centers definition approach (random / predefined centers / etc.)
- Distance calculation approach (Euclidian / Cosine)



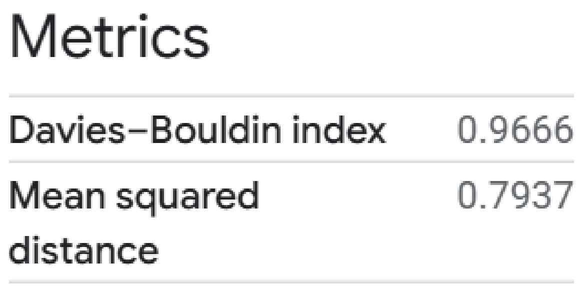
```
CREATE OR REPLACE MODEL`positive-rush-471618-c5.pca_checks.pca_k_means`
OPTIONS
  ( MODEL_TYPE='KMEANS',
    NUM_CLUSTERS=3,
    KMEANS_INIT_METHOD='RANDOM',
    STANDARDIZE_FEATURES = TRUE,
    DISTANCE_TYPE = 'EUCLIDEAN'

    ) AS
SELECT
  principal_component_1,
  principal_component_2
FROM
  `positive-rush-471618-c5.pca_checks.user_features_pca`
```

6. Clusters interpretation

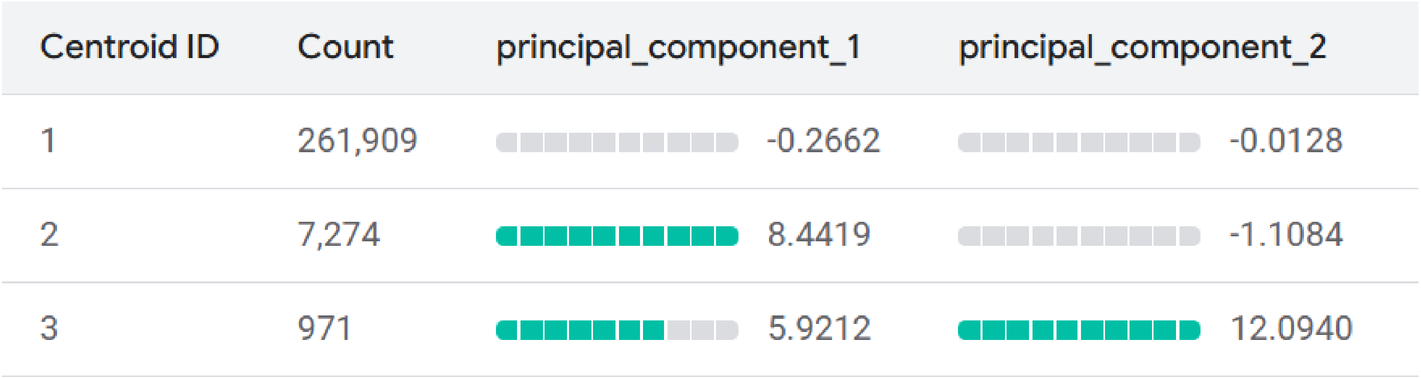
We should check:

- Each cluster volume
- K-means-related clustering metrics
- Descriptive statistics across target metrics (after merging clusters data with the main table)



Metrics

Davies–Bouldin index	0.9666
Mean squared distance	0.7937



Centroid ID	Count	principal_component_1	principal_component_2
1	261,909	-0.2662	-0.0128
2	7,274	8.4419	-1.1084
3	971	5.9212	12.0940

7. Personas definition

A blue ocean of opportunities:

- Comparing user flows
- Funnels performance comparison
- Definition of specific patterns in visited pages and performed actions
- Session-level and user-level metrics analysis
- A lot of more!

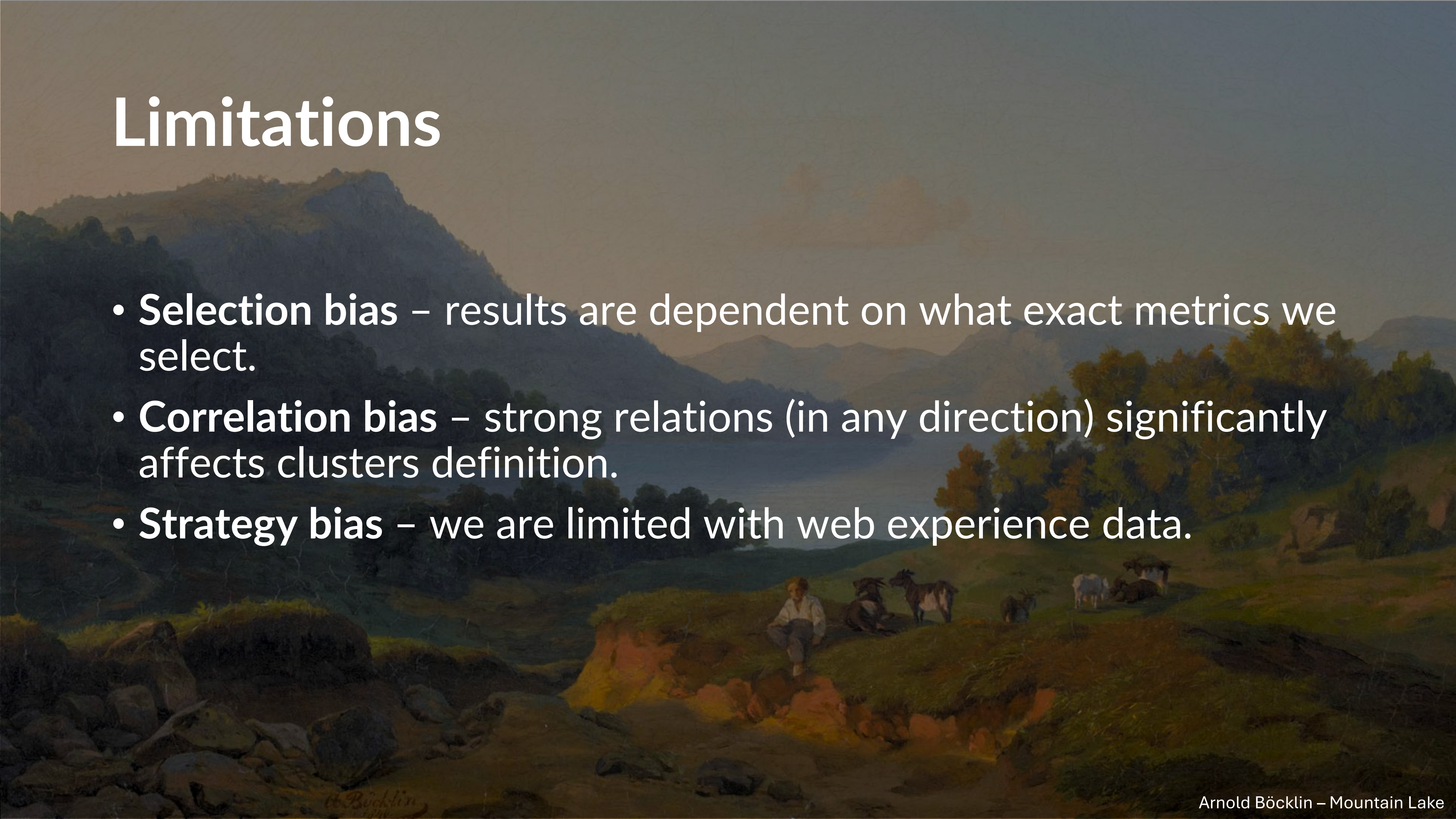
Nicolas Régnier – Self-portrait with an easel

What can we do with that next?

Jacques-Laurent Agasse – Landing at Westminster Bridge

Limitations

- **Selection bias** – results are dependent on what exact metrics we select.
- **Correlation bias** – strong relations (in any direction) significantly affects clusters definition.
- **Strategy bias** – we are limited with web experience data.



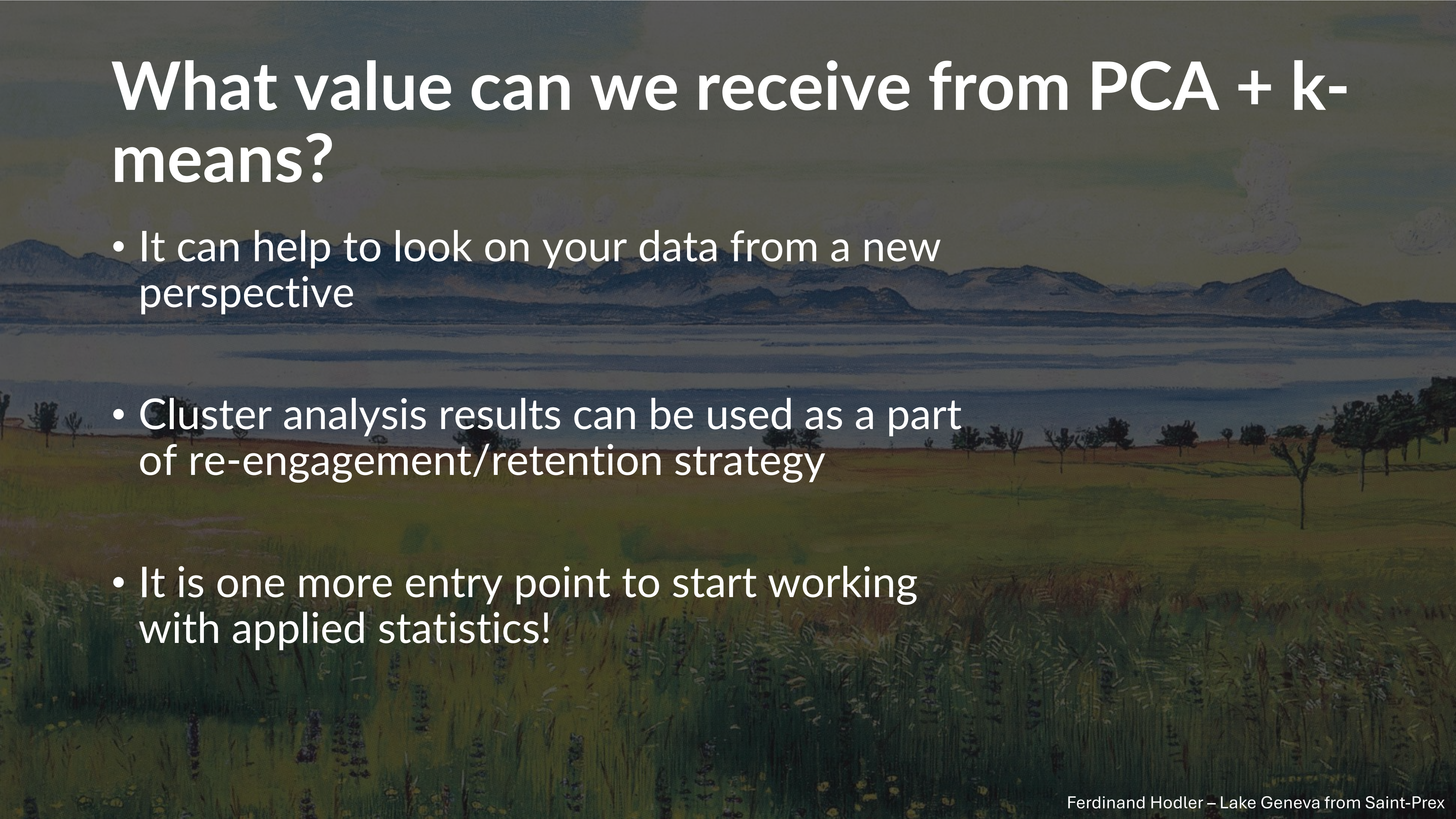
Arnold Böcklin – Mountain Lake

What value can we receive from PCA + k-means?

- It can help to look on your data from a new perspective

- Cluster analysis results can be used as a part of re-engagement/retention strategy

- It is one more entry point to start working with applied statistics!



Ferdinand Hodler – Lake Geneva from Saint-Prex

Let’s connect!

- GitHub with scripts: https://github.com/Lunthu/ga4_pca
- Linkedin: https://www.linkedin.com/in/victor-laputsky-44199bab/
- Kaggle: https://www.kaggle.com/lunthu