Data Barbarians: PCA-driven User Engagement Measurement

Victor Laputsky

Lead Data Analyst / Analytics Team Lead

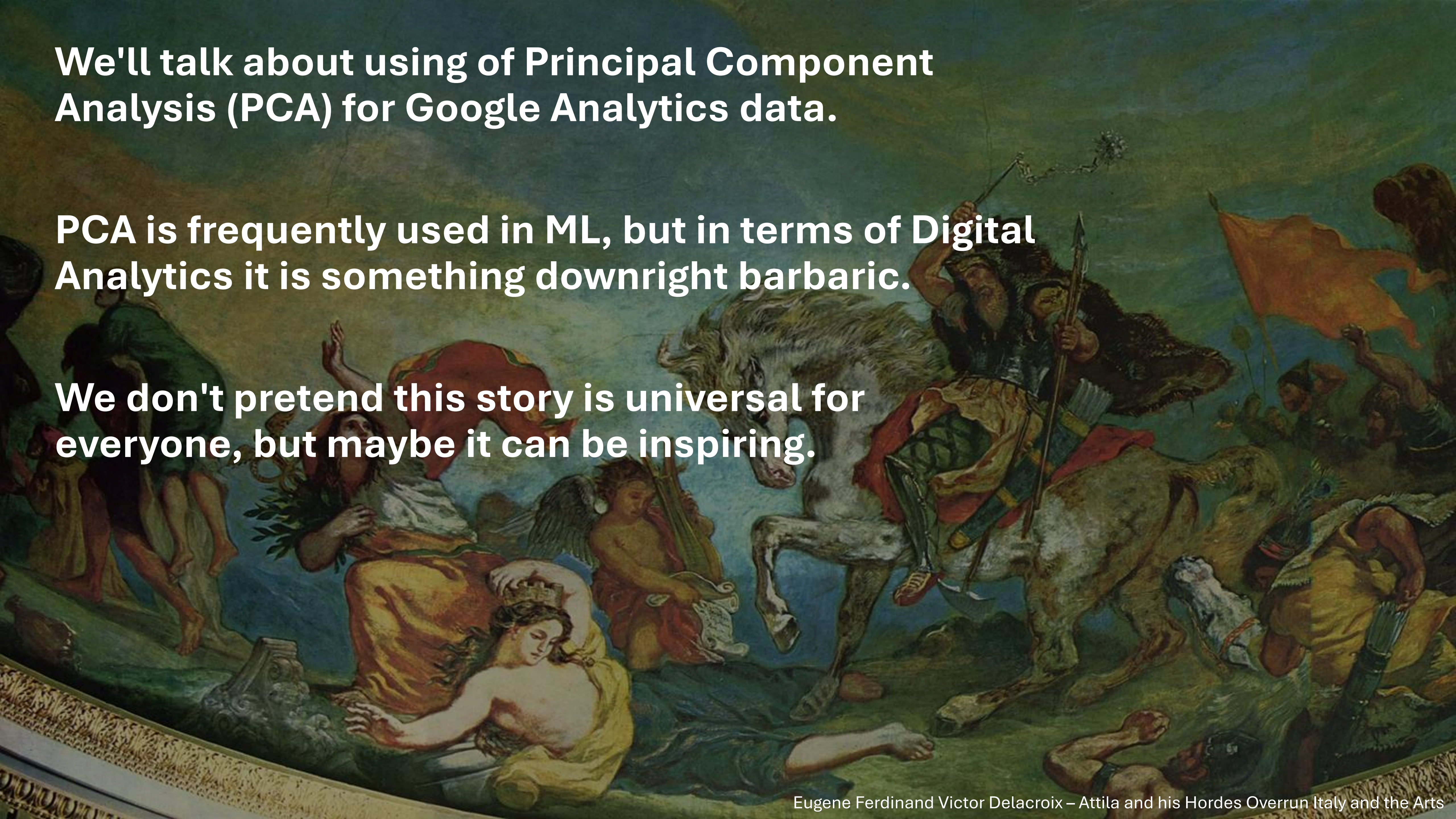
https://www.linkedin.com/in/victor-laputsky-44199bab/

Ulpiano Checa – Invasion of the Barbarians

We'll talk about using of Principal Component Analysis (PCA) for Google Analytics data.

PCA is frequently used in ML, but in terms of Digital Analytics it is something downright barbaric.

We don't pretend this story is universal for everyone, but maybe it can be inspiring.

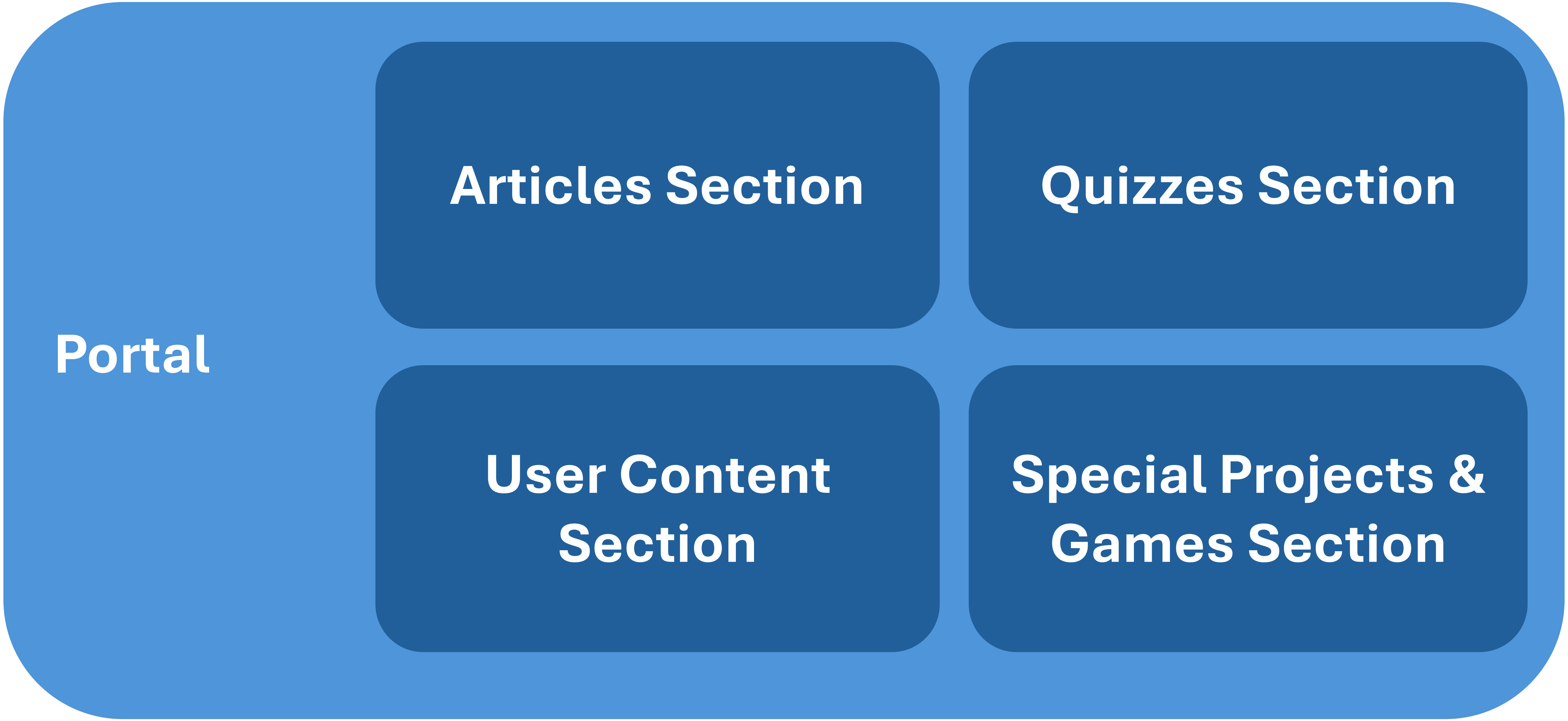


Eugene Ferdinand Victor Delacroix – Attila and his Hordes Overrun Italy and the Arts

Let’s start!

Thomas Cole – The Course of Empire: The Consummation of Empire

Let’s imagine that we have a media portal!



Each section has its own development team, core audience and engagement metrics

Portal:

- Website-level Metrics 1
- Website-level Metrics 2
- Website-level Metrics 3

Articles Section

- Metrics 1
- Metrics 2
- Metrics 3

Quizzes Section

- Metrics 1
- Metrics 2
- Metrics 3

User Content Section

- Metrics 1
- Metrics 2
- Metrics 3

Special Projects & Games Section

- Metrics 1
- Metrics 2
- Metrics 3

How can we measure user engagement?

Pieter Bruegel the Elder – The Fight Between Carnival and Lent

Let’s use basic GA4-driven engagement metrics

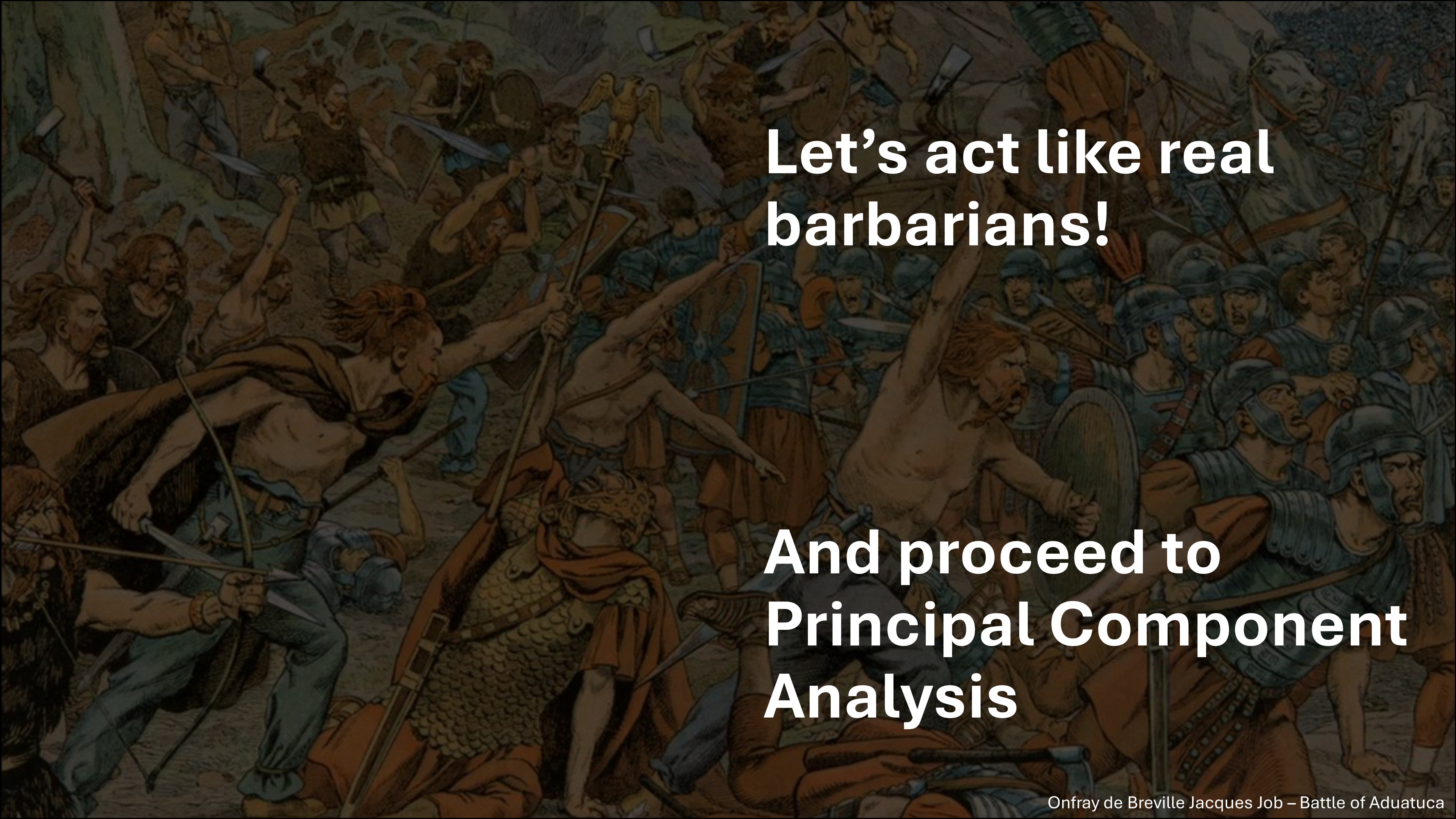
Sometimes it is too generic/non-indicative

Let’s rely on website-level metrics

It cannot say how separate sections perform

Let’s rely on section-specific metrics

It doesn’t show a full picture



Let’s act like real barbarians!

And proceed to Principal Component Analysis

Onfray de Breville Jacques Job – Battle of Aduatuca

Principal Component Analysis is (almost) your personal data archiver

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

- PCA transforms selected metrics to abstract dimensions/indexes
- In terms of digital analytics, principal components can be used in:
  - metrics patterns reveal
  - user segmentation
  - building of abstract-level metrics

We already can do it!

- BigQuery ML
- Google Colab (pandas/numpy + scikit learn)
- R
- IBM SPSS
- A lot of other variants!

1. Data Collection

User behavior tracking handled by GTM + GA4

2. Data Preparation/ Transformation

Transforming of raw data to actionable 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. Results validation & evaluation

Pieter Bruegel the Elder – The Harvesters

Data Preparation

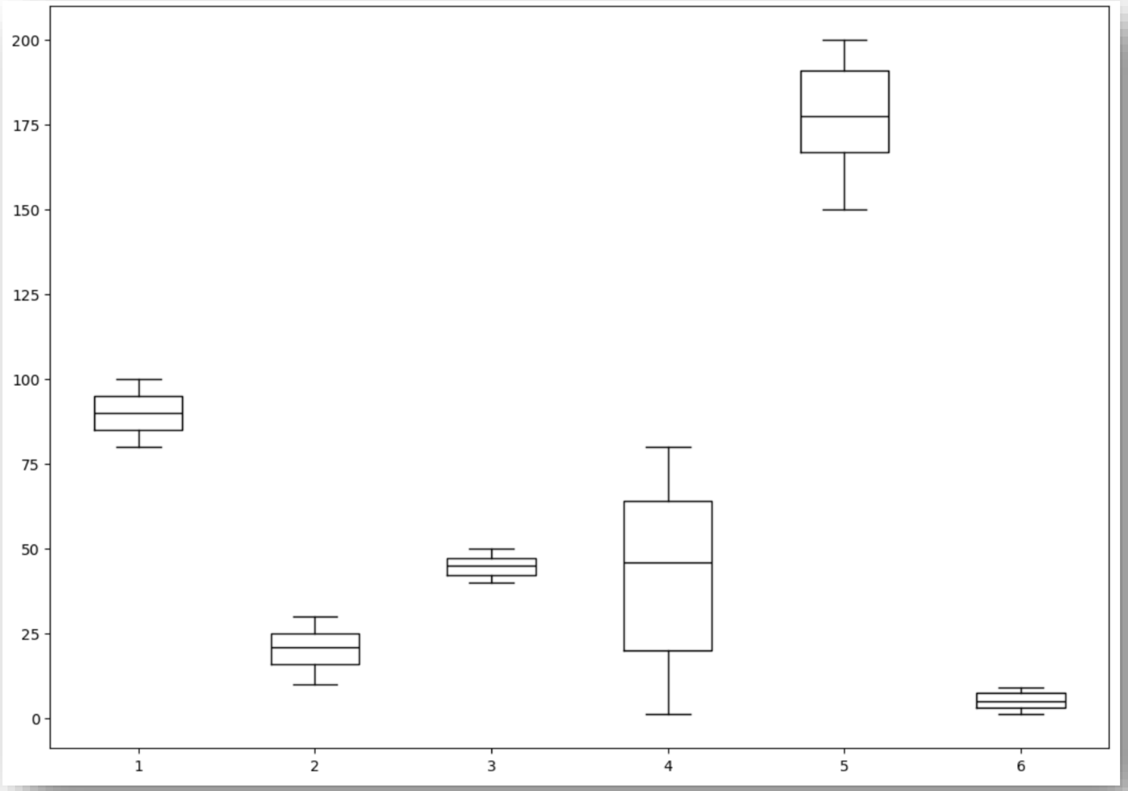
Making of session-based or user-based tables with engagement metrics as columns

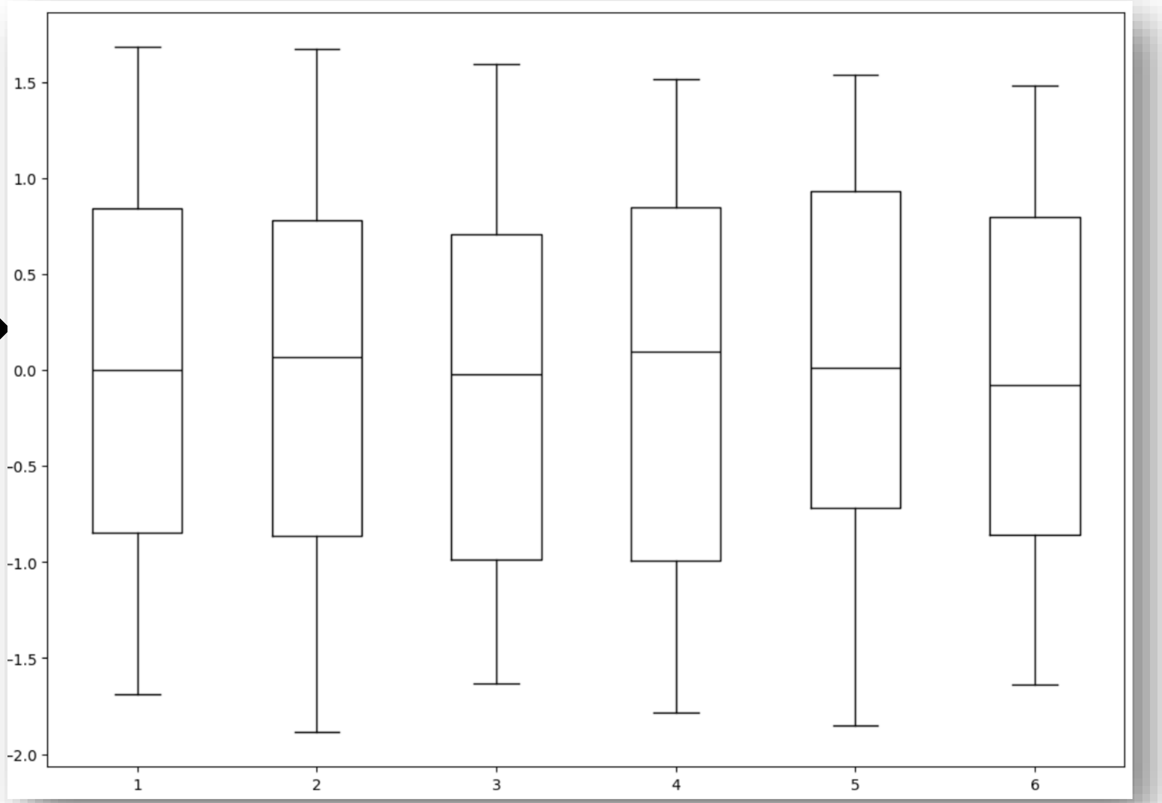
Limitations:

- It is applicable only for numerical or ordinal variables
- It is not so effective in non-linear cases – in that case better use kPCA or other DR methods

```sql
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;
```

Data cleaning & standardize





Can be done with:

- Built-in BigQuery ML or Scikit-learn options
- Z-score normalization
- Exclusion of outliers (99^{th} or 95^{th} percentile)
- Log1p approach

PCA Model Initialization

In BigQuery ML, we can either select number of components or define target explained variance

```
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`;
```

Checking results

Top things to check:

- Explained variance (bigger is better)
- Pivot between principal components and initial metrics

Limitation: In case our web product changes significantly over time, we need to re-train our PCA model

Metrics

Total explained variance ratio	0.8546

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

feature	pc1	pc2
z_metrics1	0.52	-0.07
z_metrics2	0.35	0.10
z_metrics3	0.49	-0.06
z_metrics4	0.49	-0.02
z_metrics5	0.13	0.96
z_metrics6	0.33	-0.25

What can we do with that next?

Fra Carnevale – The Ideal City

Gentile Bellini – Procession in St. Mark's Square

Nicolas Bertin – Hercules fighting Achelous

A path of wisdom

Source: https://dl.acm.org/doi/10.1145/3341981.3344222

	PC1 (AbandQs)	PC2 (AbandCs)	PC3 (DeepSERP)	PC4 (Pace)	PC5 (NLQs)	PC6 (SlowCs)
QueriesWOBooks	0.87	0.09	0.27	0.04	0.06	-0.03
QueriesWOClicks	0.86	-0.03	0.21	0.21	-0.04	0.02
Queries	0.85	0.16	0.10	-0.29	0.15	0.04
RepeatedIntentQs	0.85	0.14	0.24	-0.02	0.02	0.12
QuickReforms	0.83	0.00	0.09	-0.18	-0.20	-0.05
UniqueQueries	0.82	0.15	0.09	-0.16	0.37	0.00
QueriesWOMouse	0.71	0.02	-0.08	0.15	-0.16	-0.09
UniqueQueryTerms	0.62	0.04	-0.05	0.04	0.36	0.30
QueriesWOScrolls	0.60	0.16	-0.44	-0.33	-0.03	-0.04
Clicks	0.16	0.83	0.22	-0.37	0.07	-0.18
ClicksWOBooks	0.18	0.79	0.25	0.13	0.02	-0.24
UniqueURLs	0.04	0.76	0.08	-0.05	0.12	0.09
CompletionTime	0.05	0.57	0.08	0.53	0.08	0.34
MouseWOClicks	0.34	0.18	0.85	-0.11	0.14	0.06
Mouseovers	0.16	0.40	0.82	-0.09	0.07	-0.02
Paginations	-0.09	0.18	0.78	0.01	-0.09	0.10
ScrollDistance	0.36	-0.06	0.77	-0.03	0.22	0.12
AvgTimeBWEvents	-0.26	-0.05	-0.07	0.83	-0.03	0.31
Bookmarks	-0.02	0.27	-0.07	-0.78	-0.01	0.09
TimeToFirstBook	0.06	0.19	-0.17	0.70	-0.12	0.22
QuestionQueries	0.02	0.08	0.13	0.01	0.83	-0.08
AvgQueryLength	0.01	0.10	0.03	-0.09	0.82	0.10
TimeToFirstClick	0.04	-0.02	0.02	0.12	0.06	0.82
Avg1stClickTime	0.05	-0.13	0.34	0.30	-0.04	0.70

Metrics patterns definition

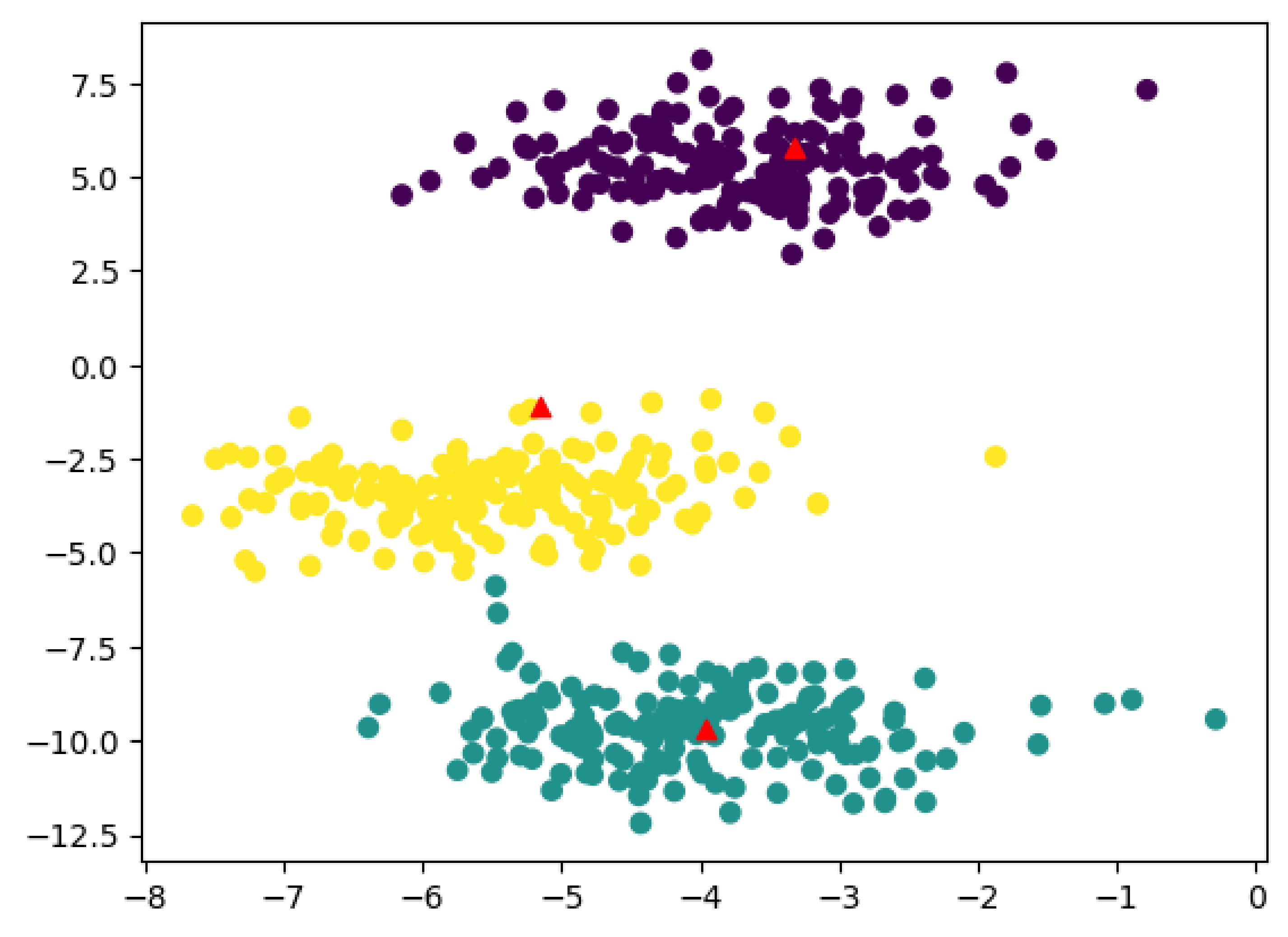
Checking metrics numeric values + grouping them by principal components

Fra Carnevale – The Ideal City

A path of might

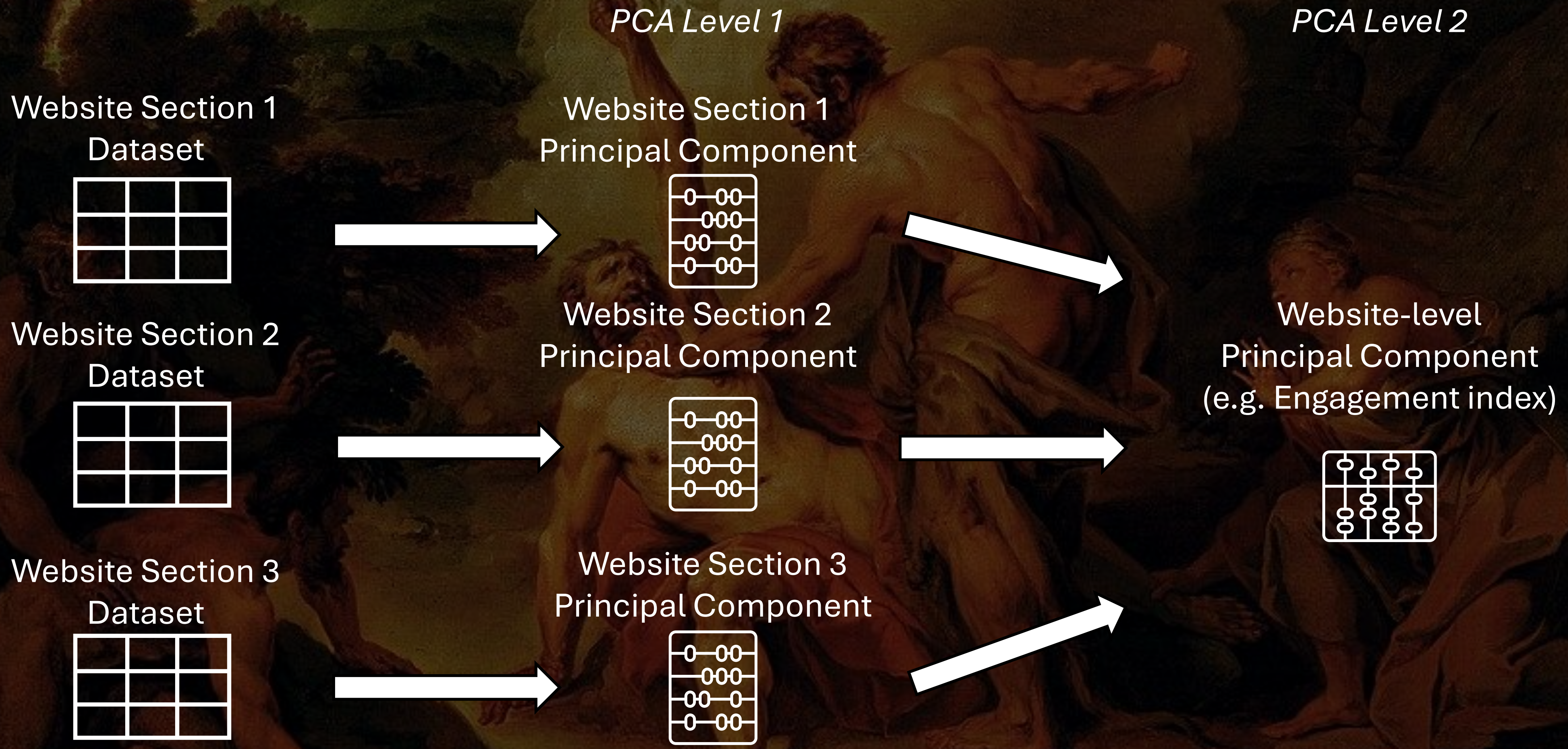
Segmentation based on principal components

Saving of principal component values in a separate table + application of k-means clustering



Gentile Bellini – Procession in St. Mark's Square

A path of courage – I totally would not recommend it!



Nicolas Bertin – Hercules fighting Achelous

Summary

Giovanni Paolo Panini – Modern Rome

Why is it good to try PCA?

- It is another good entry point to start working with applied statistics!
- It can help to uncover additional patterns to user engagement understanding
- Findings of PCA then can be used in terms of propensity modelling
- **We need to do something barbaric before transition to Ancient Greeks of Digital Analysis**

Raphael – School of Athens

Let’s connect!

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