UNIVERSITÀ DI TRENTO

Generating Marketing Personas

An automated approach based on extracting insights from social networks

Supervisor
Alberto Montresor
Daniele Miorandi

Student
Thomas Reolon

Degree in Computer Science
Department of Information Engineering and Computer Science

Academic Year 2019/2020

About this Project

External internship at U-Hopper:

Big Data Analytics

Business Intelligence

Chatbot

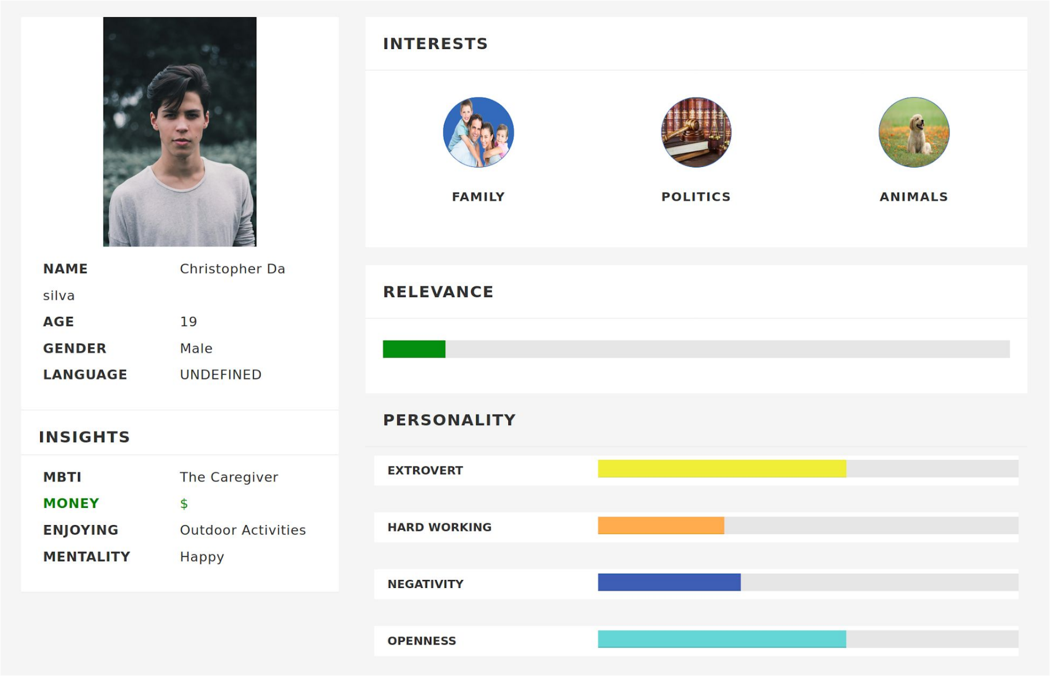
IoT solutions

Artificial Intelligence solutions

Marketing Personas

pros[1]**:**
higher click through rates
boost in ROI
websites' effectiveness

cons:
high costs
high time to develop



[2] Ashley Wilson-Rew, 12 Statistics That Prove the Value of Creating Buyer Personas, 2015

Purpose

is it possible to generate personas using social media data?

develop a prototype capable of:

1. Extracting **insights** from users' activities
2. **Clustering** the insights
3. Generating **personas** using the clusters
4. Respecting the requirements: **modularity** & **scalability**

State of the Art

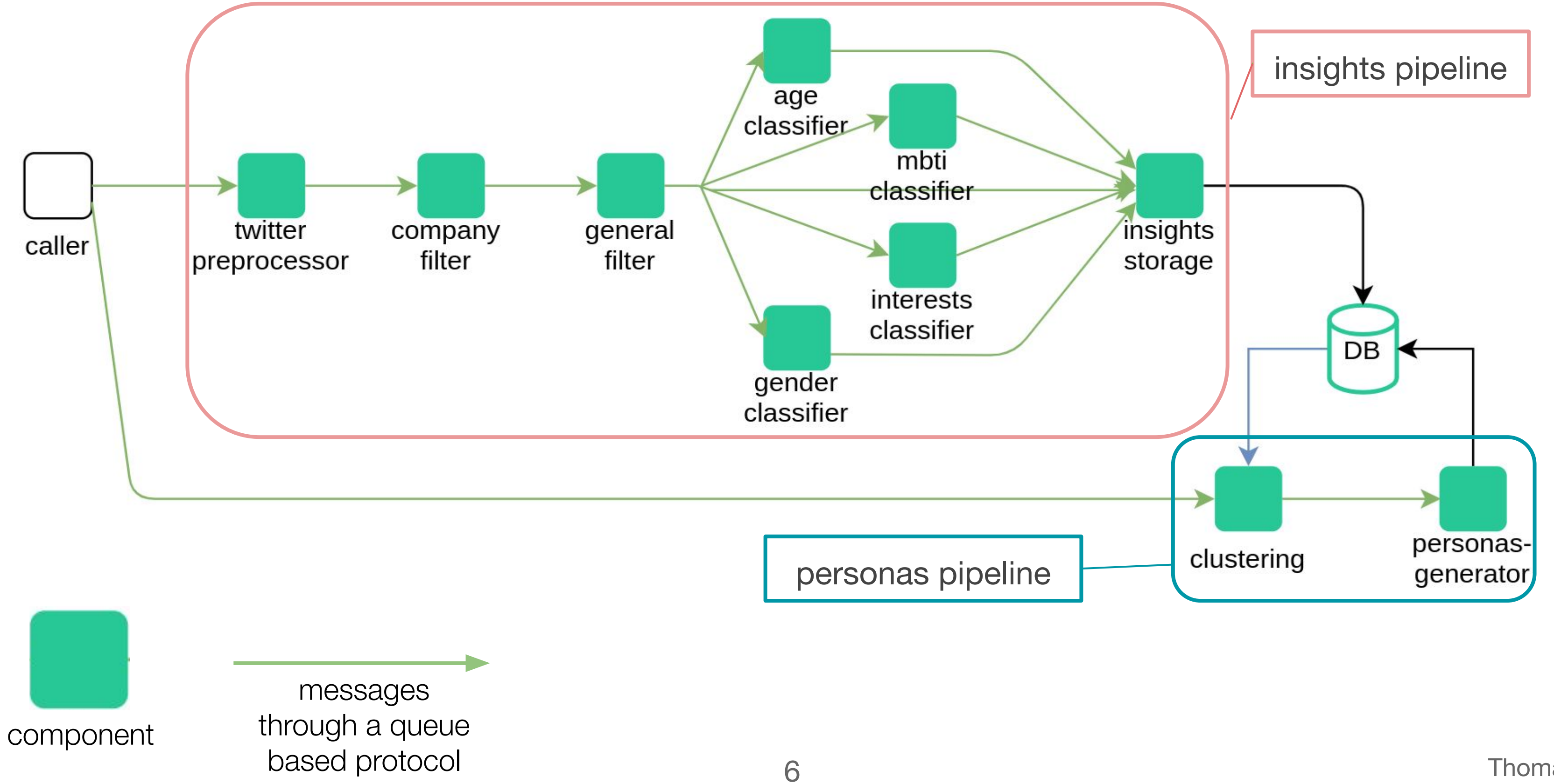
Koponen[2]: generating marketing personas **clustering** users' behaviours collected from a **website**

age
gender
personality
interests

wide literature about extracting insights from social networks

[2] M. Koponen, Developing marketing personas with machine learning for educational program finder, 2017

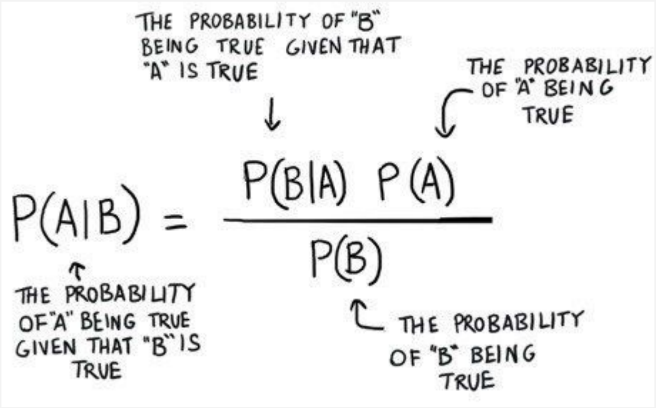
Solution


insights pipeline
caller
twitter preprocessor
company filter
general filter
age classifier
mbti classifier
interests classifier
gender classifier
insights storage
DB
clustering
personas-generator
personas pipeline
component
messages through a queue based protocol

Components: insights

- Preprocessors:
  - Twitter preprocessor
- Filters:
  - Company filter
  - General filter
- Classifiers:
  - **MBTI personality**
  - OCEAN personality [IBM]
  - **Interests texts**
  - **Interests pictures**
  - Interests Watson [IBM]
  - **Age**
  - **Gender**

Personality & Age & Gender Classifiers



Naive Bayes Classifier

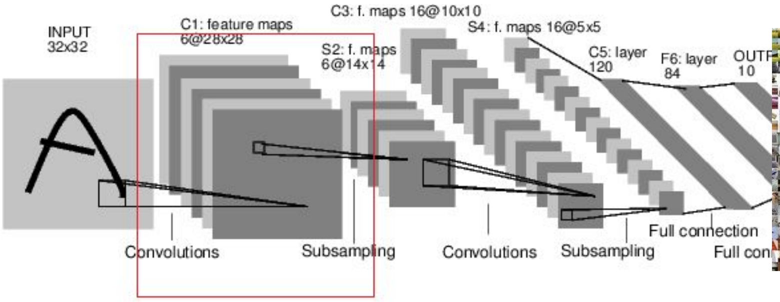
Pretrained Wide ResNet [3]

	name	gender
1	Emma	F
2	Olivia	F
3	Noah	M
4	Liam	M
5	Sophia	F

Map: first_name → gender

[3] Yusuke Uchida, Age Estimation in Pytorch, 2019

Interests Classifier



object detection

tapoi

WIKIPEDIA
The Free Encyclopedia

Whoosh

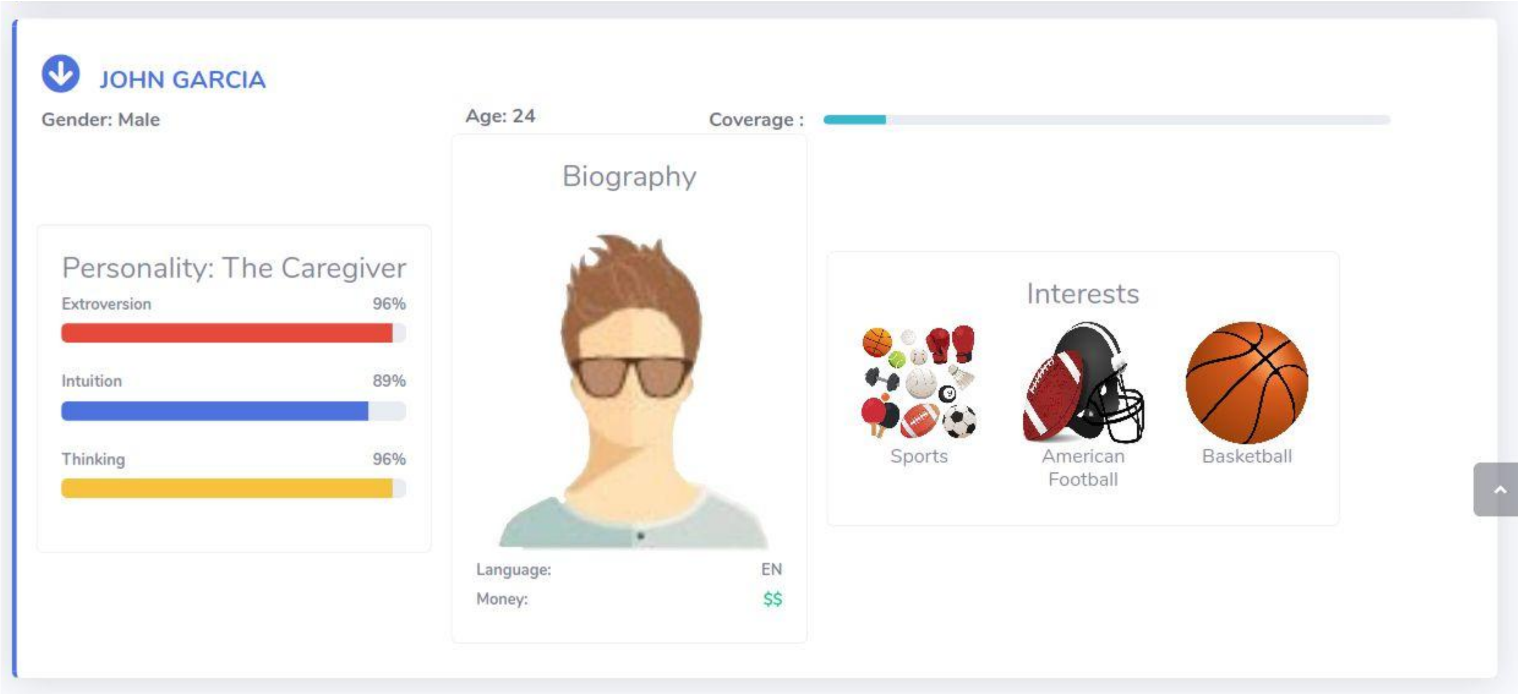
NLU

Components: Personas

- Clustering
  - MIMOSA
  - General
  - Hierarchical
- Presentation
  - **personas generator**

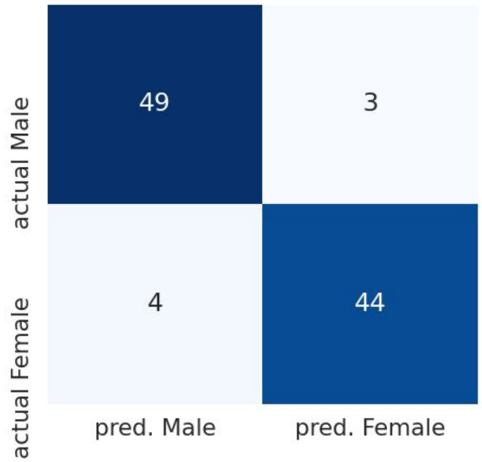
There are other types of components, but they are not fundamental (logger, collector, …)

Personas Generator



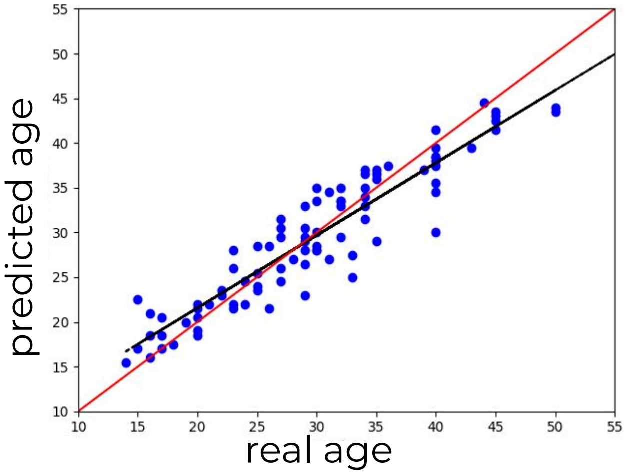
Evaluation: insights

GENDER



prob. err = 0.07

AGE



avg. error = 3 years

PERSONALITY

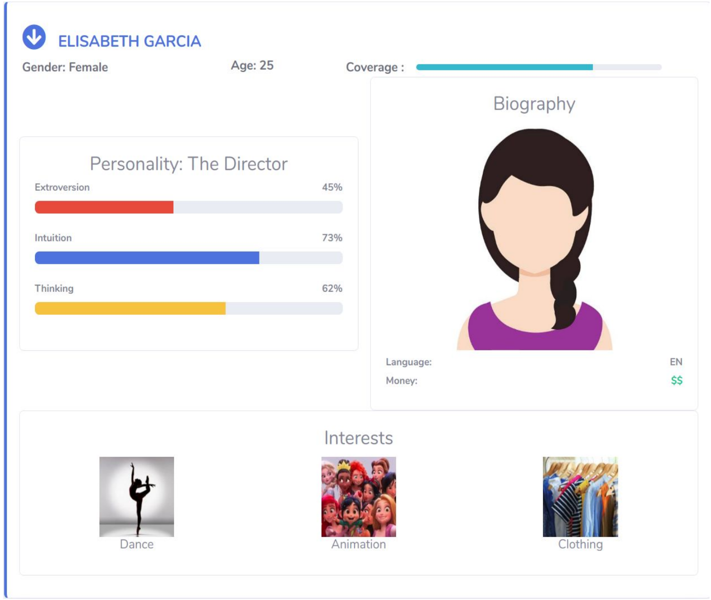
MBTI class	Precision	Recall
Thinking	0,78	0,74
Sensing	0,74	0,80
Extraversion	0,70	0,83
Judging	0,65	0,85

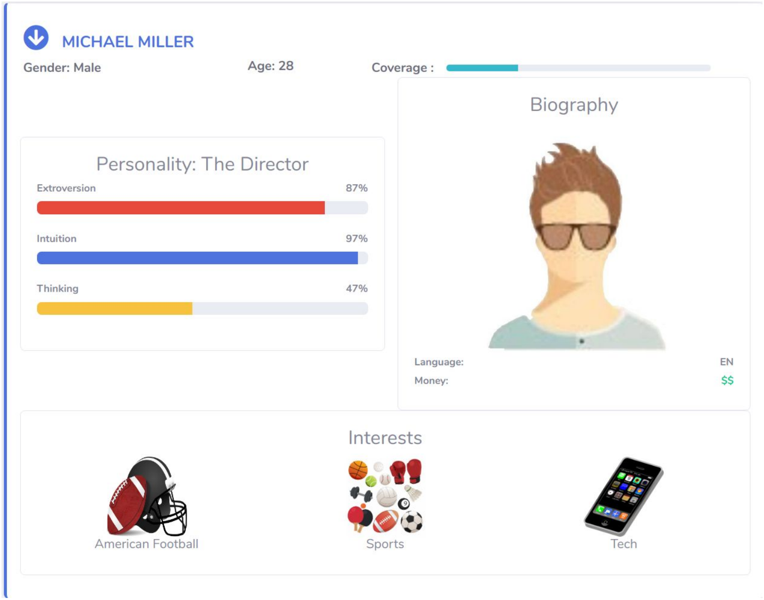
avg. prob. err = 0.28

Evaluation: personas

ground truth: Twitter page about Makeup — Twitter page about American Football

prediction:





Summary

- Distributed system (scalability)
- Independent processes (modularity)
- Python & docker (portability)
- ML techniques to process data
- Wrapped by a web-service (collector with restful API, personas presented as html)

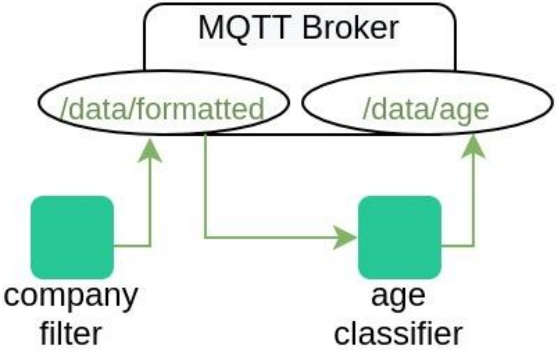
Future Work

New classifiers:	job, salary, favourite brands, preferred channels
Upgrade classifiers:	interests
Increase scalability:	substitute Mosquitto (MQTT) with Apache Kafka

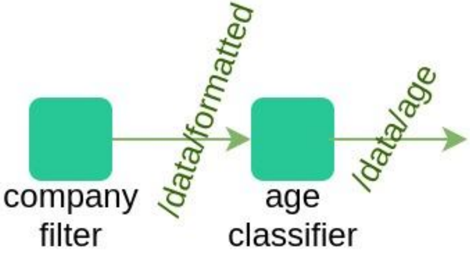
Thank you

+: the **message passing** protocol

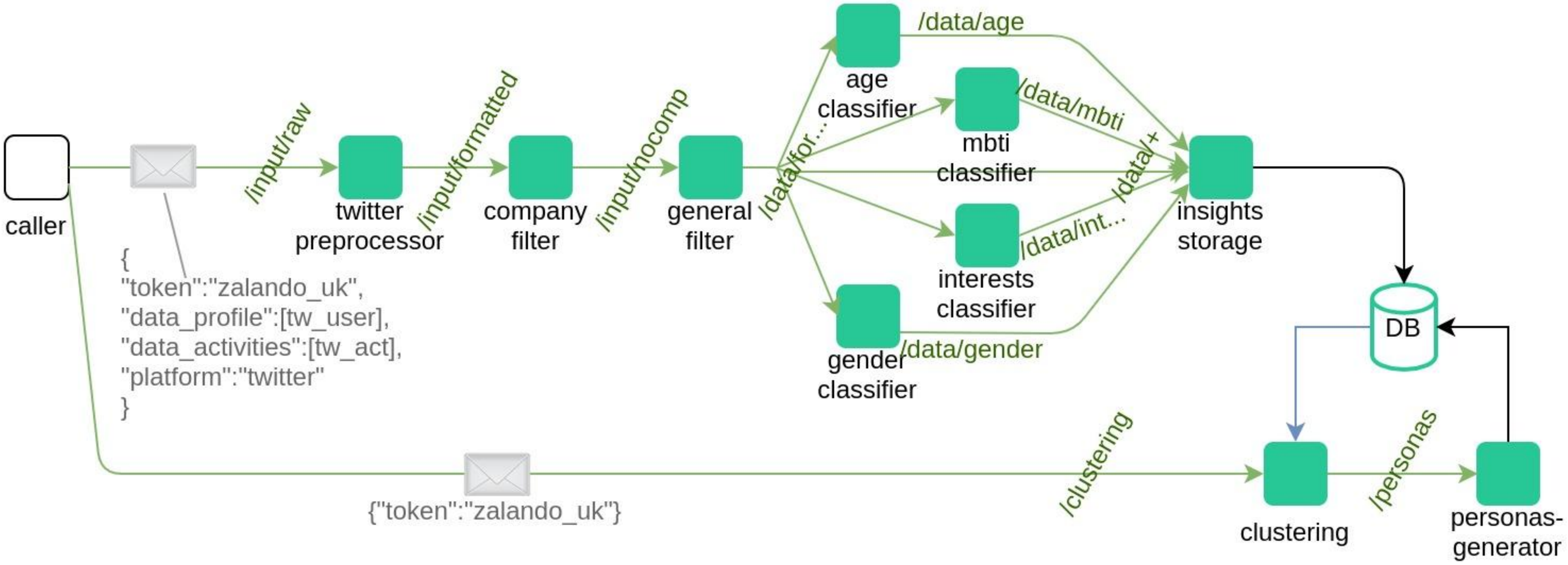
long way to represent how the components communicate



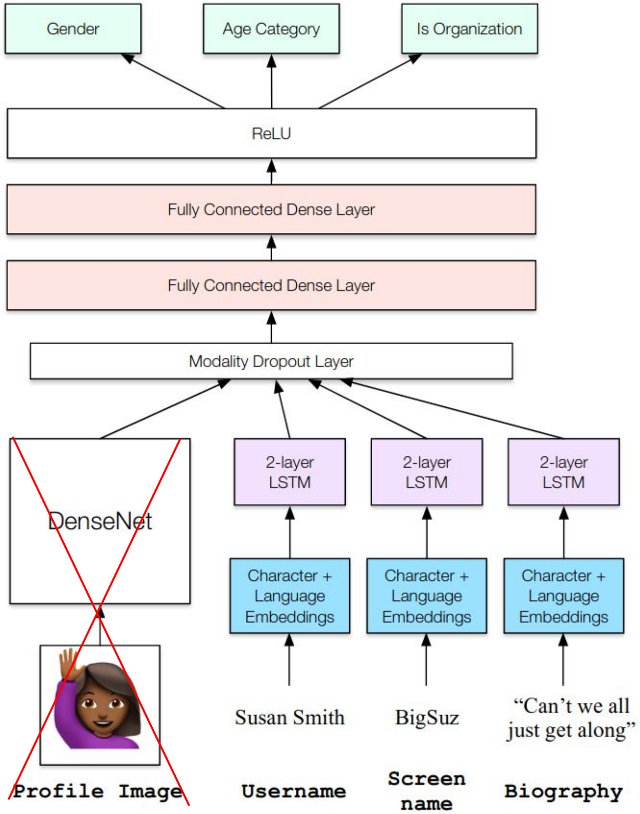
a more compact way to show the same communication



+: Solution


caller
{
"token":"zalando_uk",
"data_profile":[tw_user],
"data_activities":[tw_act],
"platform":"twitter"
}
/input/raw
twitter preprocessor
/input/formatted
company filter
/input/nocomp
general filter
/data/for...
age classifier
/data/age
mbti classifier
/data/mbti
interests classifier
/data/int...
gender classifier
/data/gender
/data/+
insights storage
DB
{"token":"zalando_uk"}
/clustering
clustering
/personas
personas-generator

+: Company Filter



Zijian Wang et al.

[Demographic Inference and Representative Population Estimates from Multilingual Social Media Data]()

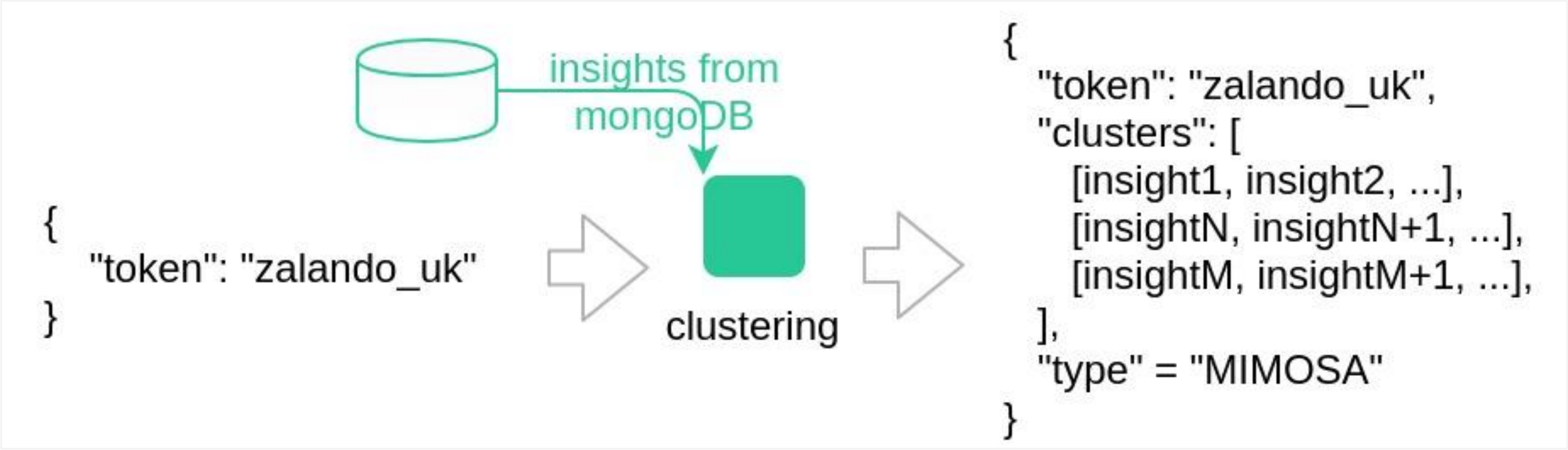
+:Insights Storage


```
{
  "_id" : NumberLong(1869560826726337887),
  "img_topics" : {        computed by: obj-detect-interests
    "Animals" : 0.0,
    "Music" : 0.0,
    "Health Problems" : 0.0,
    "War" : 0.0,
    "Clothing" : 0.0,
    "Sports" : 0.0277777777777778,
    "Drinks" : 0.0,
    "Rich" : 0.0,
    "Cosplay" : 0.0,
    "Office" : 0.0,
    "Travel" : 0.0,
    "Family" : 0.0,
    "Baby" : 0.0
  },
  "interests" : {        computed by: NLU-interests
    "Tech" : 0.0742912752107073,
    "Academics" : 0.0,
    "Animation" : 0.0793321633762743,
    "Tennis" : 0.097546949087905,
    "Biking" : 0.0,
    "Baseball" : 0.0946241806448701,
    "American Football" : 0.0890734428972284,
    "Soccer" : 0.0,
    "Dance" : 0.0,
    "Activism" : 0.0,
    "Nature" : 0.0,
    "Animals" : 0.0,
    "Food" : 0.0742286485371892,
    "Basketball" : 0.0986046997150346,
    "Religion" : 0.0,
    "Cars" : 0.0790913512389902
```

```
  "token" : "theffballers",
  "latest_activity" : "2020-05-27 16:12:40",
  "platform" : "twitter",

  "age" : 39,        computed by: company-filter
                                  + age-classifier

  "gender" : 0.995,        computed by: company-filter
                                        + gender-classifier

  "mbti" : {        computed by: mbti-classifier
    "T" : 0,
    "S" : 1,
    "E" : 1,
    "J" : 1
  },

  "ocean" : null,
  "needs" : null,        dismissed components: IMB API
  "nlu" : null,

  "language" : {        computed by: twitter API
                                     + pycld2
    "it" : 0,
    "en" : 1,
    "fr" : 0,
    "de" : 0,
    "sp" : 0,
    "un" : 0
  }
}
```


Thomas Reolon

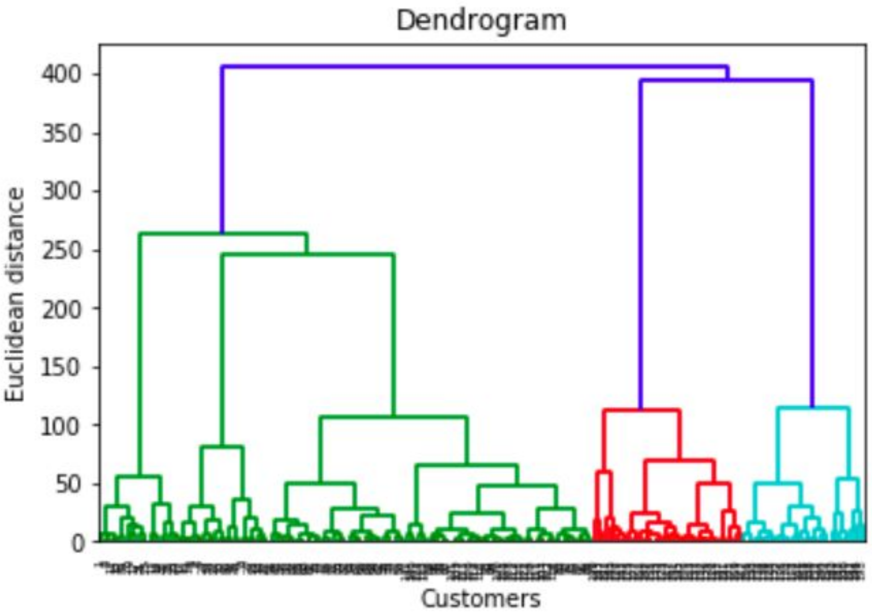
+: Clustering

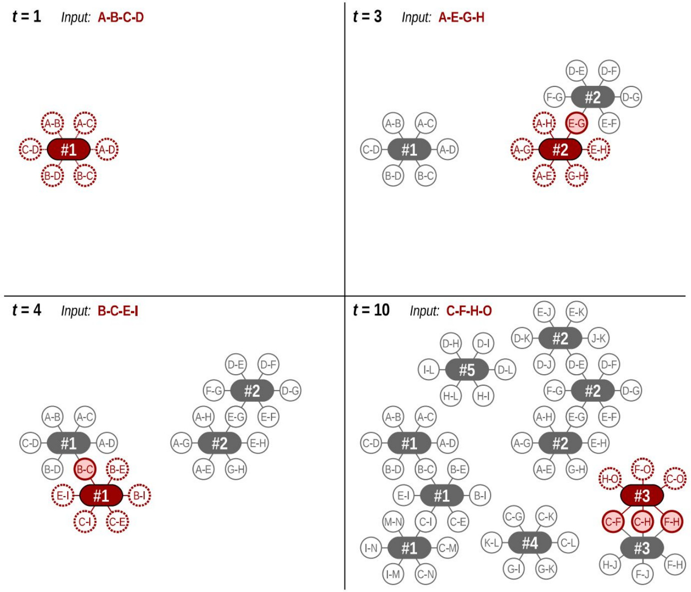


Clustering algorithms: MIMOSA[5], hierarchical, general

[5] J. A. Marshall and L. C. Rafsky, Exact clustering in linear time, 2017

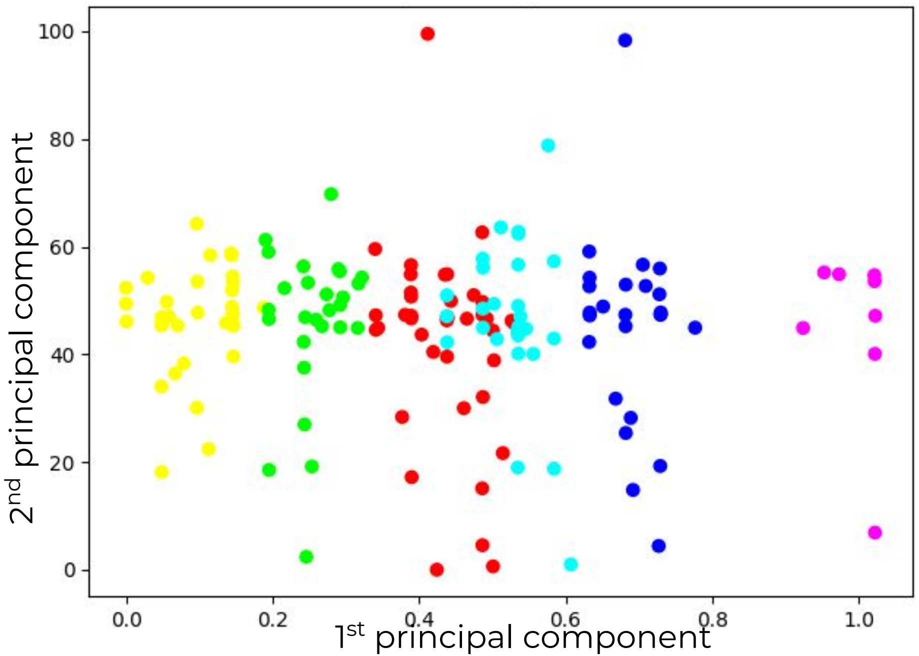
+: Clustering (hierarchical & MIMOSA)





+: **Clustering**

Visual evaluation of the insights plotted on a bidimensional space



plot of the clusters found with the general clustering for the account *theffballers*

Companies must know their Customers

Not knowing your customers can lead to serious mistakes (eg. GAP's rebrand in 2010)[1]

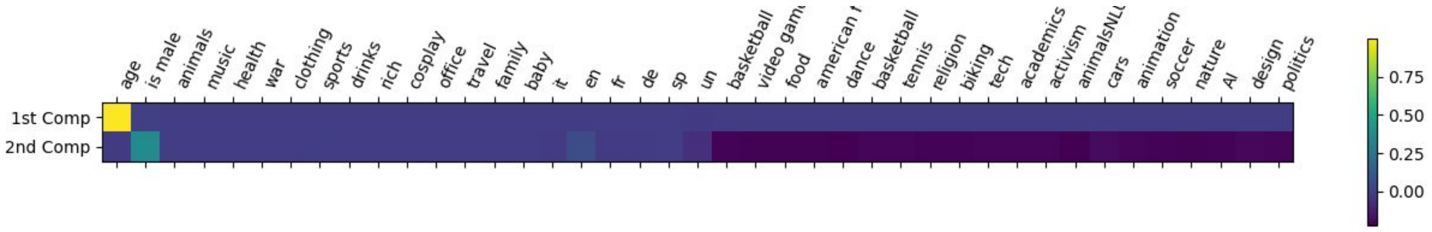
GAP — Gap — GAP

1986 - 9th October 2010 | 9th October - 12th October 2010 | 12th October 2010 - Present

[1] Kelly O'Nail, Rebranding - Successes and Failures, 2017

+: Visualization

PCA on 4 accounts (900+ accounts)



PCA on zalando_uk account: 224 users