Analysis of niche tourist behaviour based on social network data

Ca' Foscari University of Venice
Department of Environmental Sciences, Informatics and Statistics

Computer Science Master's Thesis
Year 2021-2022

Graduand — Lorenzo Padoan
Supervisor — Claudio Lucchese
Co-supervisor — Alessandra Raffaetà

TABLE OF CONTENTS

1 Introduction

Motion Analytica

- Start-up
- Mobility & Tourism Analytics
- Telco data
- Expansion to social network data thanks to this work

Goals

Given a dataset of Tripadvisor reviews provide:

- A method that allows for the identification of tourist niche behaviors in Italy
- Implement an ETL that incorporates the identified method

2 Data Description

2) Data Description

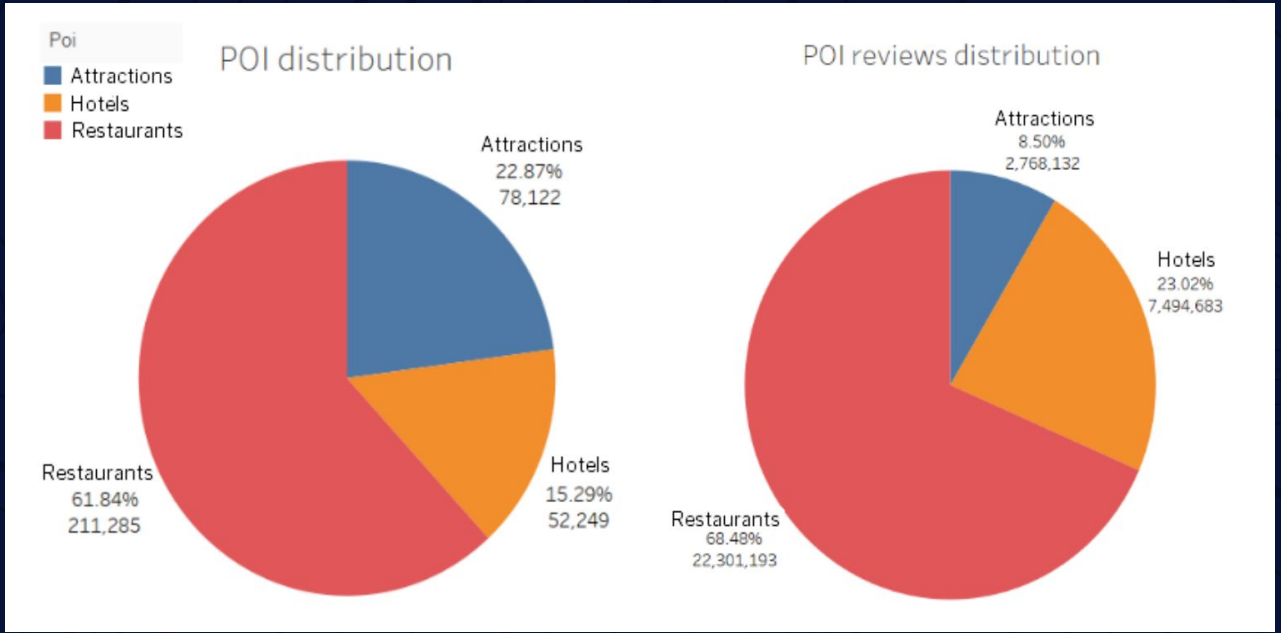
Entity Relationship Diagram

2) Data Description

POI & POI reviews comparison

- More than 30 millions of reviews
- More than 300000 POIs tracked
- Number of Attractions > Hotels
- But number of reviews Attractions < Hotels

Dataset shows consistency with actual TripAdvisor usage

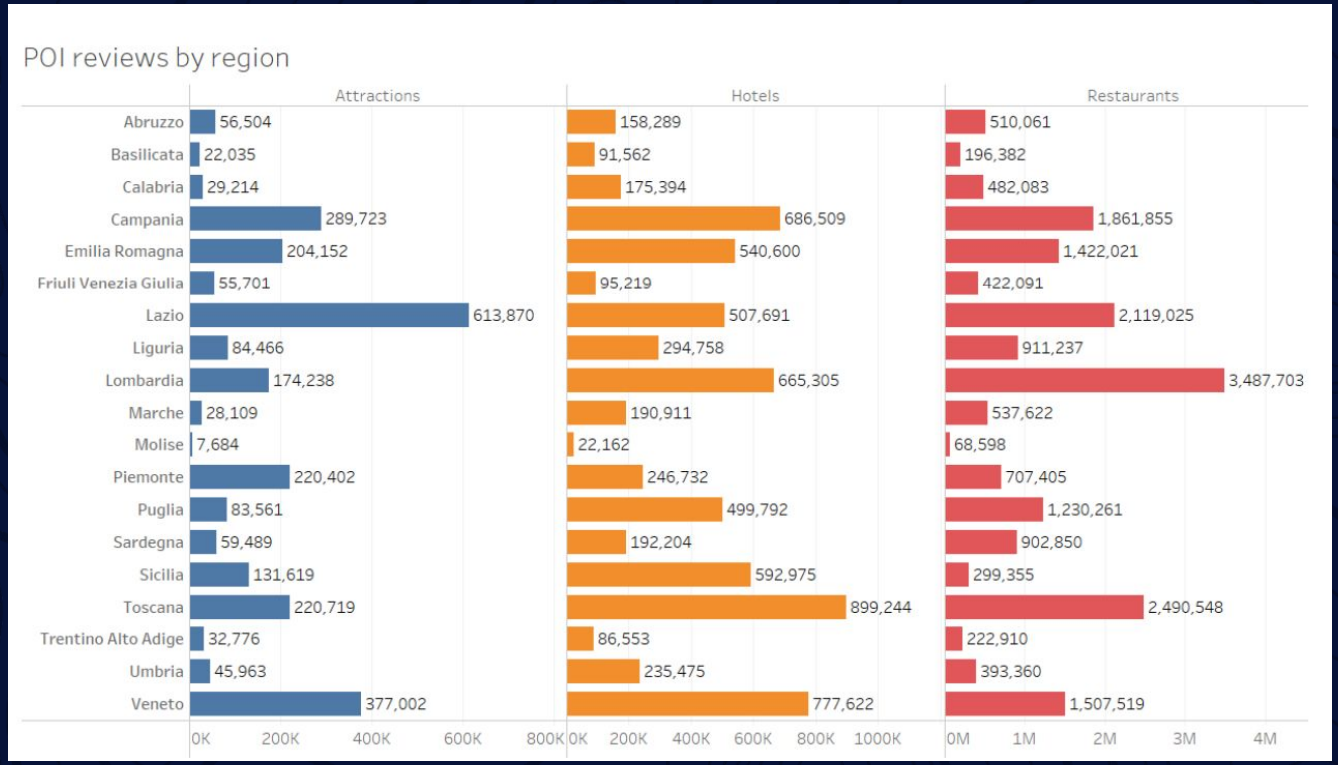


2) Data Description

POI Reviews by region

- Top 5 regions by number of reviews:
  - Veneto
  - Lazio
  - Lombardia
  - Toscana
  - Campania
- These are the top regions by number of visitors according to regional data

The distribution of reviews by region in the dataset is in agreement with actual visit data

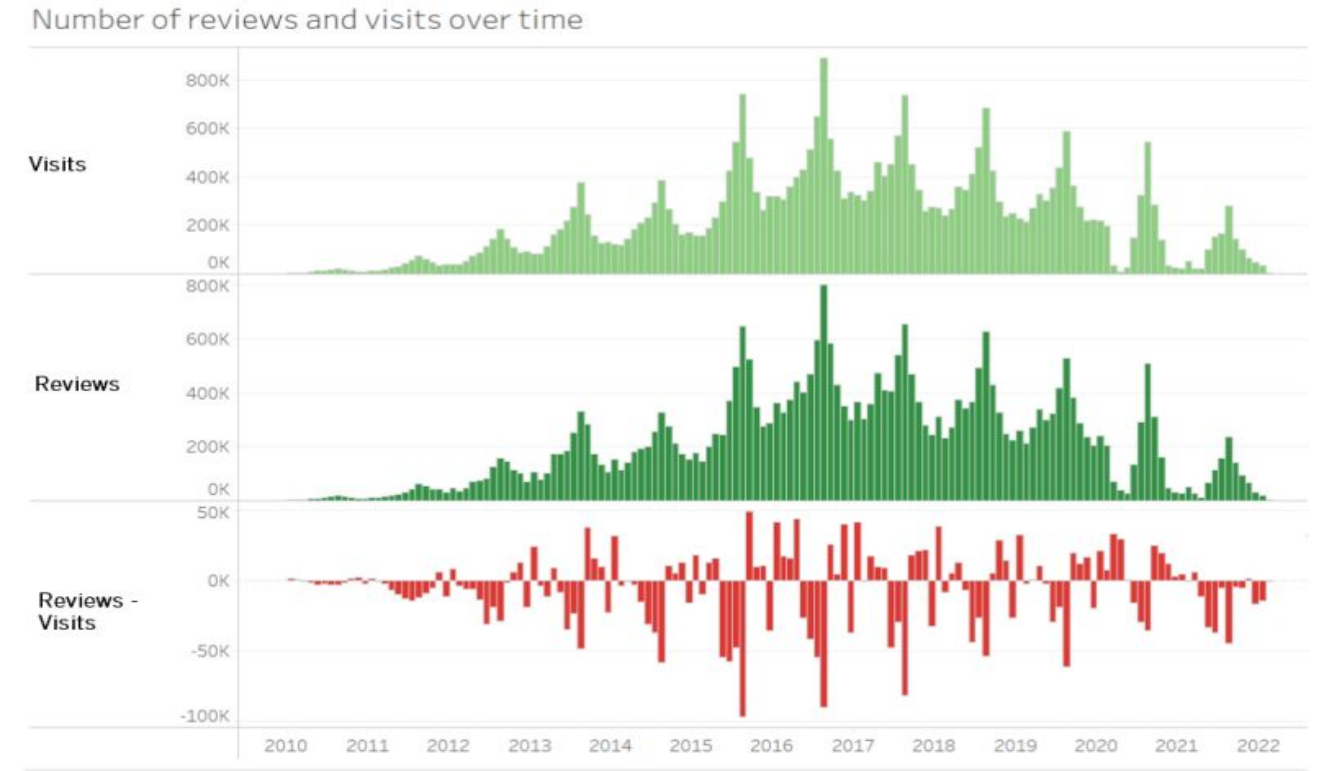


2) Data Description

Reviews & Visits overtime

- The peaks correspond to the month of August
- There is a discrepancy between the day the POI is visited compared to when it is reviewed

The distribution of reviews over time corresponds to the real trend. Summer brings more people on vacation, which means more visits and reviews



EDA considerations

- Very large dataset
- The reviews in the dataset trace the tourist visits
- This dataset is assumed to be representative of the preferences of tourists

3 Data Transformation

3) Data Transformation

Origin Reviewer Field

- The nation of origin of reviewers is vital for Motion Analytica business purposes
- Problem: During the TripAdvisor sign-up process for a long time the origin field was treated as free text
- Because of this many people in this field indicate only the city, only the country or both. Very often introducing typos.

To solve this problem, a software module has been implemented as part of the ETL pipeline that from the origin indicated by the user is able to trace back to the country, expressed through the ISO 3166 standard. The implementation details are part of the ETL section.

Input	Output
Mosca, Russia	Russian Federation
Moscov	Russian Federation
Russia	Russian Federation
Москва, Россия	Russian Federation

3) Data Transformation

Time-Origin-Destination Matrix

Intuitions behind this data rearrangement

- Different nationalities of reviewer have different ways of behaving in different places in different times
- The aggregate of reviews with respect to nationality to a time and location provides an indicator of a specific group of reviewers' preference for some POIs over others.

The 3 main dimensions of refined data

- **TIME**: Refers to a particular month, expressed in mm-yyyy
- **ORIGIN:** It is the origin country of the reviewer, expressed with ISO 3166
- **DESTINATION**: Refers to the Italian province.

3) Data Transformation

TOD matrix snippet

Time	Origin	Destination	Musei d'arte	....	Hotel 3 stelle	...	Ristorante di Pesce
2019-08	Germany	Venice	7	....	10	...	8

Interpretation

in August 2019, Germans in the province of Venice left 7 reviews to art museums, 10 to 3-star hotels, 8 to seafood restaurants.

TOD matrix details

- **Time, Origin, Destination** are the main dimensions of the dataset
- The other dimensions make up the specific type of POIs that TripAdvisor provides
- The TF-IDF algorithm is applied to the data in this form

3) Data Transformation

TF-IDF background

- Term Frequency-Inverse Document Frequency is a statistical metric
- It grows proportionately with the frequency of a term in a document but is adjusted by the number of documents that include the word
- As result common terms in every text are ranked poorly

TF-IDF formulas

Definition 1.

$$TF(t,d) = \log(1 + \ freq\ (t,d))$$

Definition 2.

$$IDF(t,D) = \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right)$$

Definition 3.

$$TF - IDF(t,d,D) = TF(t,d)IDF(t,D)$$

3) Data Transformation

TF-IDF adaptation for TOD matrix

- *t is a POI type*
- *d* is identified by < Time,Origin,Destination >
- *freq(t,d)* is the value in row *d* =< Time,Origin,Destination > and column *t*
- *D* is the set of all the combination of values < Time,Origin,Destination > in the dataset
- *N* is the cardinality of *D* or in matrix terms the number of rows
- *count(d ∈ D : t ∈ d)* corresponds to the amount of rows given the column *t* greater than 0

$$TF(t,d) = \log(1 + \ freq\ (t,d))$$

$$IDF(t,D) = \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right)$$

$$TF - IDF(t,d,D) = TF(t,d)IDF(t,D)$$

3) Data Transformation

TF-IDF adaptation for TOD matrix

Interpretation

- Low TF-IDF value indicates a POI type that is not frequently reviewed or is also very common in the dataset
- A high TF-IDF value may indicate niche behavior

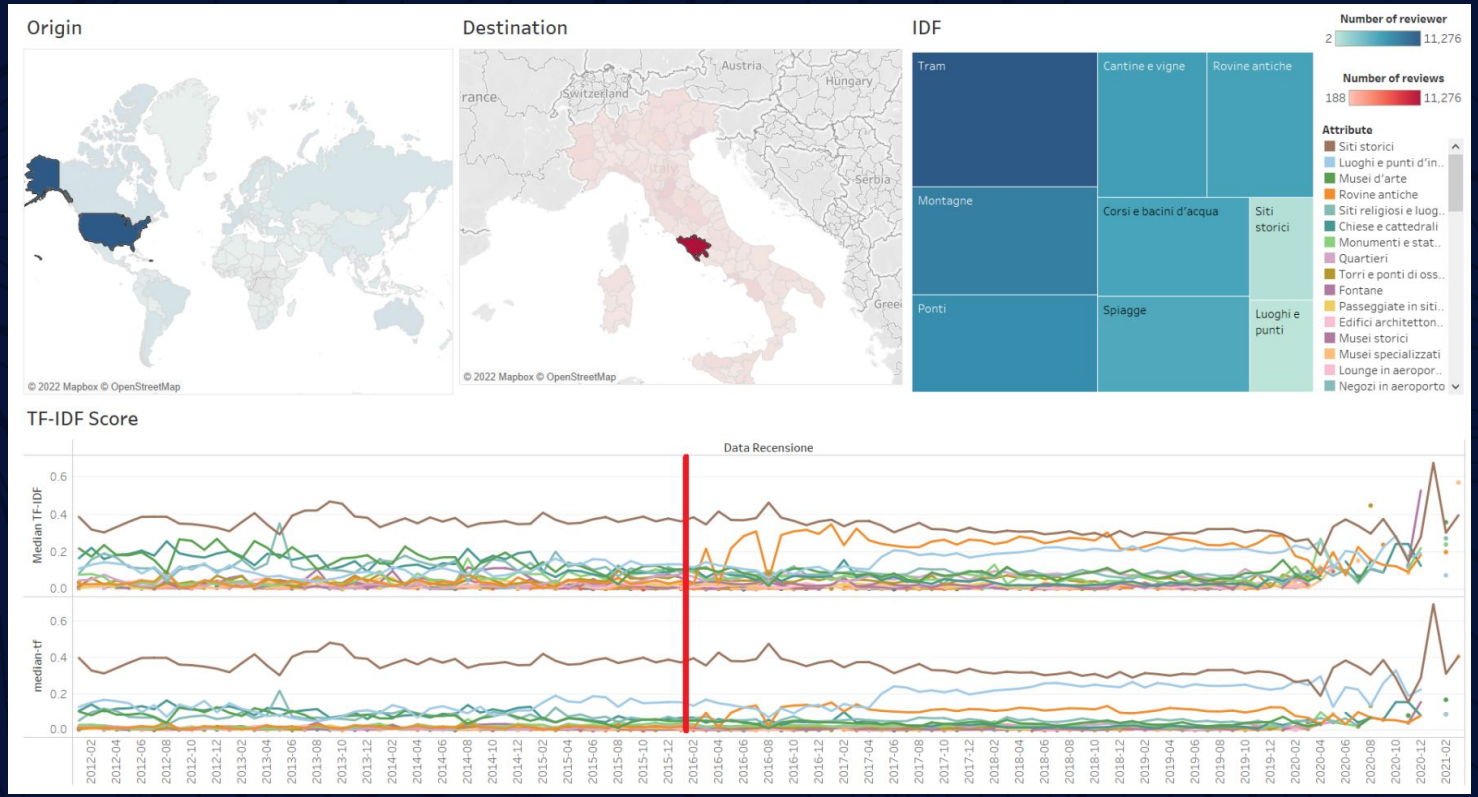
A niche behavior is a moderate number of reviews from a specific nationality in a specific location at a specific time that are uncommon in the TOD matrix.

4 Data Visualization

4) Data Visualization

Dashboard design & Niche Behavior

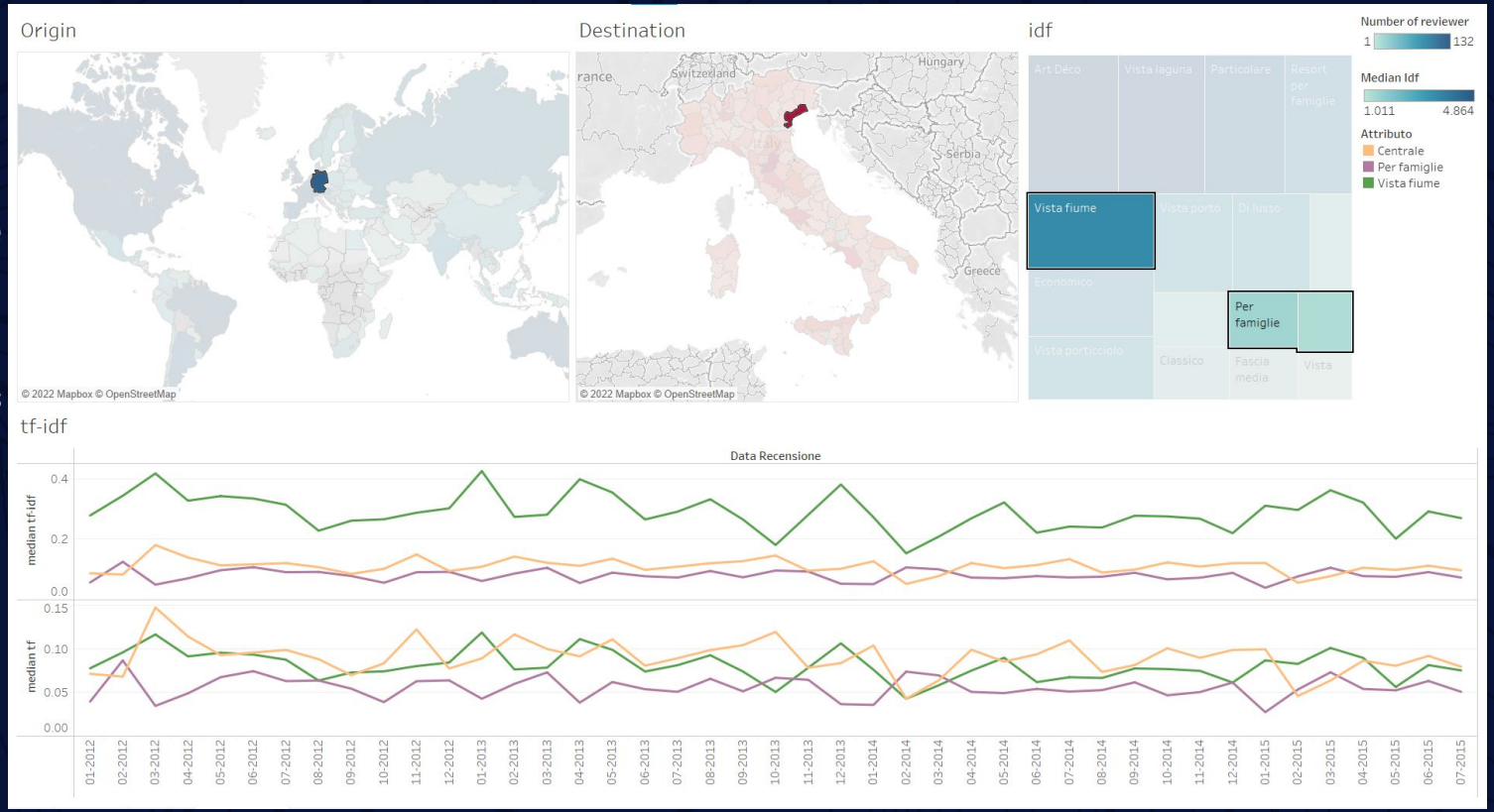
- Implemented interactive dashboard using Tableau
- Reviewers from the US in the province of Rome manifest niche behavior starting from 02-2016 with respect to the Ancient Ruins attraction



4) Data Visualization

Another example of niche behavior

- Niche tourist behavior for Hotel style attribute
- Reviewers from the Germany in the province of Venice manifest niche behavior the Hotels with a river view



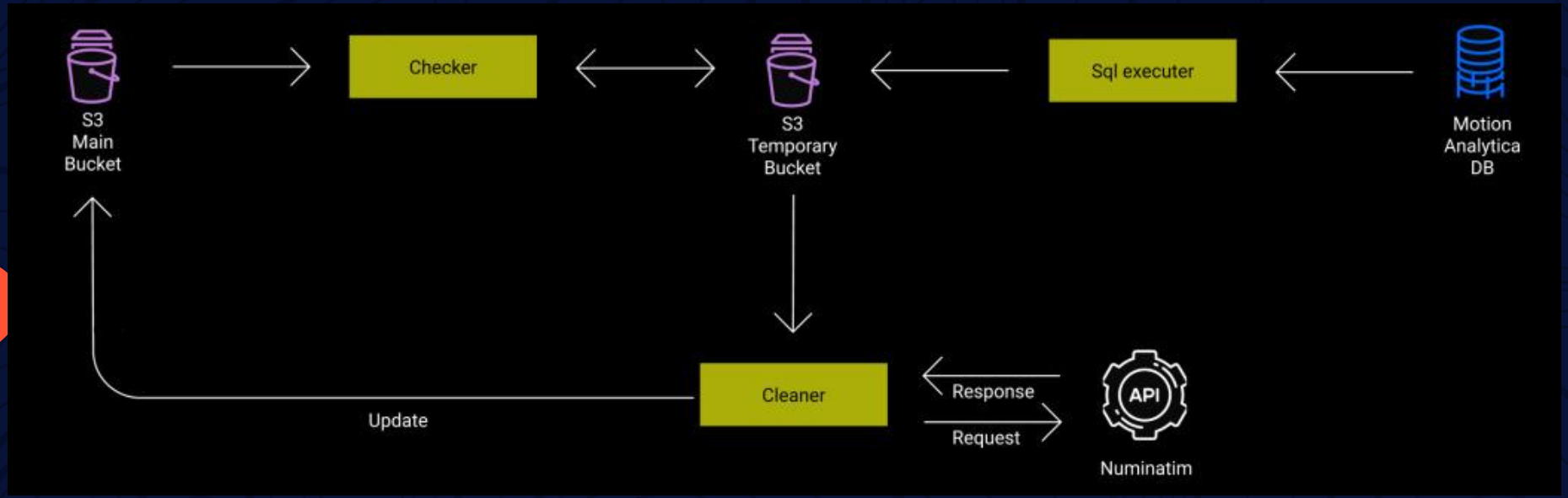
5 ETL Architecture

ETL system requirements

- **Business Needs**:
  Data in the form of a TOD matrix, to be updated once a month, once new data from TripAdvisor has been deposited in the corporate database
- **Data Latency**:
  The results of the TF-IDF algorithm applied to the TOD matrix must be available within one day of depositing the updated TripAdvisor data in the corporate DB
- **Business Intelligence Delivery Interfaces**:
  The refined data must be compatible with the Tableau software in which the dashboards have been implemented

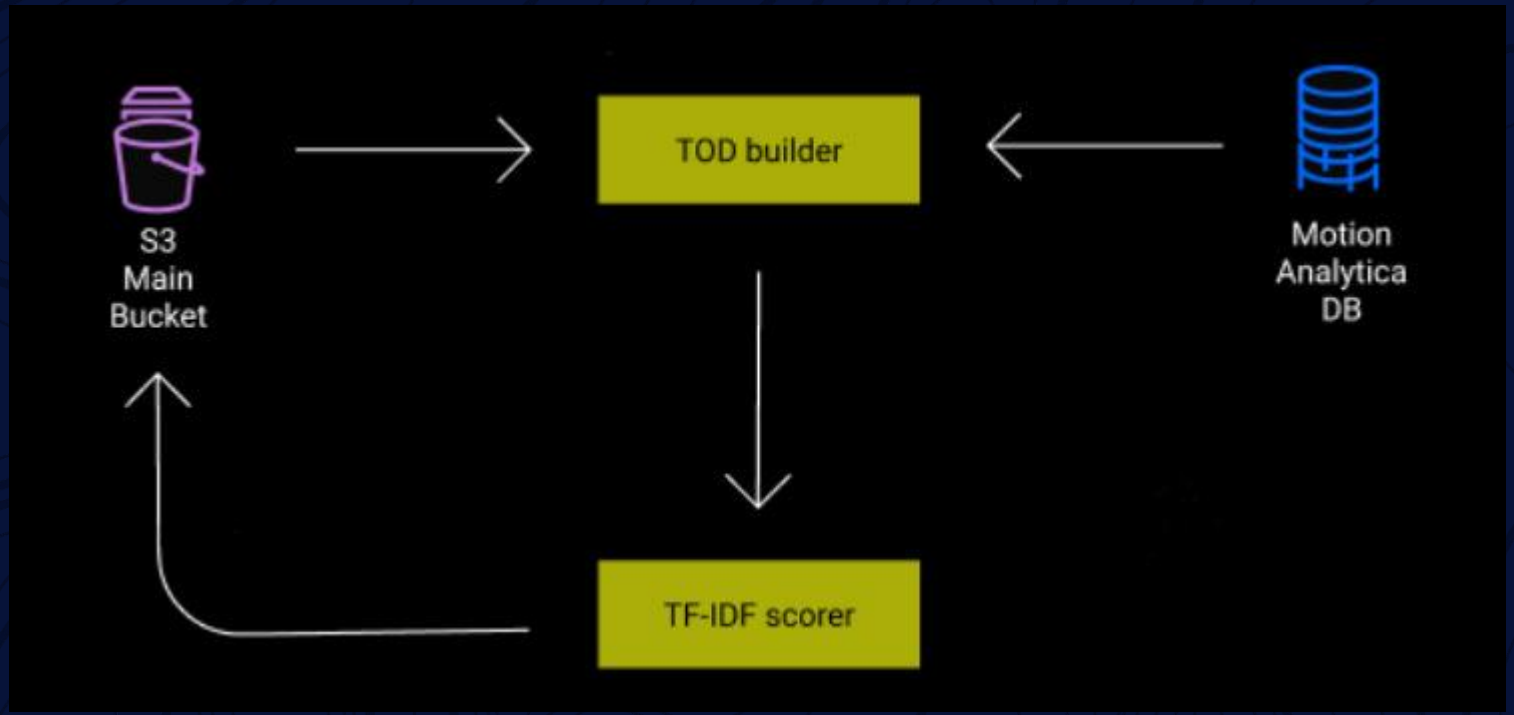
5) ETL Architecture

Origin Cleaner Module



5) ETL Architecture

Transformation-Evaluation Module



6 Conclusion & Future Work

Conclusion

- An approach has been found that can identify niche behaviors of tourists in Italy
- Implemented an ETL system that uses the repurposing of the TF-IDF algorithm to the TOD matrix context
- Motion Analytica has found the work satisfactory and some of their largest clients including Regione Lombardia have shown interest in this particular refined data

Future Work

- Integrate this data with telco data to provide new insights

Q&A Session