Automatic Generation of Marketing Personas

Extracting insights from social networks

Supervisor
Alberto Montresor

Co-Supervisors
Carlo Caprini
Daniele Miorandi

Bachelor Degree in Computer Science

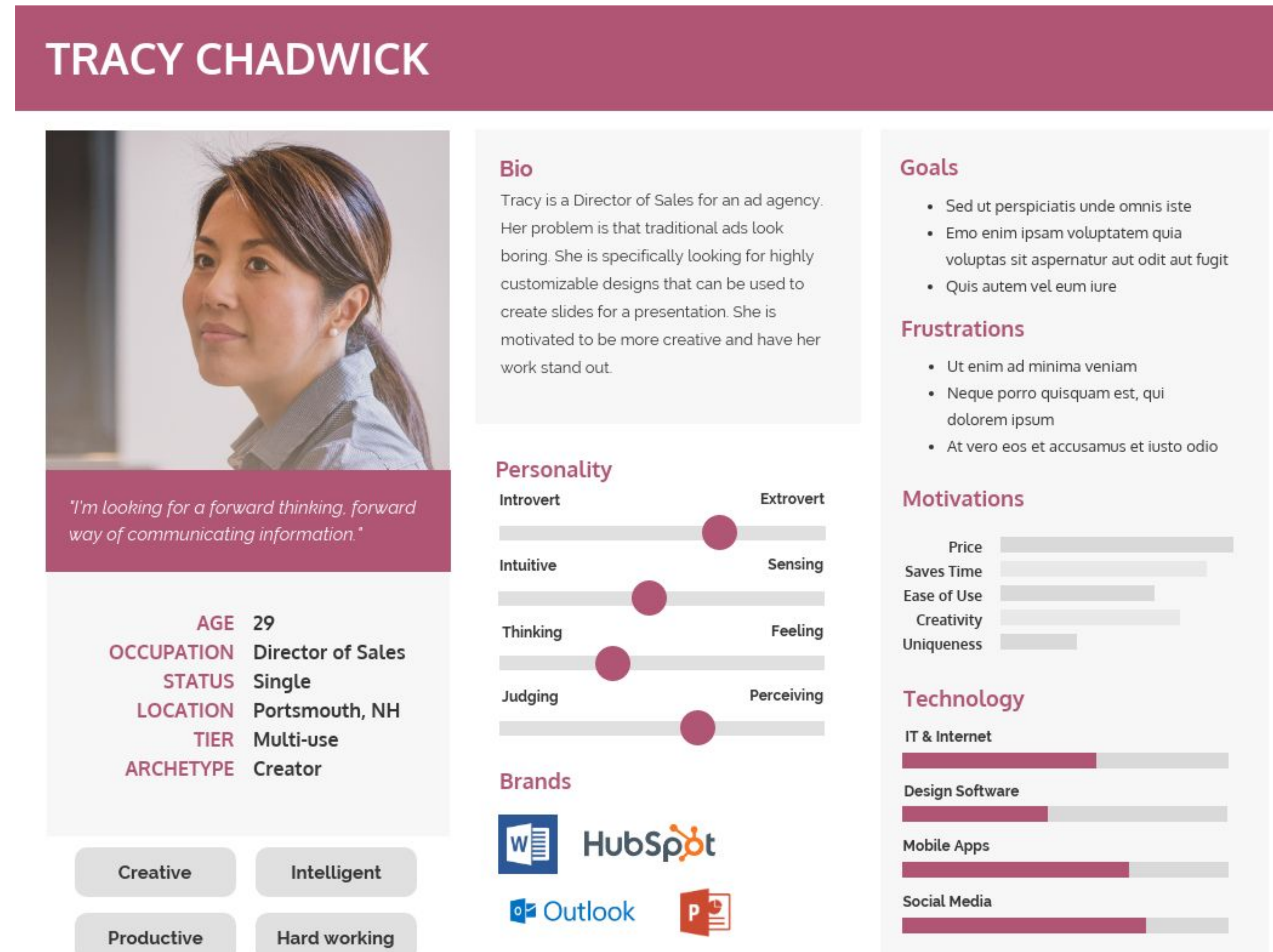
Department of Information Engineering and Computer Science

Academic Year 2020/2021

Student
Nicola Farina

Marketing Personas

A persona is a fictional character that communicates the primary characteristics of a group of users. [1]



pros:

- personalized customer experience
- easier to plan marketing campaigns

cons:

- long time to create
- high costs

[1] https://www.smartinsights.com/persuasion-marketing/marketing-personas/

Research Question

Is it possible to **automatically** generate marketing personas through the use of **machine learning**?

Automatically:

- no need for user input

Machine learning:

- clustering
- classification

State of the Art

Current focus on using **social network data,** from which can be extracted:

demographics

- **gender**
- **age**
- **location**
- **job**
- **income**

behavioral insights

- **personality**
- **interests**
- **attitude**

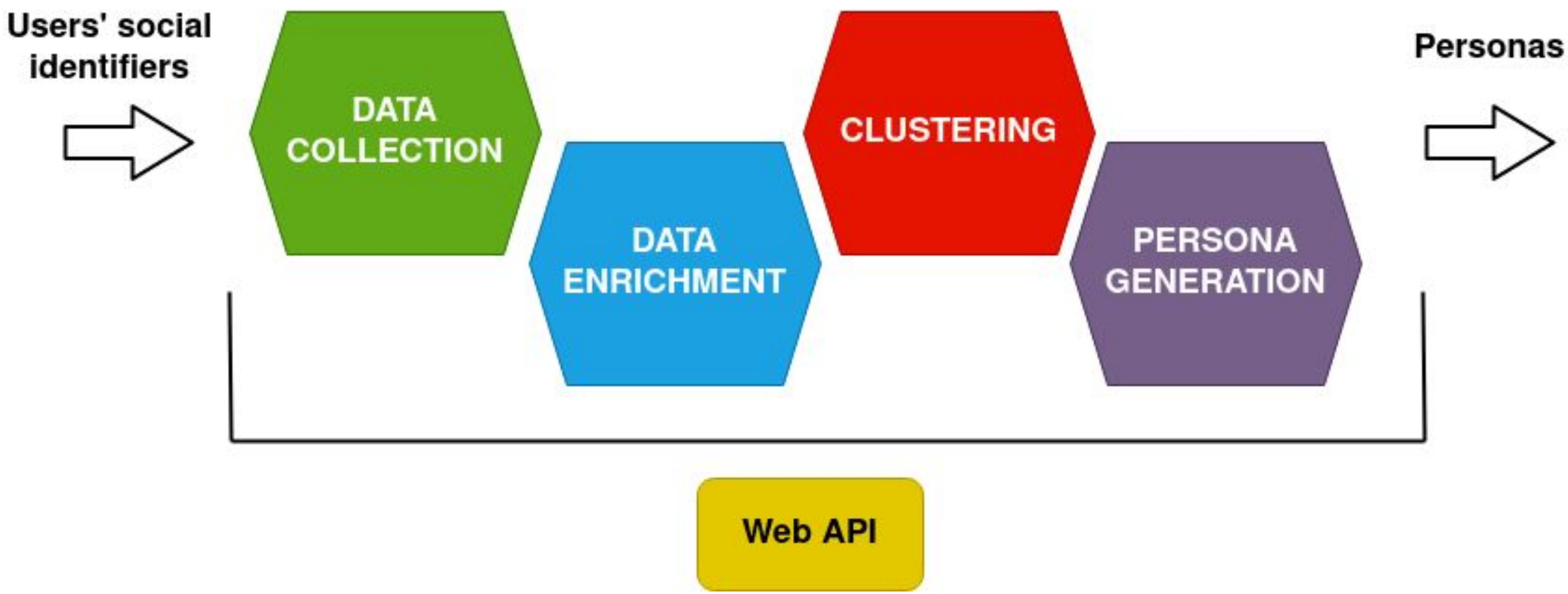
→ customers **grouped** based on such insights

Legal Basis:

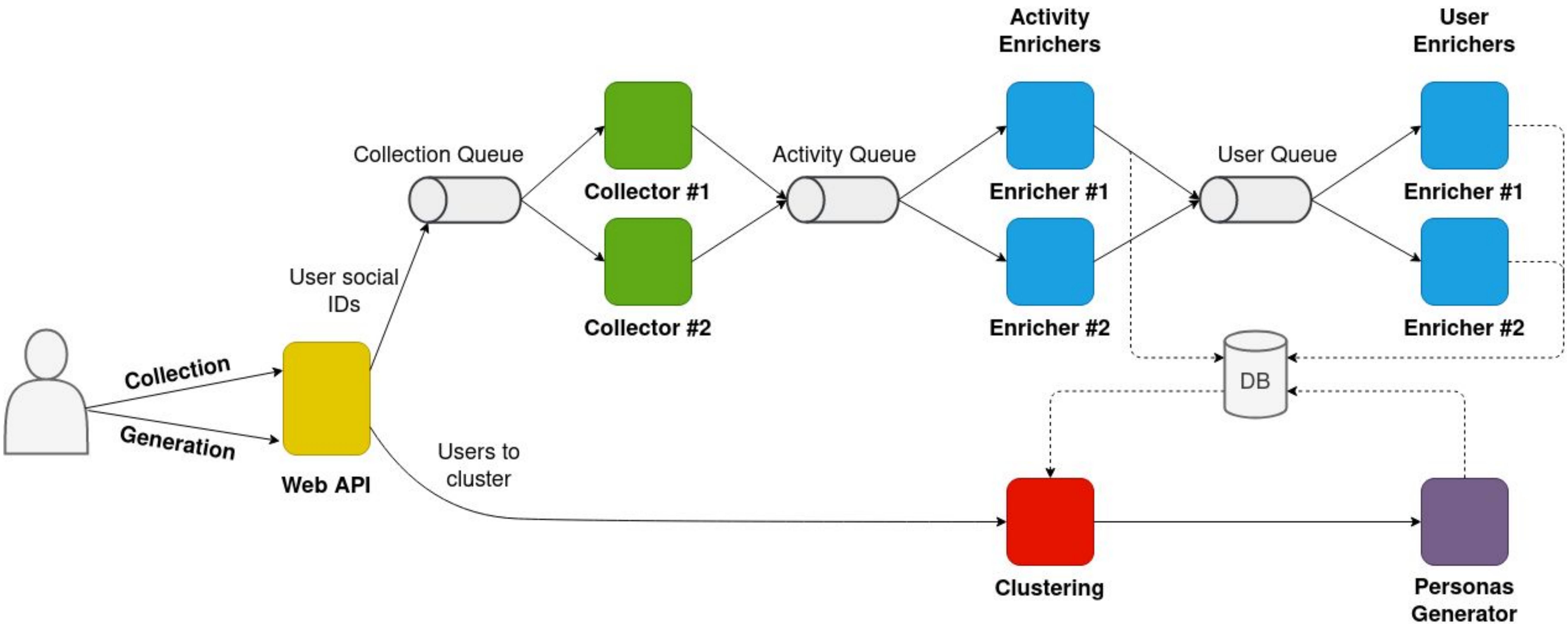
- GDPR compliance

GDPR | General Data Protection Regulation

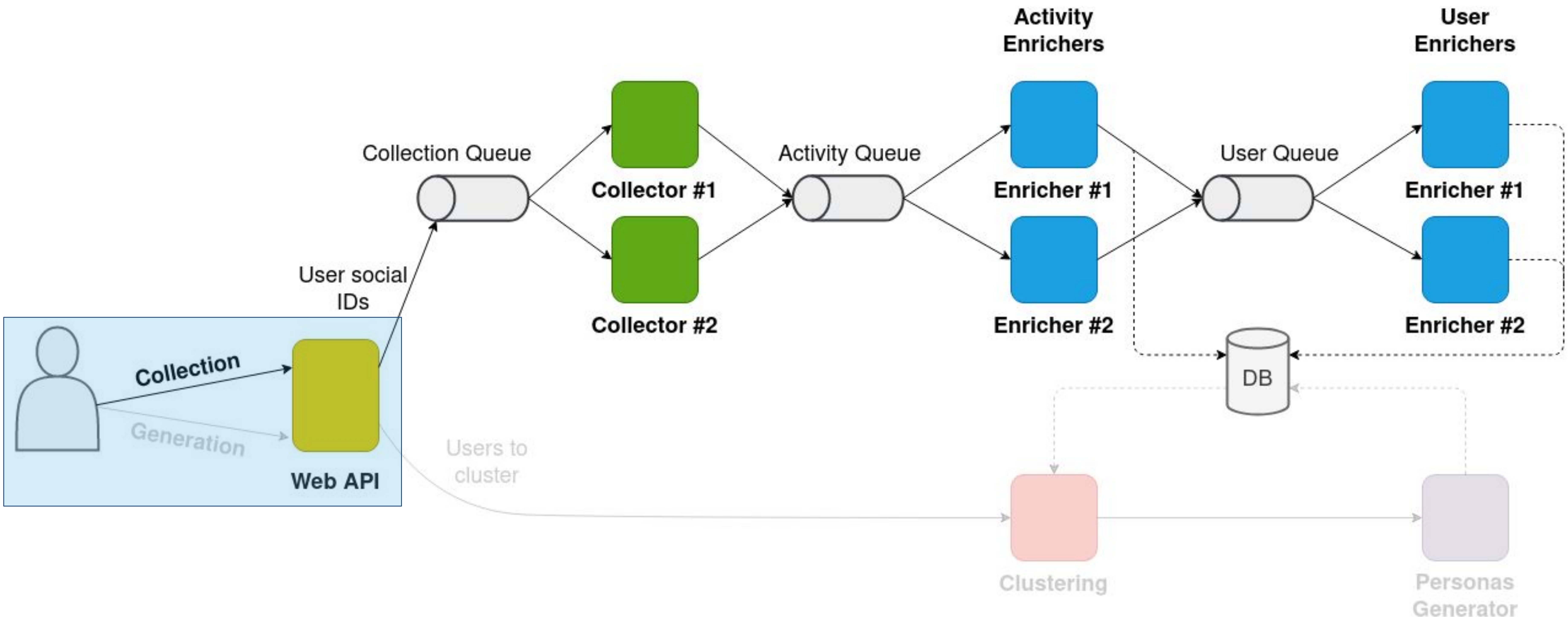
Solution design



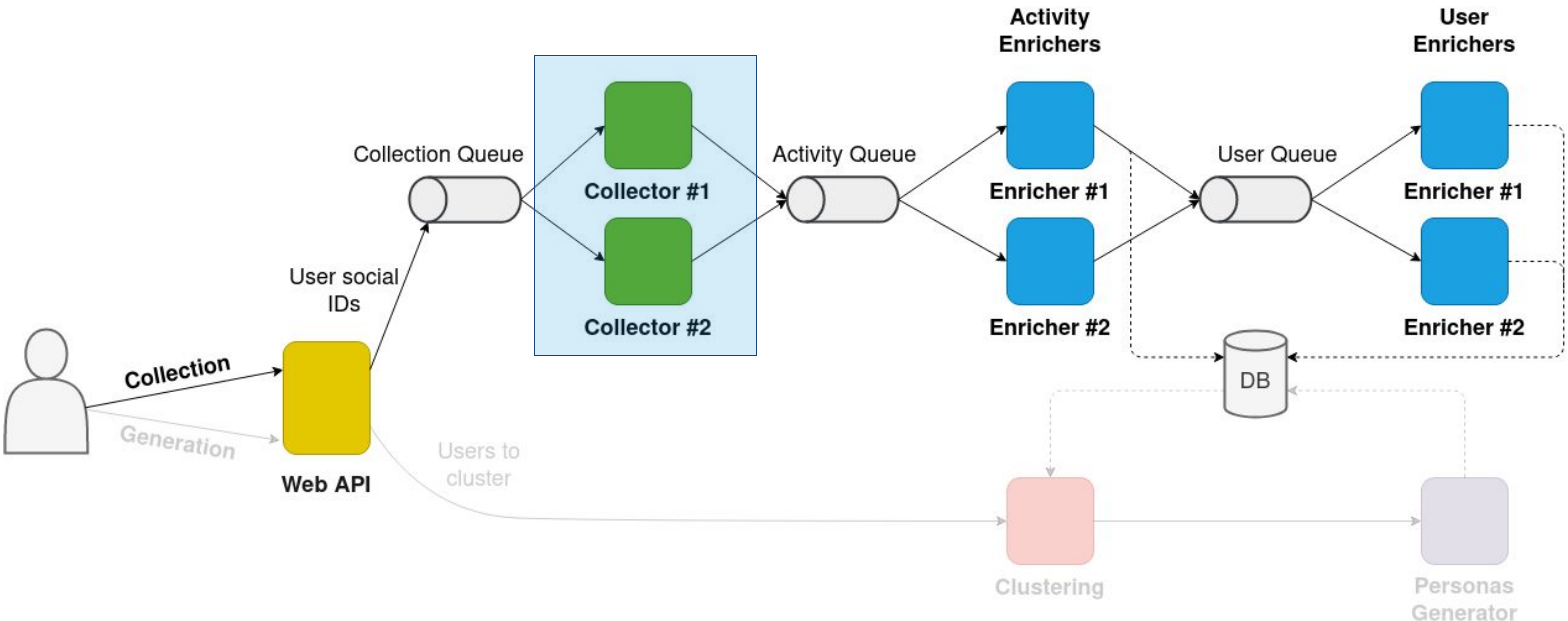
System Architecture



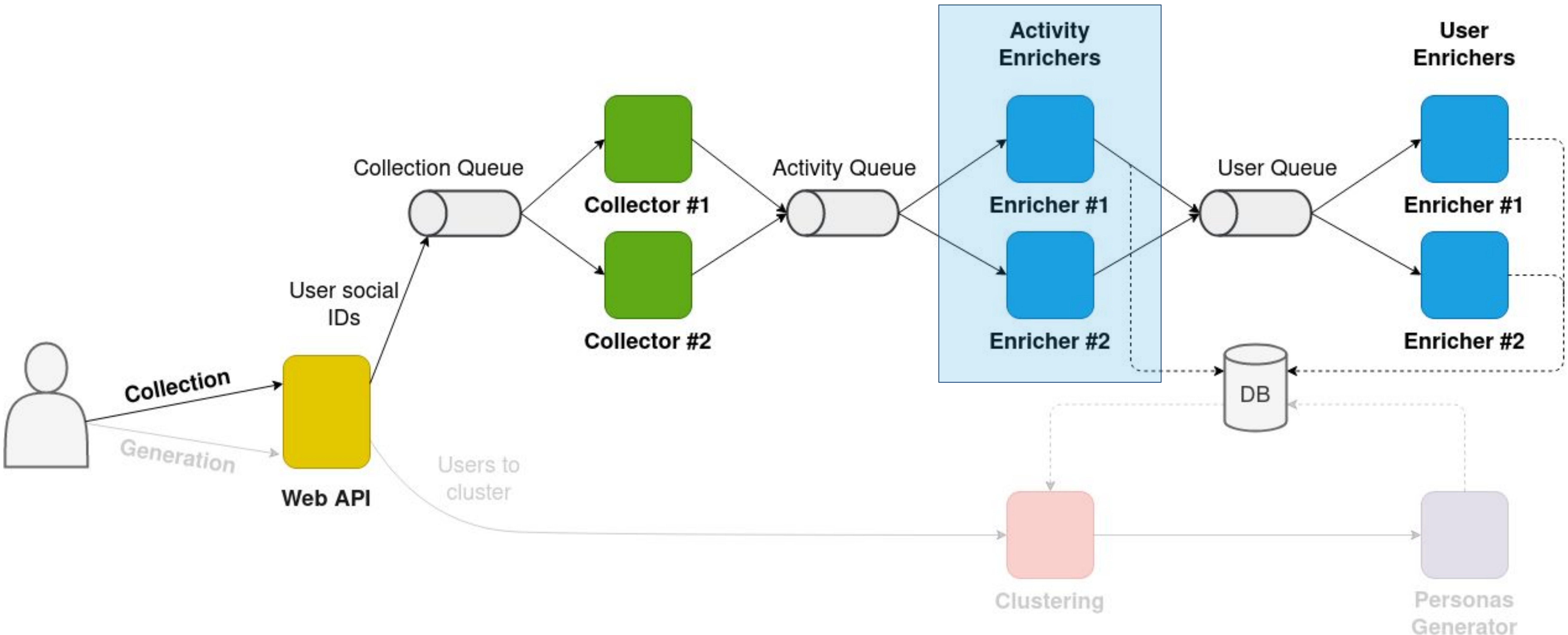
System Architecture



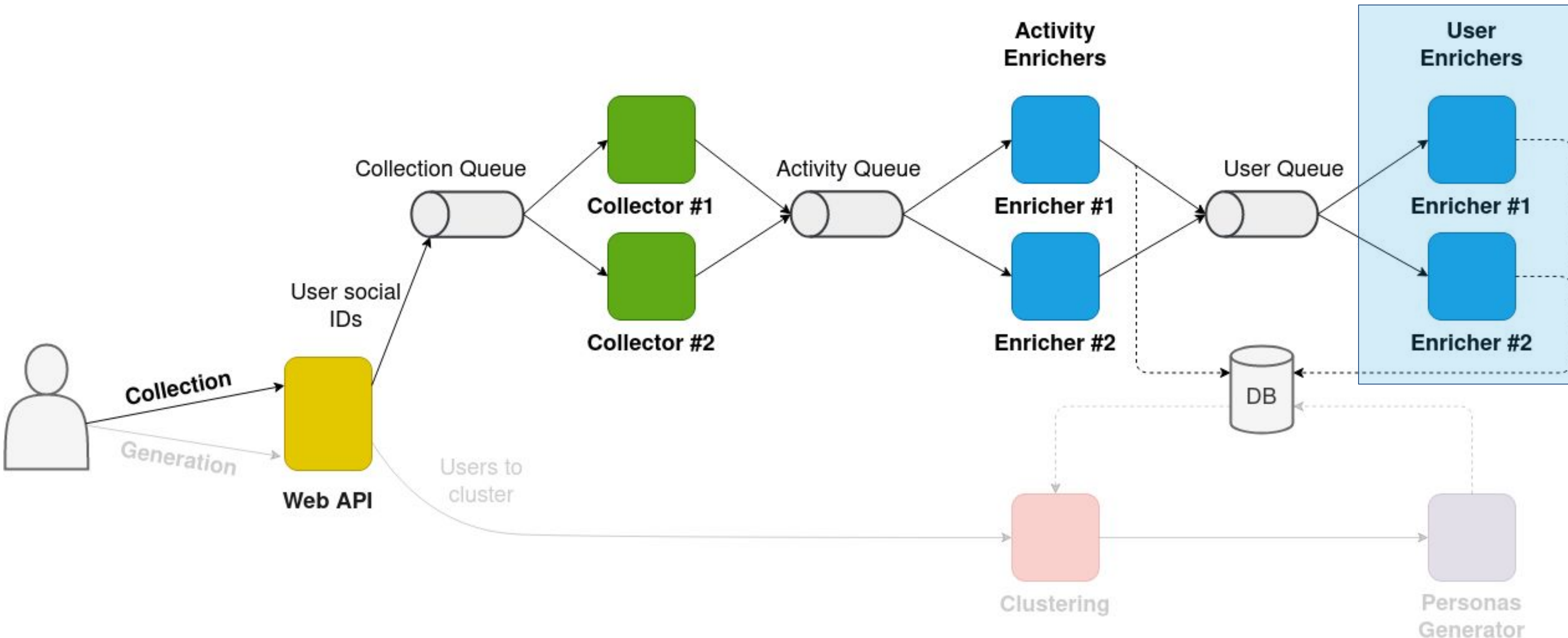
System Architecture



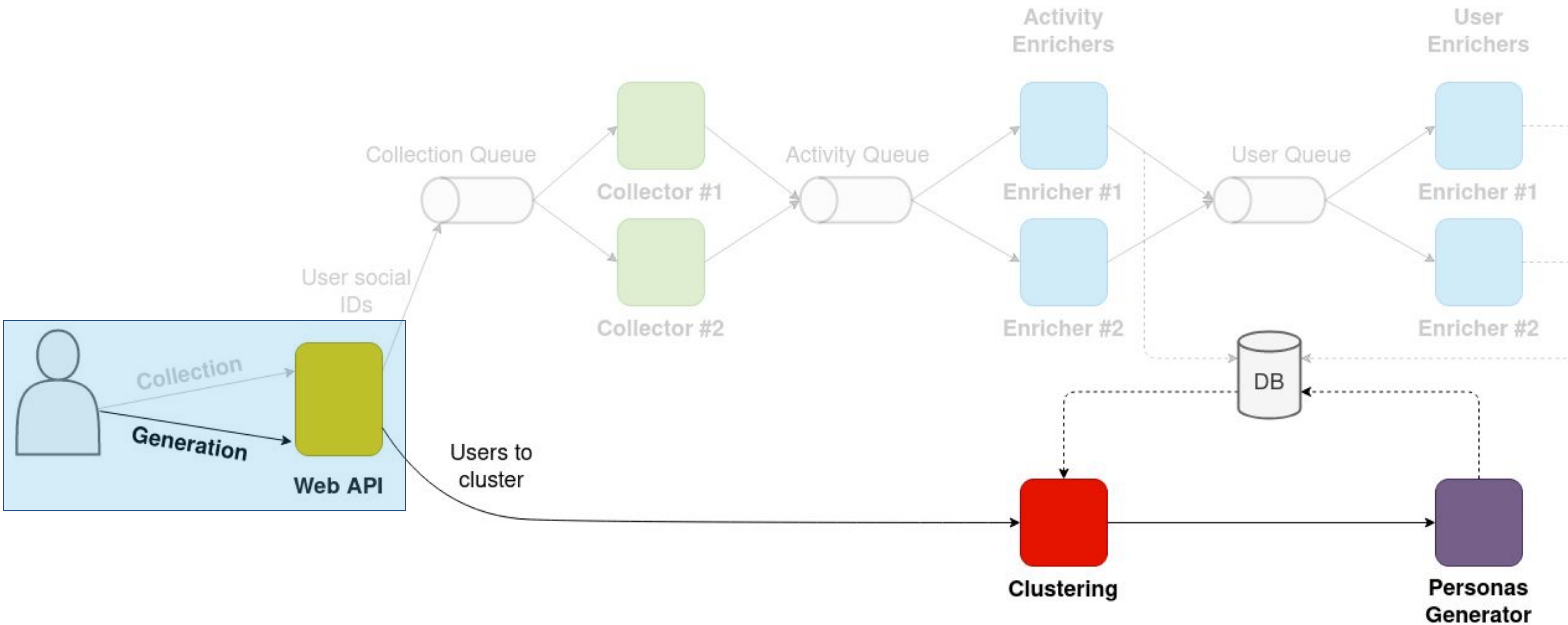
System Architecture



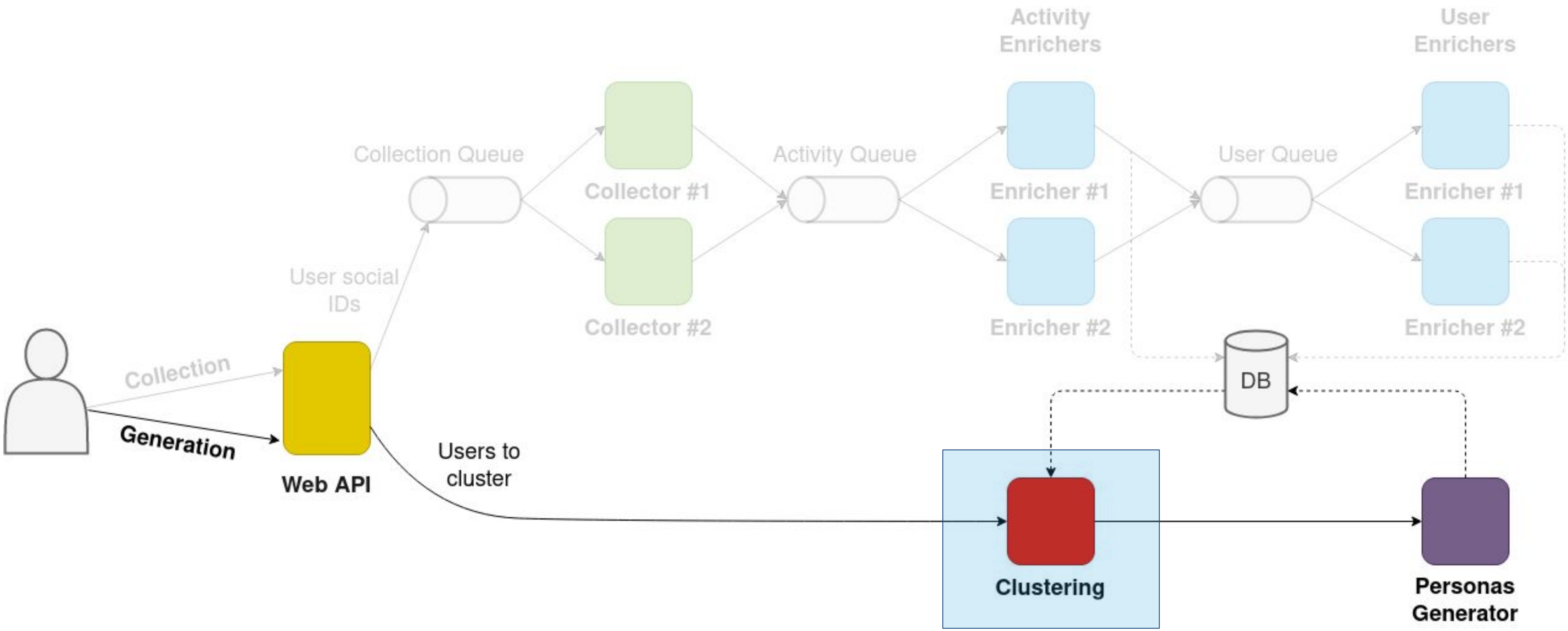
System Architecture



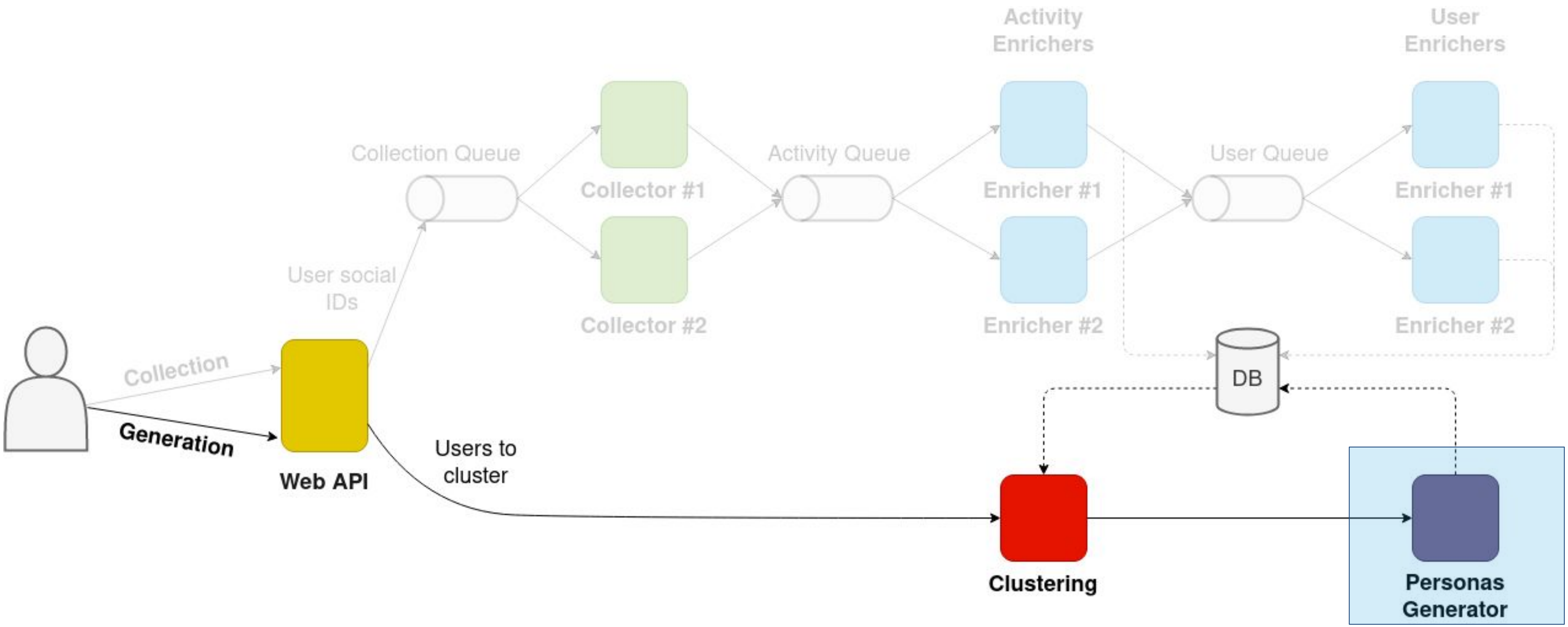
System Architecture



System Architecture

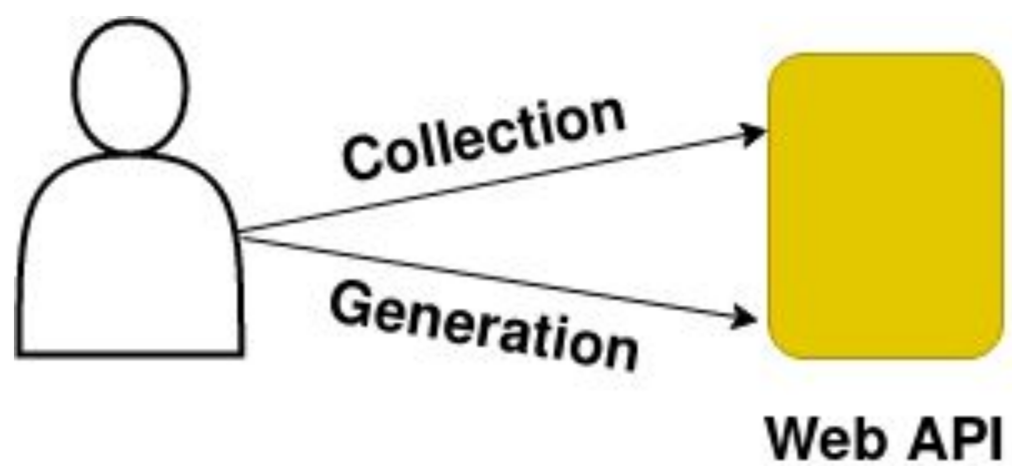


System Architecture

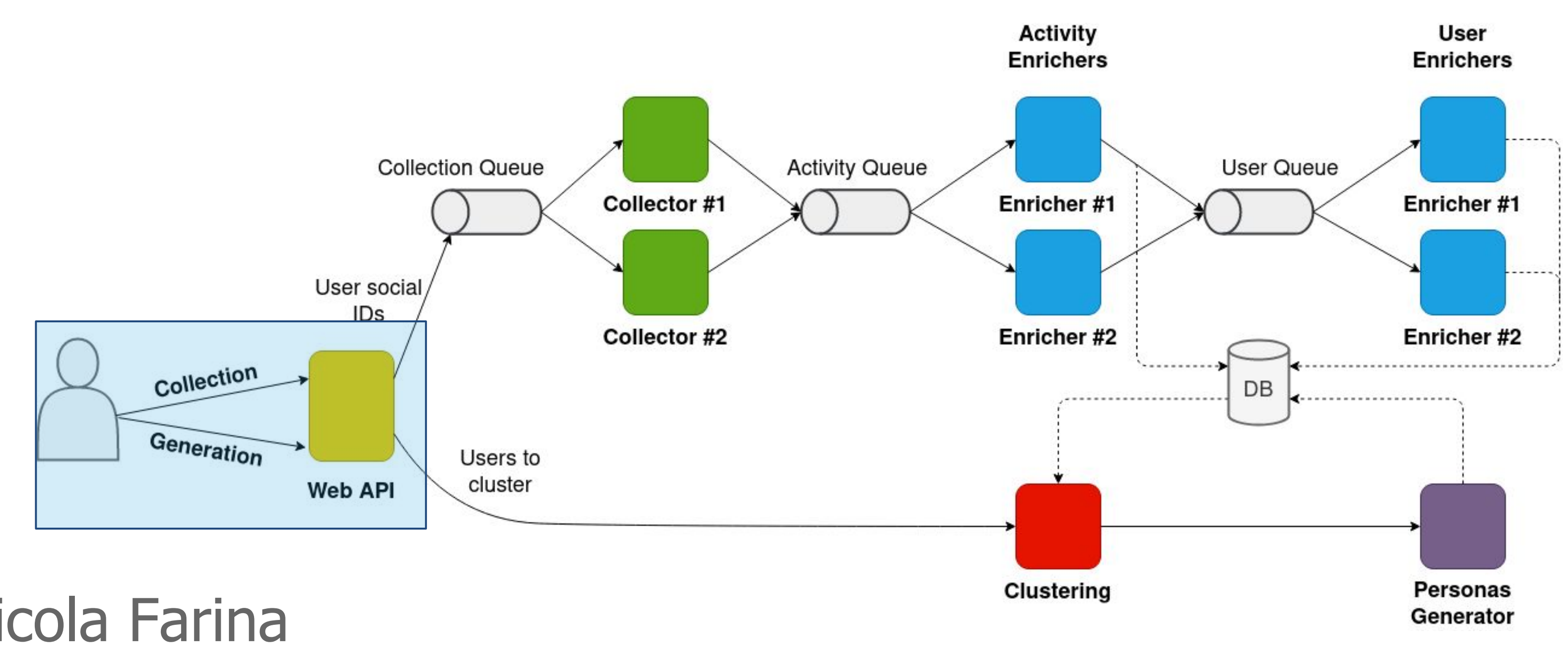


Web API

- RESTful
- Framework: Flask
- Security: JWT



More operations and endpoints are provided (authentication, information retrieval, status checking…)



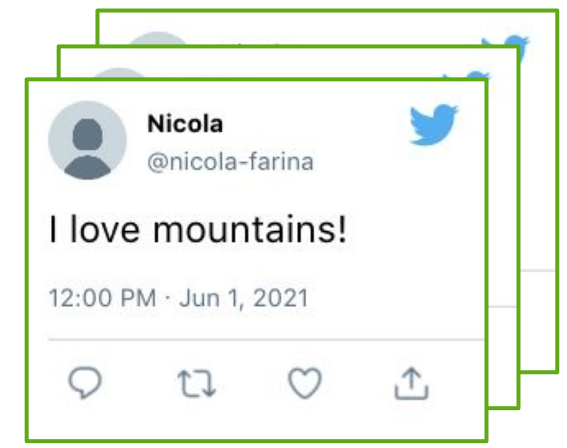
Collection

- Twitter API

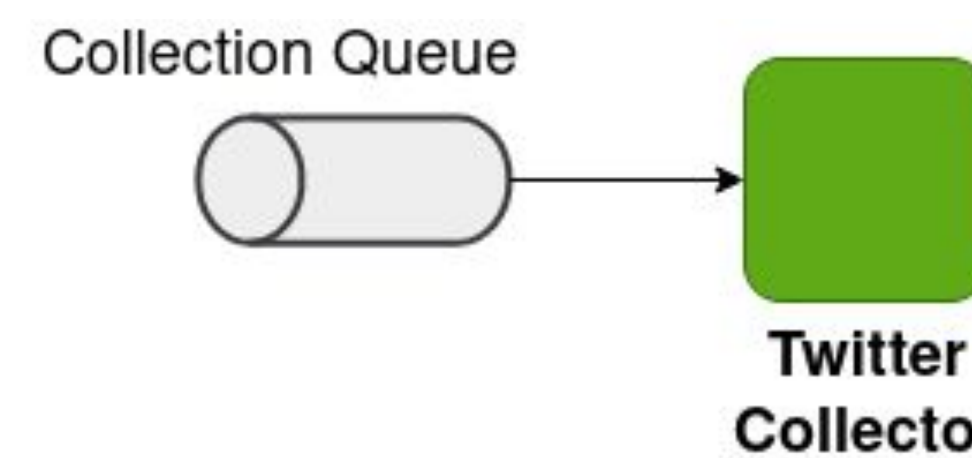
User social ID

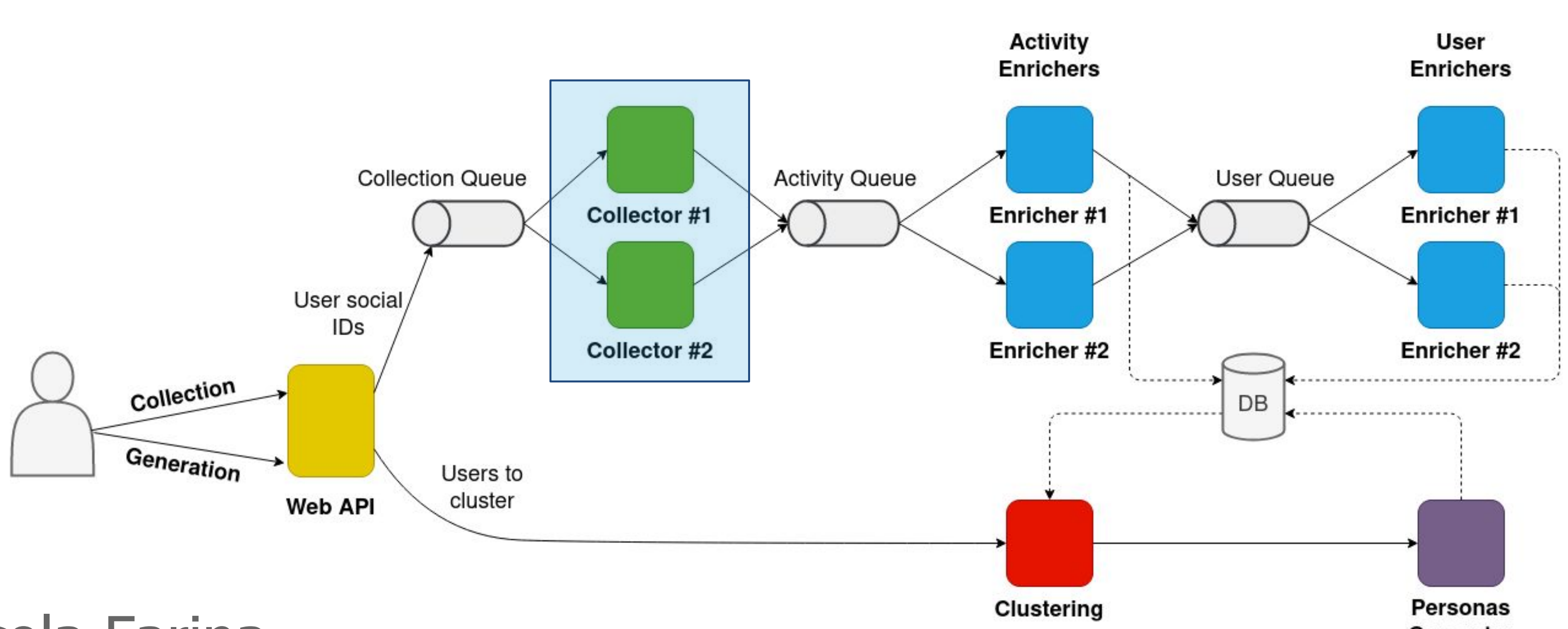
1446281586

Latest tweets



Profile data





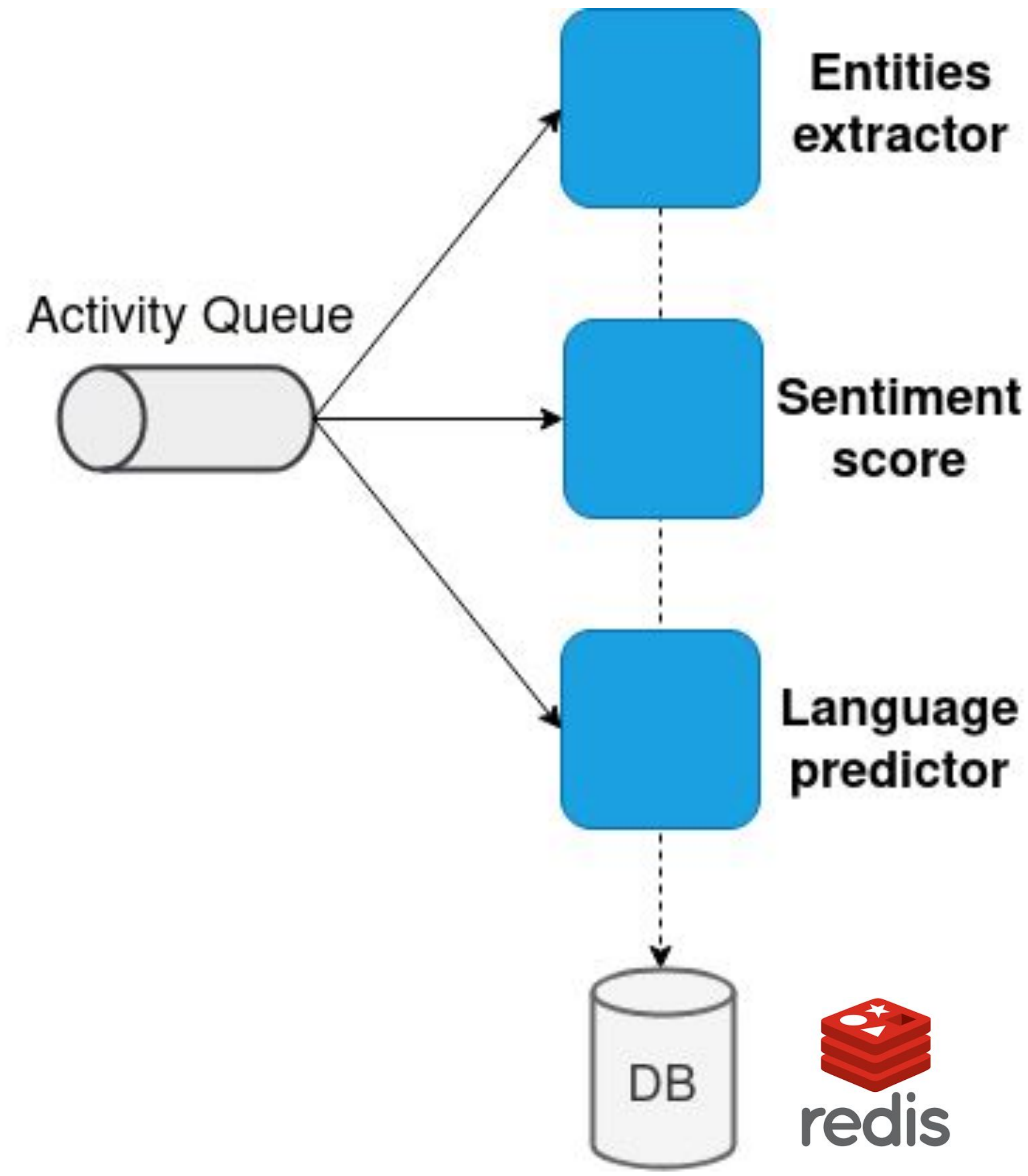
Activity enrichment

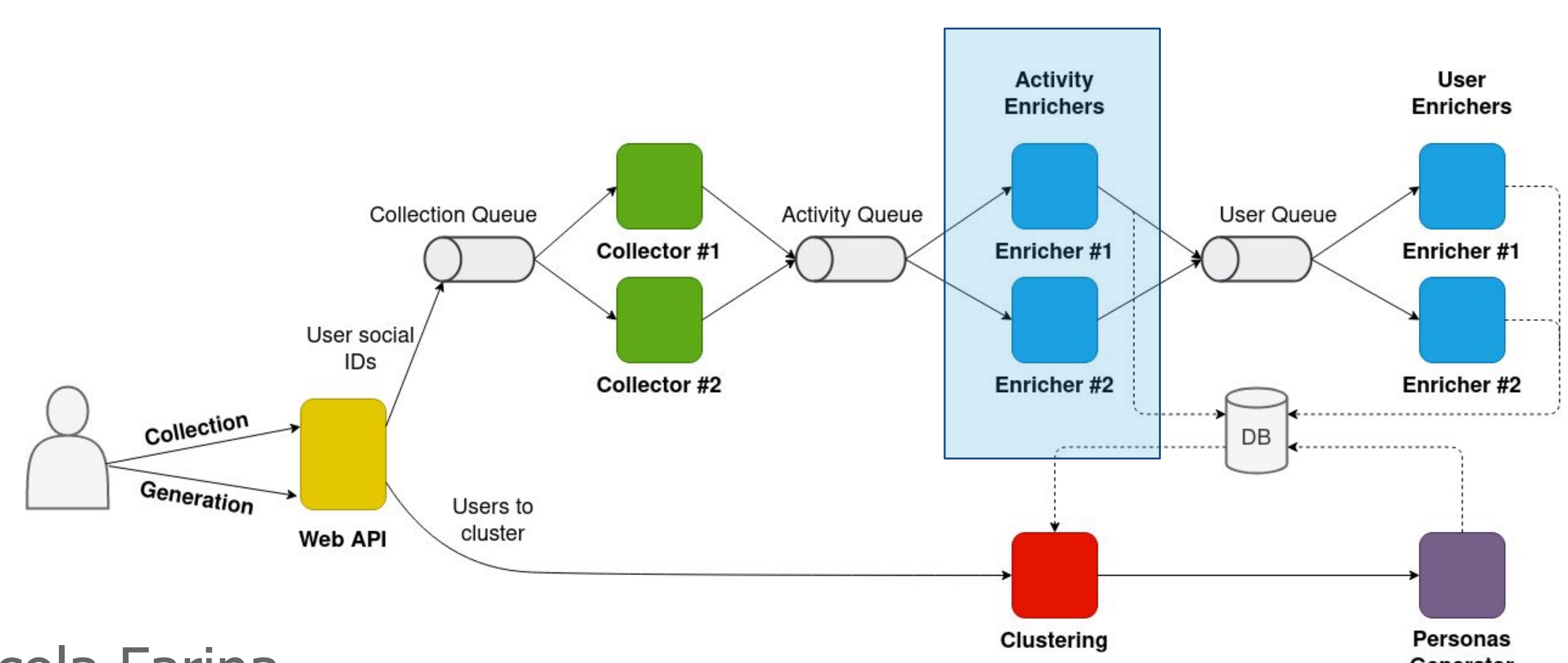
- Dandelion API

Entities: [Mountain]

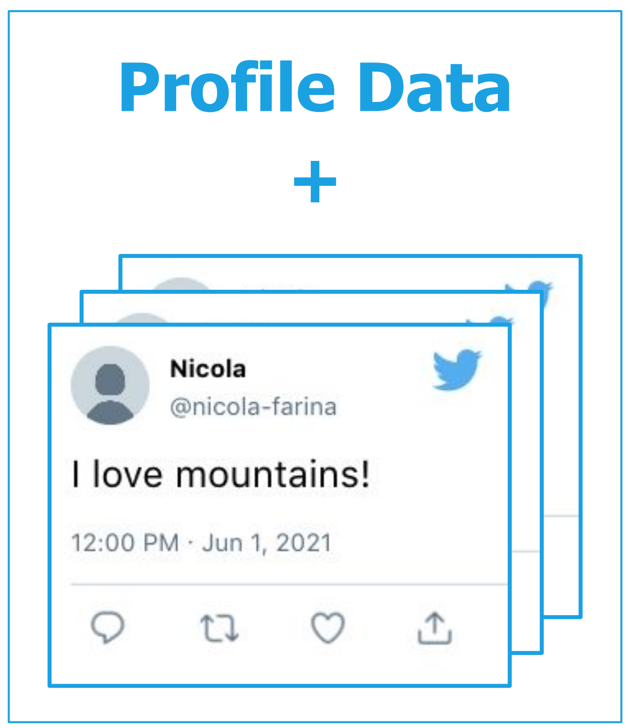
Sentiment: 1.0

Language: en





User enrichment



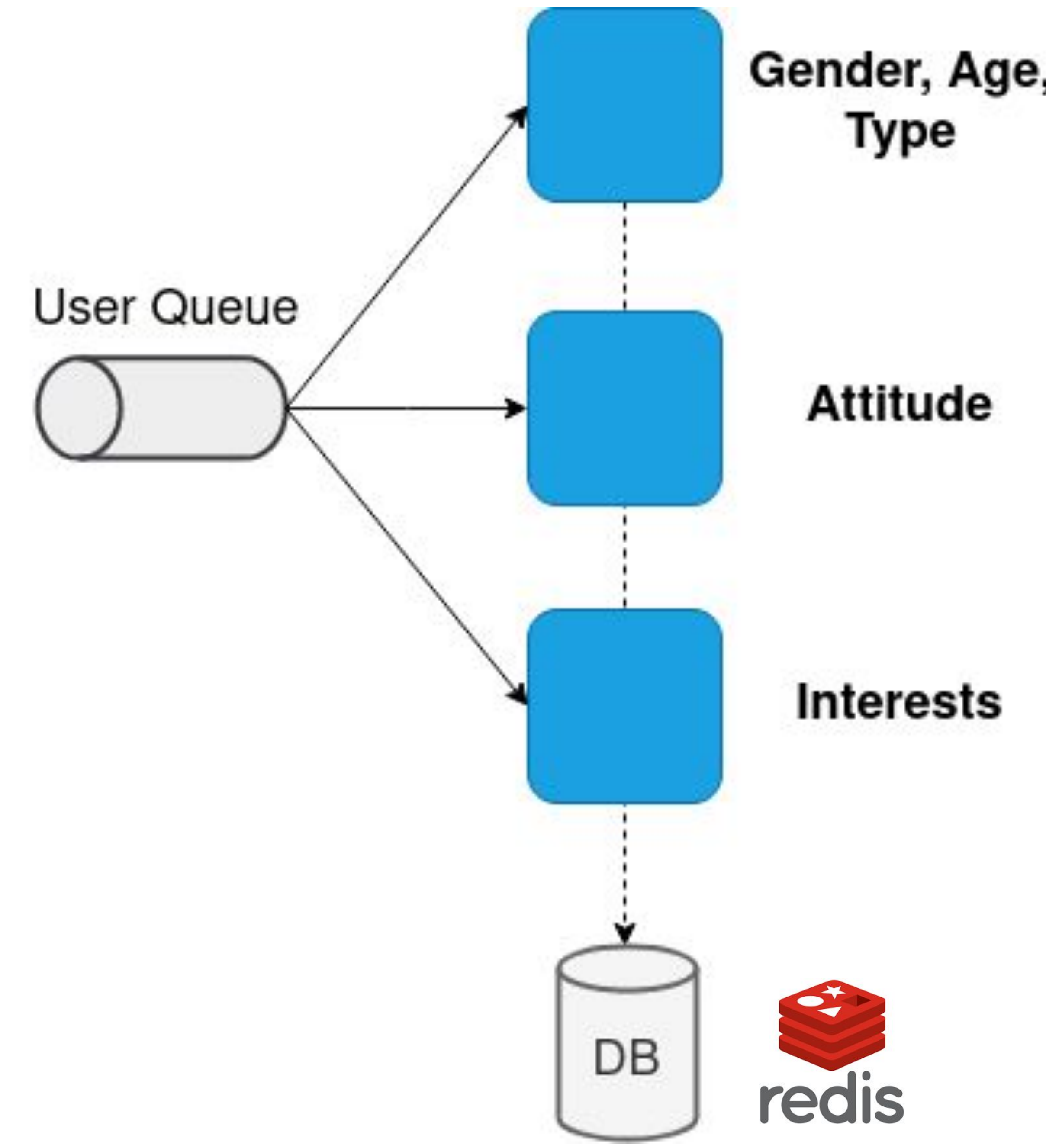
Gender: male

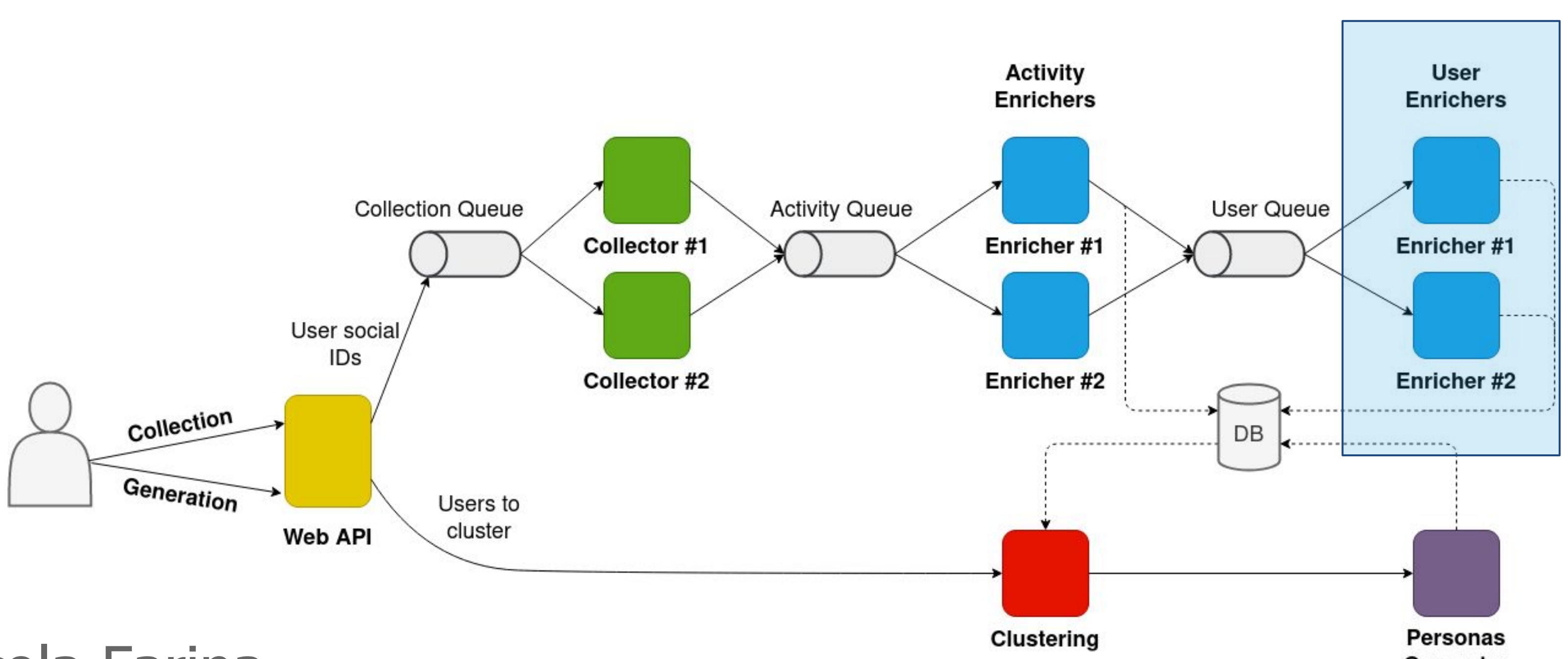
Age: 19-29

Type: person

Attitude: 0.95

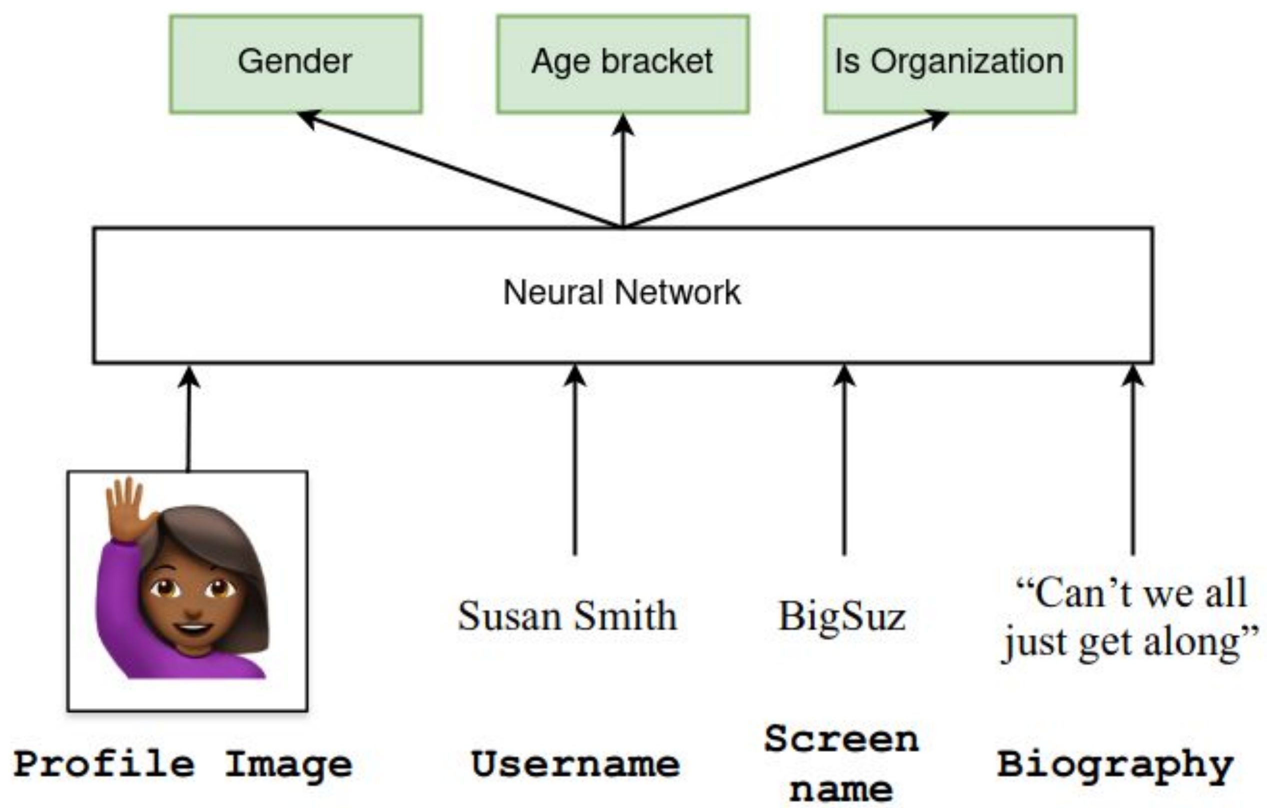
Interests: [Geography: 0.5…]





Gender, age, type, attitude

Gender, Age, Type [2]



Gender (alternative)

Map: First name -> gender

Name	Gender
Marco	M
Stefania	F
Gabriela	F
John	M

Attitude

Average sentiment score

[2] Wang et al., Demographic Inference and Representative Population Estimates from Multilingual Social Media Data

Interests

Football: 6
Basketball: 2
Guitar: 2
Mario Draghi: 1

Entity map

WIKIPEDIA
The Free Encyclopedia

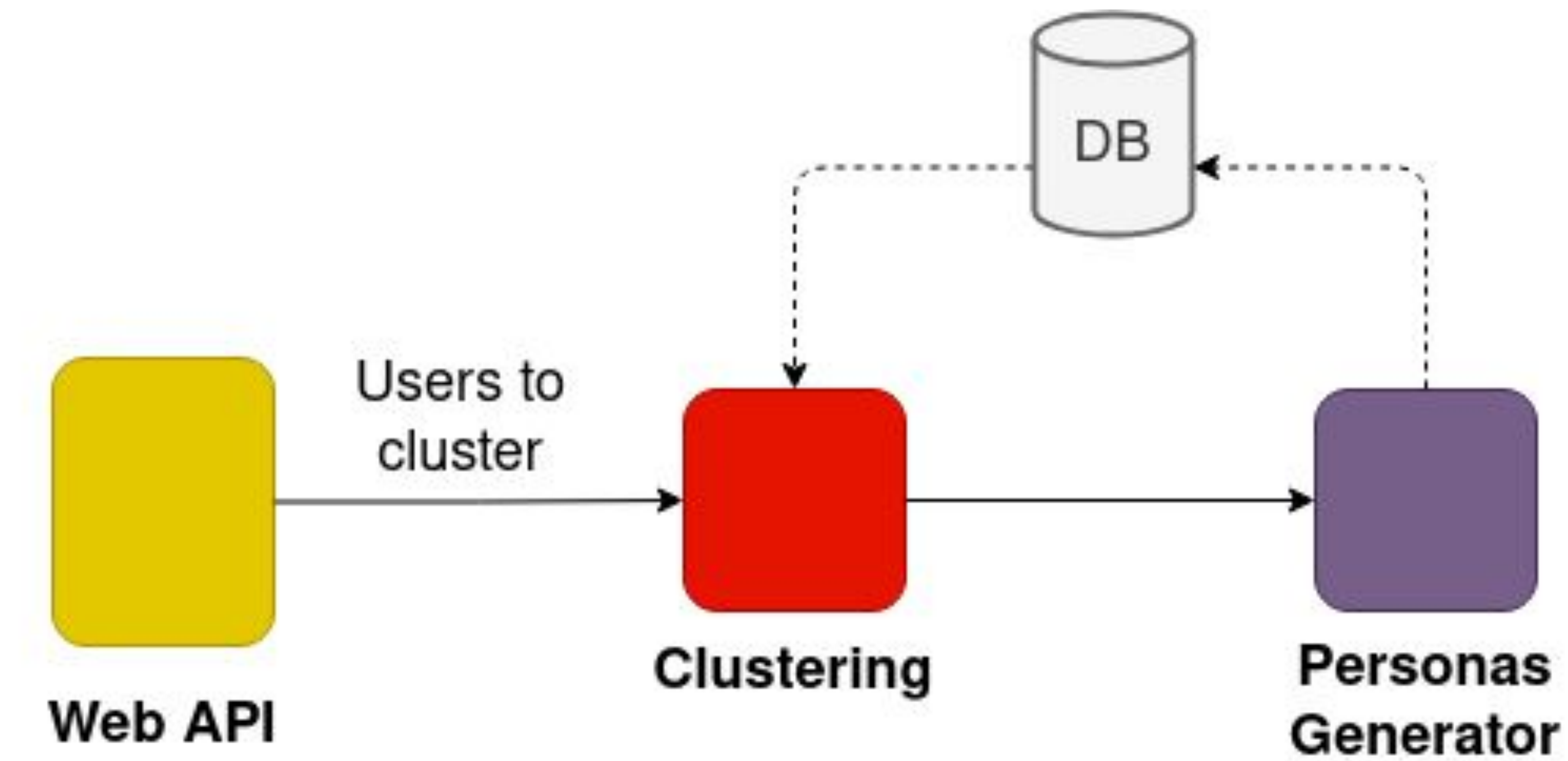
Interests

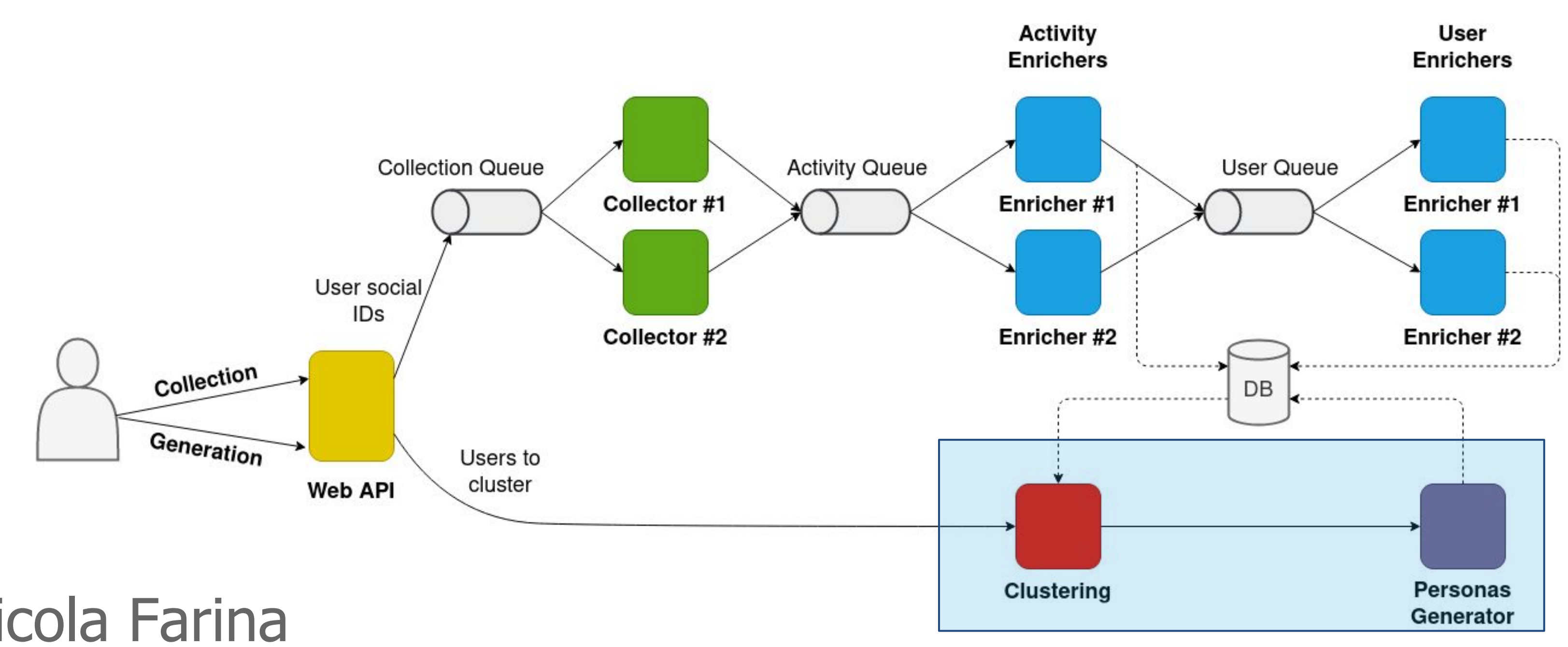
Sports: 0.4
Politics: 0.1
Music: 0.2
....

Clustering and Personas Generation modules

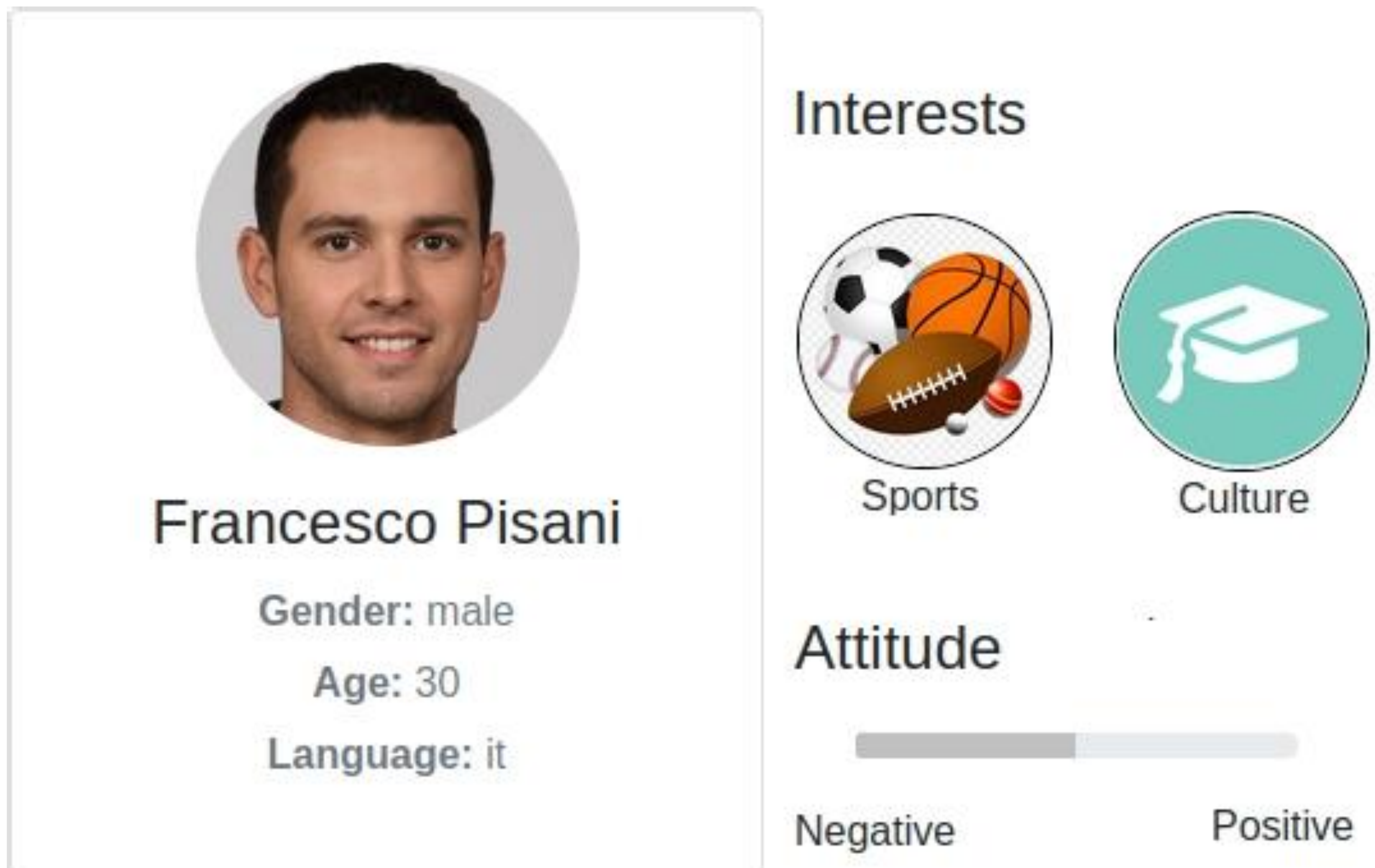
- **Clustering:**
  - K-Modes
  - Custom distance metric
  - Centroid: real user

- **Personas generation:**
  - Assign name and photo to each cluster
  - Results via API or web interface



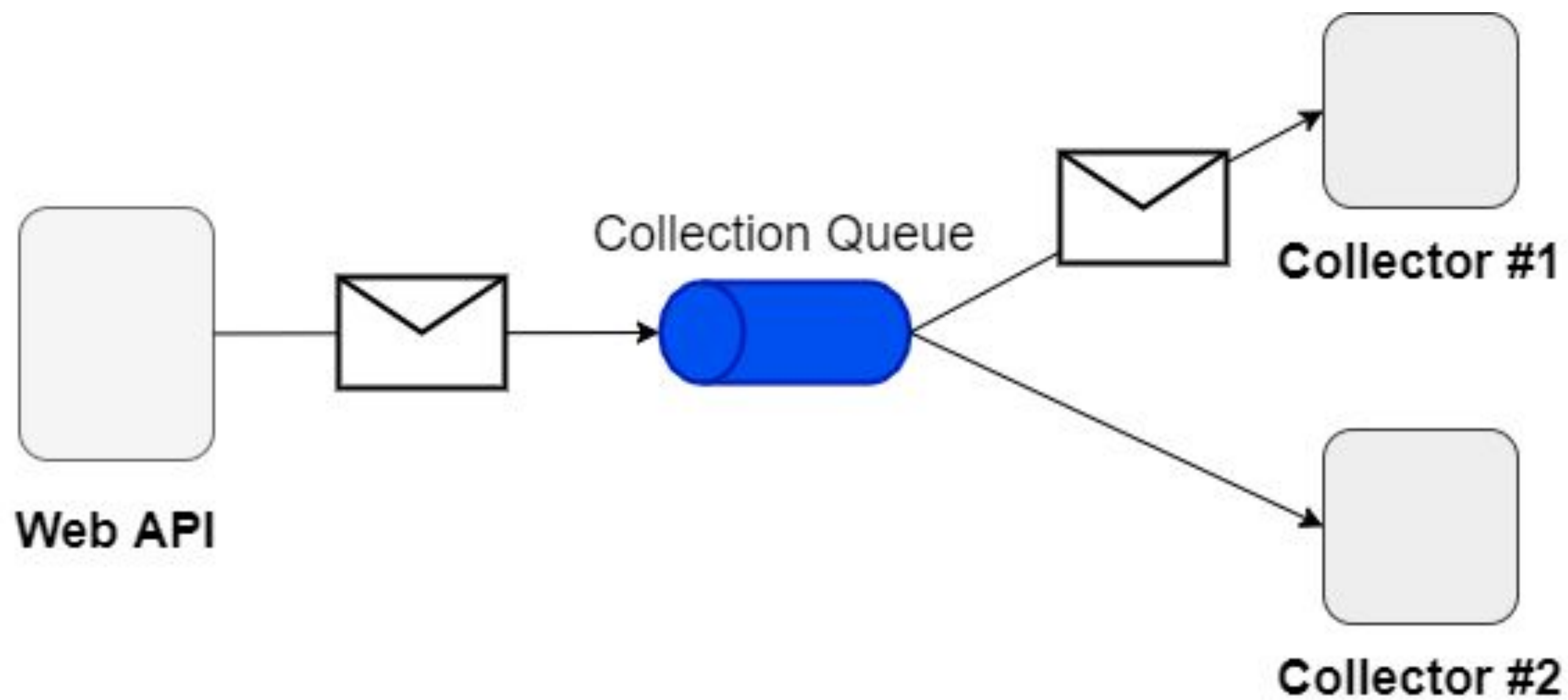


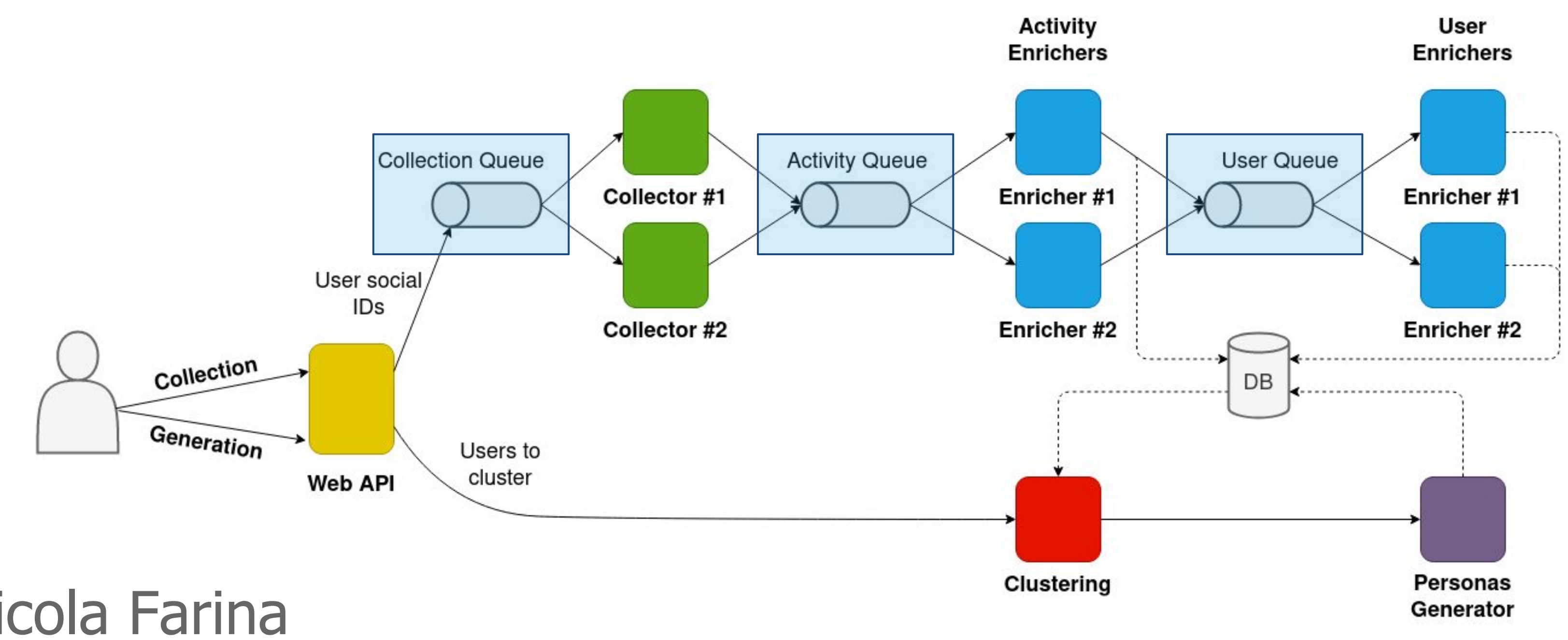
Persona


Francesco Pisani
Gender: male
Age: 30
Language: it
Interests
Sports
Culture
Attitude
Negative
Positive

Queue System

- Communication between modules
- Modularity and scalability





Evaluation: Setup

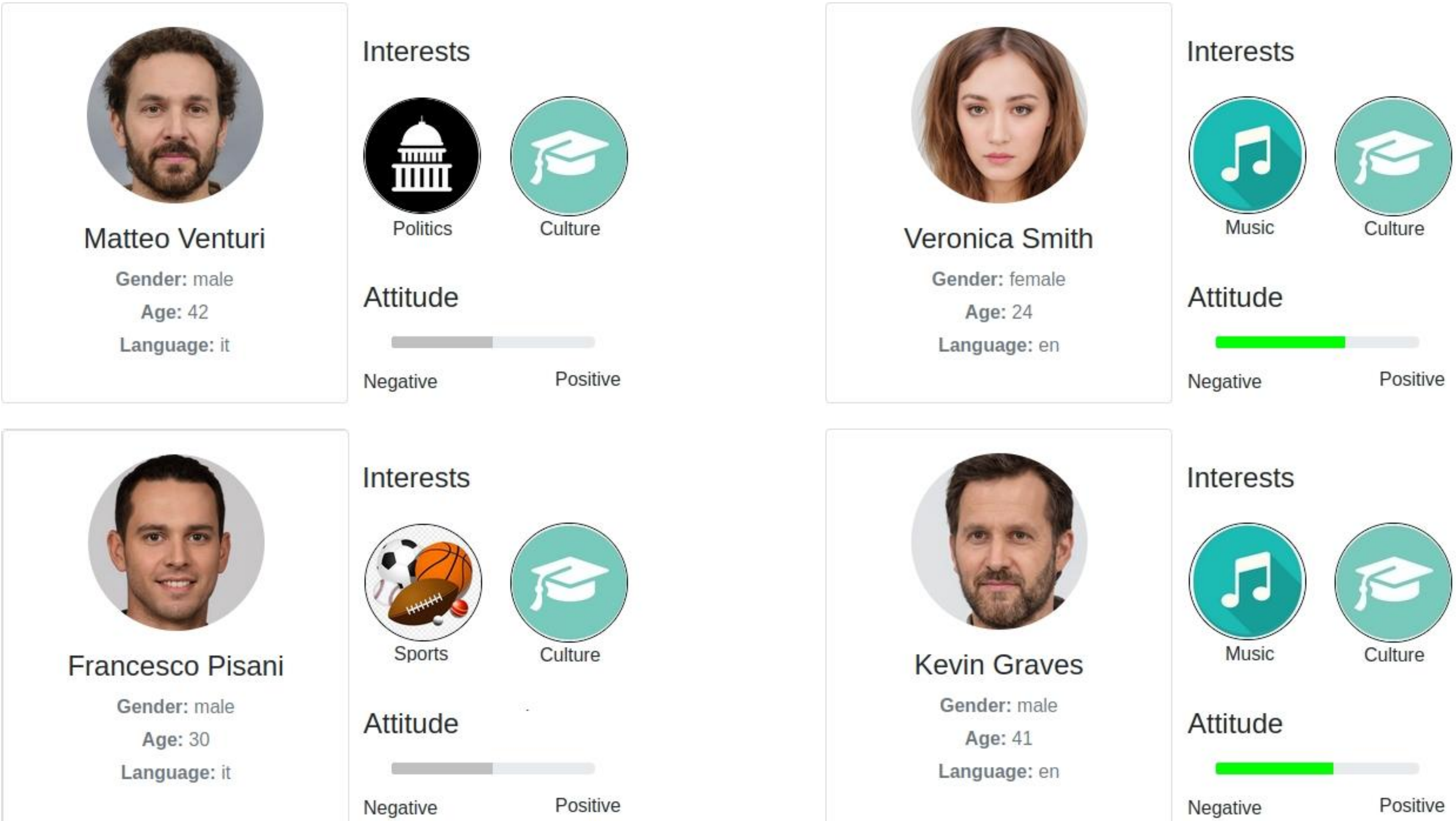
- **Objectives:**
  - Measure quality of personas
  - Tune parameters:
    - number of activities per user
- **Evaluation dataset:**
  - 90 Twitter profiles of **public celebrities**
    - 30 football players
    - 30 musicians
    - 30 politicians
  - **Ground truth:** gender, age, language, main interest

Evaluation: Optimal number of clusters

- **Trade off:** minimum number of activities (per user) for good clusters
  - Too few: **interests misclassification**
  - Too many: **API rate limitations** (4750 activities per day, Dandelion API)

Activities per user	Optimal number of clusters
20	10
50	4
100	4

Evaluation: Personas



Evaluation: Clustering metrics

Cluster	Accuracy	Precision	Recall
Politicians	0.96	0.93	0.96
Musicians (F)	0.97	0.88	1.0
Musicians (M)	0.97	1.0	0.86
Footballers	0.98	1.0	0.96
Global	**0.97**	**0.95**	**0.94**

Conclusions

- System prototype allows to **automatically** generate **accurate** marketing personas
    - Collect users' data from **Twitter** (e.g. followers of a Twitter page)
    - Enrich users with **gender, age, type, attitude, interests**
    - Cluster users, output **representative users** for each cluster
    - Generate **personas,** one for each representative user
- **Expandable** (add/remove classifiers/data sources) due to queues

Future work:

- **Provide web service with GUI**
- Add classifiers and data sources

+: Distance Metric

$$D_{Gower}(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} d_j(x_{1j}, x_{2j})$$

Ordinal features

$$d_{j,ord}(x_1, x_2) = \frac{|rank(x_{1j}) - rank(x_{2j})|}{range_j}$$

Numerical features

$$d_{j,num}(x_1, x_2) = \frac{|x_{1j} - x_{2j}|}{range_j}$$

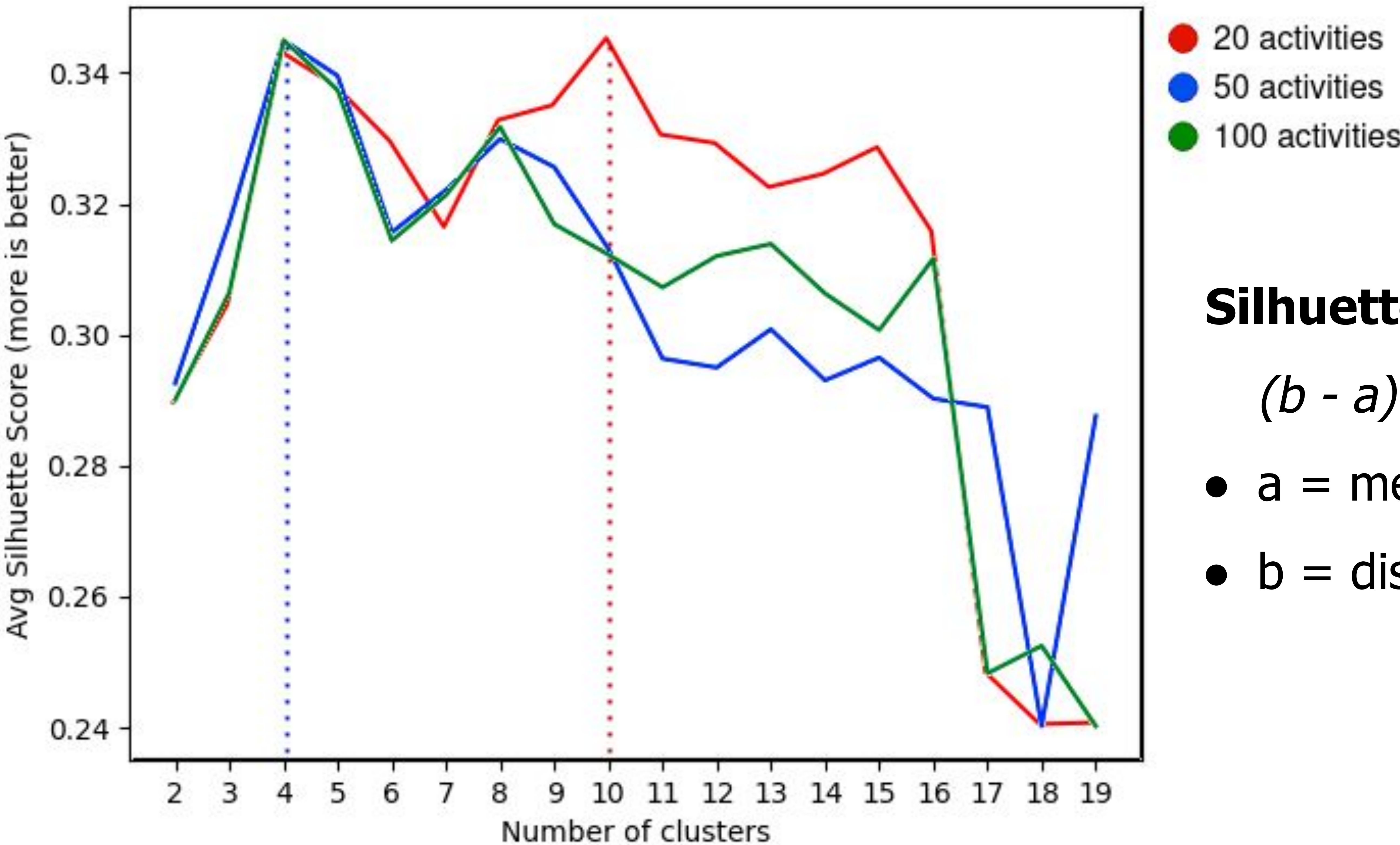
$$d_{j,nom}(x_1, x_2) = \begin{cases} 1 & \text{if } x_{1j} \neq x_{2j} \\ 0 & \text{otherwise} \end{cases}$$

Nominal features

Weights

$$Gender = 0.5 \quad Age = 0.5 \quad Language = 0.3 \quad Interests = 13 \quad Attitude = 0.3$$

+: Optimal number of clusters



Silhuette score for a sample:

(b - a) / max(a, b)

- a = mean intra-cluster distance
- b = distance to nearest cluster

+: Evaluation metrics

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(Fraction of correct predictions)

$$Precision = \frac{TP}{TP + FP}$$

(What portion of positive predictions was actually correct)

$$Recall = \frac{TP}{TP + FN}$$

(What portion of actual positives was correctly predicted)

+: Gender, age, type classificator