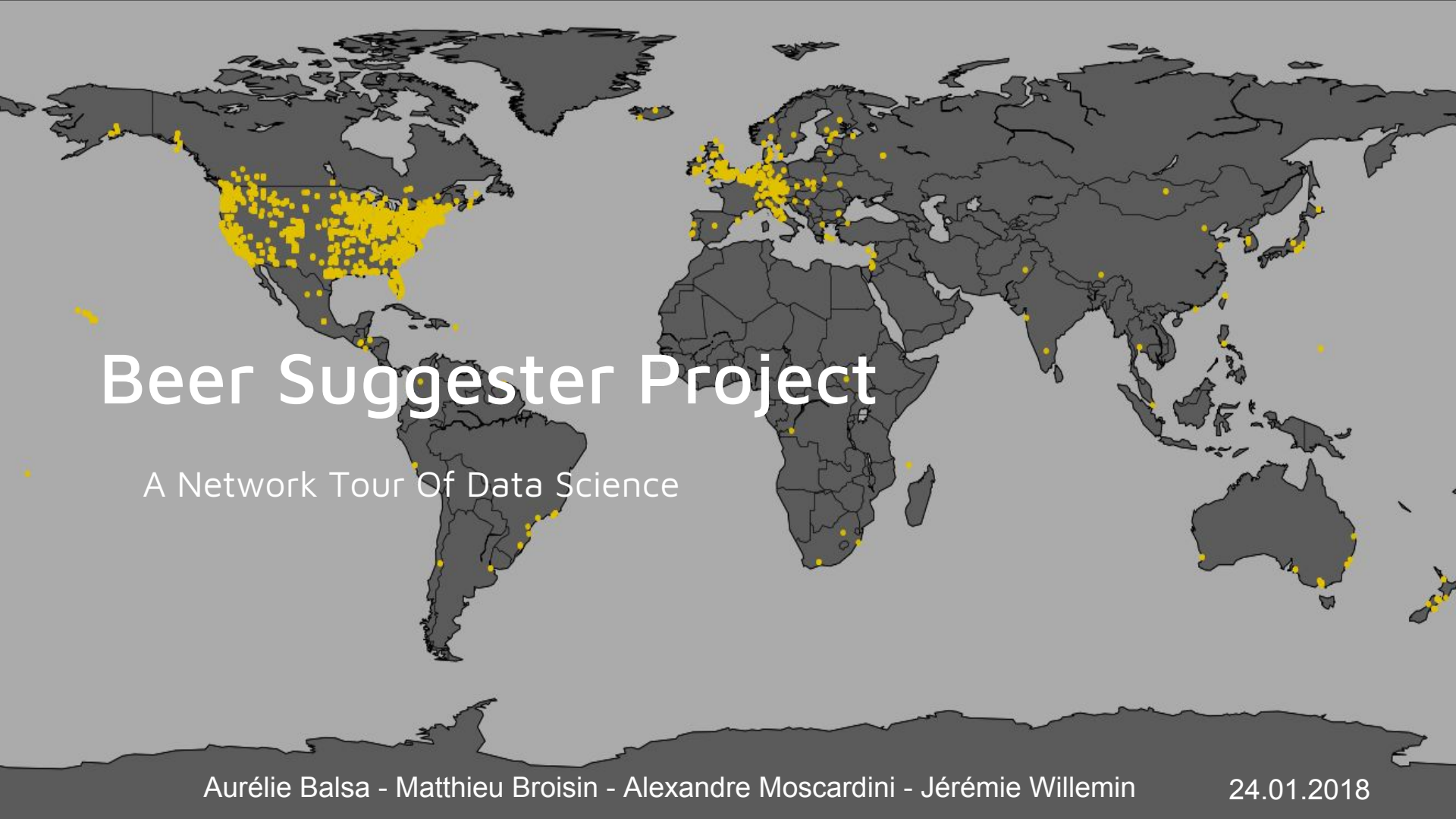


# Beer Suggester Project

A Network Tour Of Data Science

Aurélie Balsa - Matthieu Broisin - Alexandre Moscardini - Jérémie Willemin

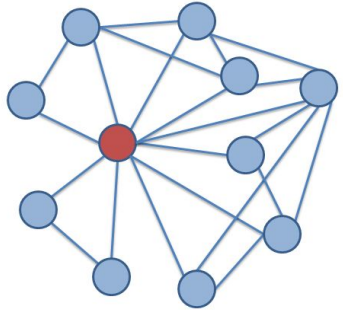
24.01.2018

# Purpose of the Project

Goal: Making **beer suggestions**

Why: Discover new refreshing drinks that the user should like.

How: Based on a **similarity network**.
User provides a reference (model) of a beer.



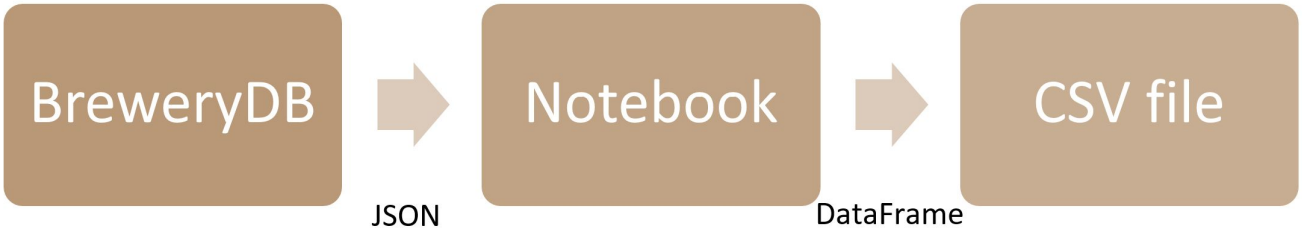
# Dataset

Data comes from **BreweryDB.com**


BreweryDB gets their data from individuals. Everyone can upload beers data on their website, but administrators validate the information to insure the accurateness of the data.

# Data Collection & Cleaning

BreweryDB provides an **API**.




We had to outcome problems of **special characters** in strings and **missing fields.**

# Beer characteristics

Typically, beers are characterized by:

- IBU: metric of bitterness
- ABV: alcohol by volume
- SRM: color metric [1-40]

SRM values

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |

We can add also where the beer is made:

- Latitude of the brewery
- Longitude of the brewery

# Data cleaning

SRM missing
or
IBU missing
or
ABV missing

| Style |
| --- |
| • abv max<br>• abv min<br>• srm max<br>• srm min<br>• ibu max<br>• ibu min |

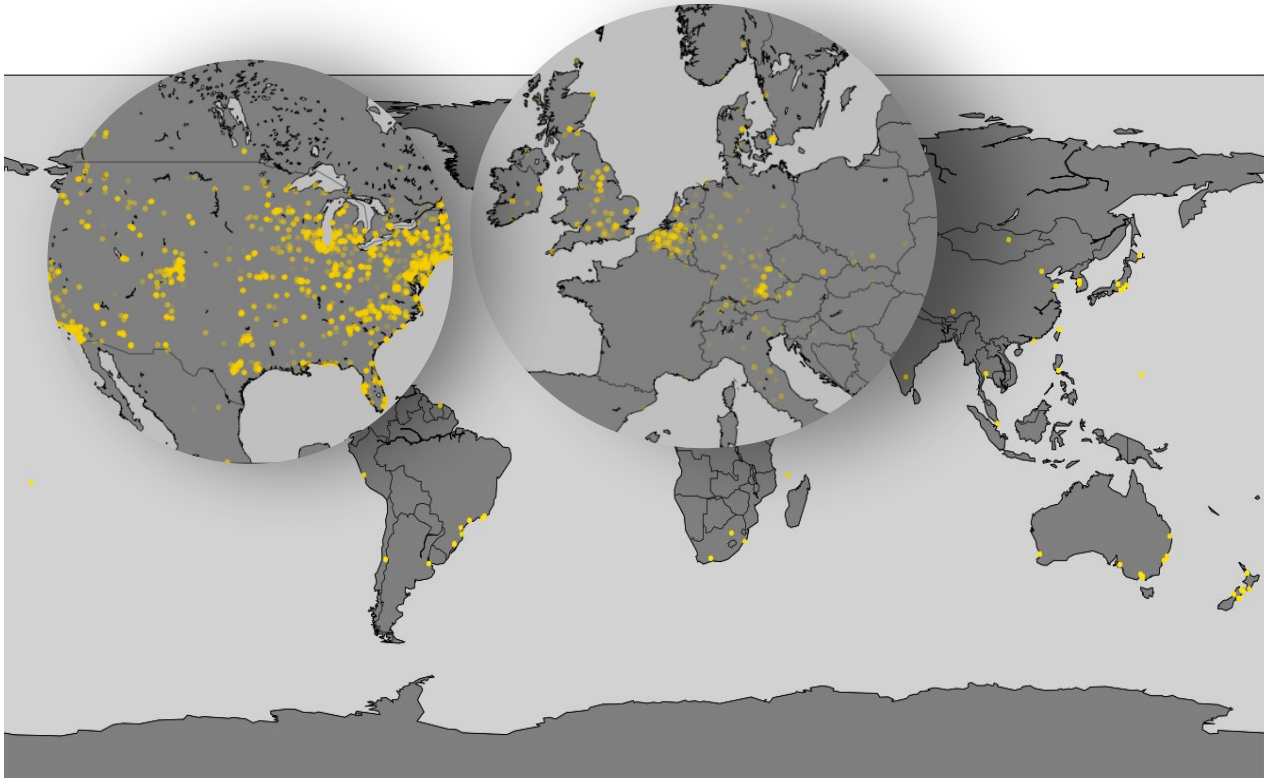
Otherwise we discarded the beers for which we didn't have neither the values or the style.

~16'500 ⟶ ~15'400

# Visualization of the Dataset

Breweries Location mostly in:

- USA
- EUROPE (UK, Belgium, Germany)



*Mercator projection breweries location*

# Distances Computation

For the **Features**, we use an euclidean distance.

$$Feature = \begin{bmatrix} abv_1 & ibu_1 & srm_1 \\ abv_2 & ibu_2 & srm_2 \\ \vdots & \vdots & \vdots \\ abv_N & ibu_N & srm_N \end{bmatrix} \cdot \underbrace{\begin{bmatrix} w_{abv} & 0 & 0 \\ 0 & w_{ibu} & 0 \\ 0 & 0 & w_{srm} \end{bmatrix}}$$

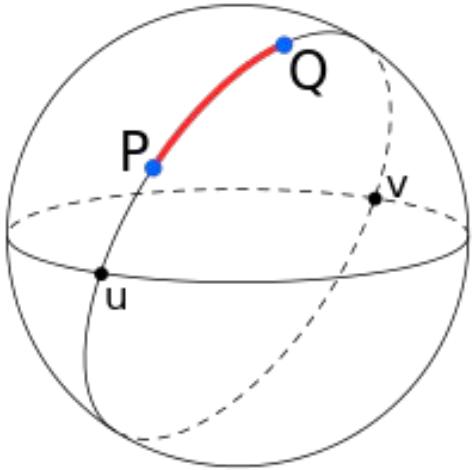
Weights that can be tuned if the user is more sensitive to one feature than another.

$$DistanceF = \begin{bmatrix} 0 & d(1,2) & \dots & d(1,N) \\ d(2,1) & 0 & \dots & d(2,N) \\ \vdots & \vdots & \vdots & \vdots \\ d(N,1) & d(N,2) & \dots & 0 \end{bmatrix}$$

$$d(b_1, b_2) = \|b_1 - b_2\|_2 = \sqrt{(b_1 - b_2)^T (b_1 - b_2)}$$

# Distances Computation

For the **location** (lat, long) we cannot use an euclidean-like distance between vectors because the earth is spherical → **haversine formula**




$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

φ: latitude
λ: longitude
R: Earth radius (6371 km)

[https://www.movable-type.co.uk/scripts/latlong.html]

# Distances

Finally, to regroup both distances:

coef. to adjust the ponderation of the localization

$$d_{i,j} = \frac{d_{i,j}^{euclidean}}{max\left(d^{euclidean}\right)} + w\frac{d_{i,j}^{haversine}}{max\left(d^{haversine}\right)}$$

Related to the features

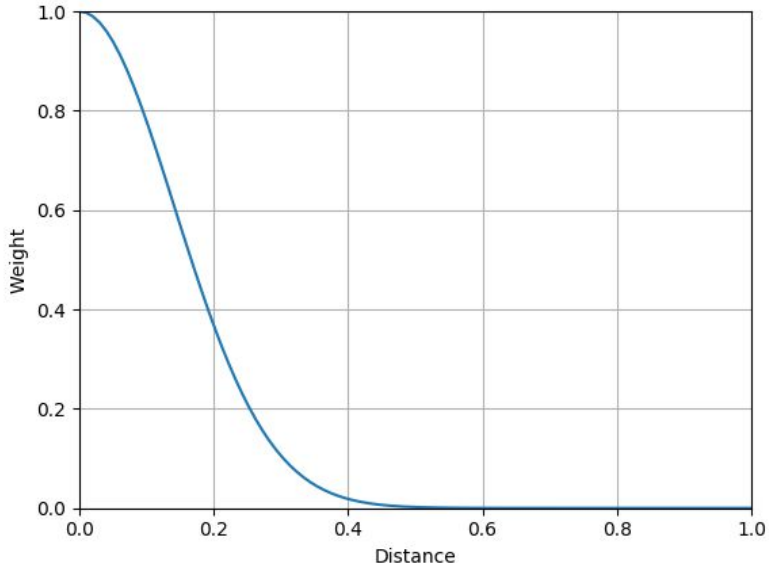
Related to the location

# Weight matrix - graph construction

In order to build the graph, we converted distances to weights with a Gaussian kernel:

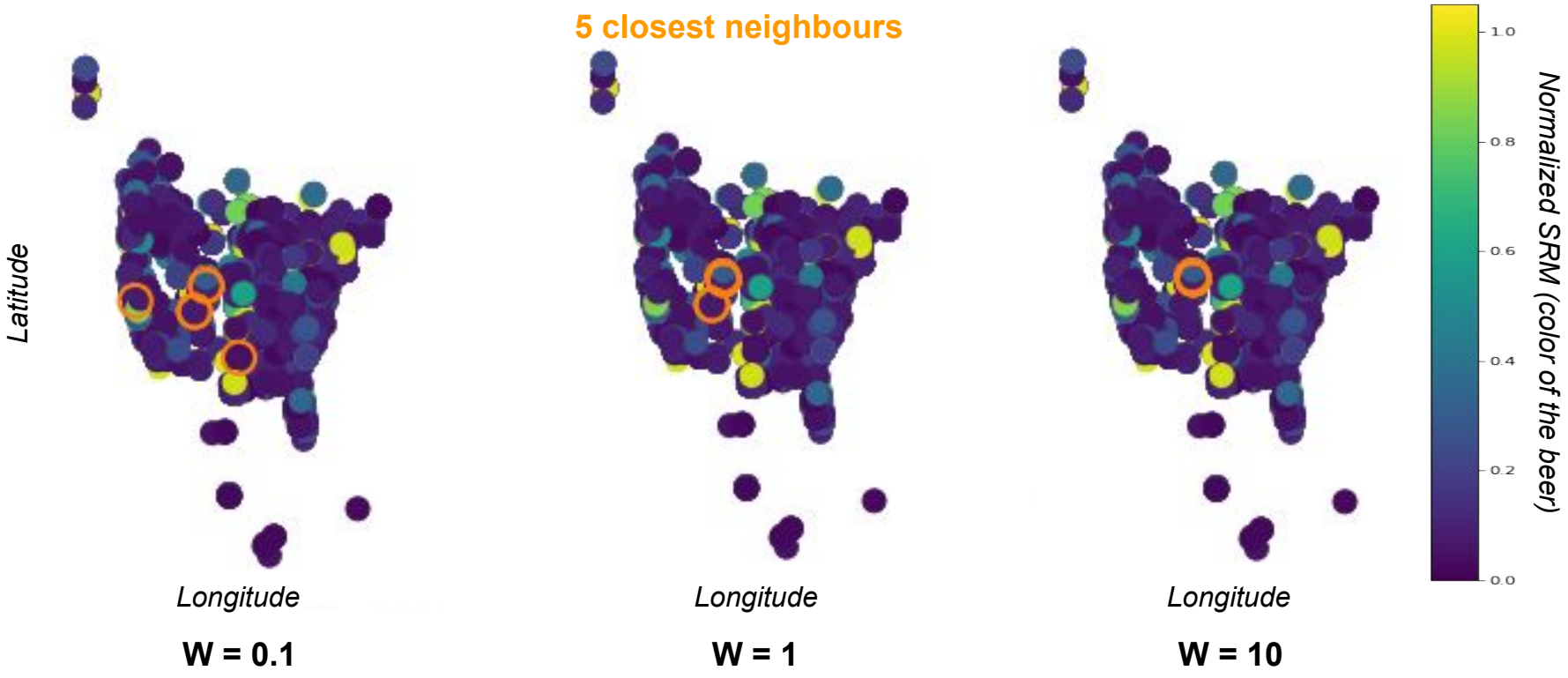
$$W_{i,j} = exp\left(\frac{-d_{i,j}^2}{\sigma^2}\right)$$

Shape of a Gaussian kernel →



# Distance ponderation Comparison

**5 closest neighbours**




**W = 0.1** **W = 1** **W = 10**

# Distance ponderation Comparison

| | abv | ibu | srmId |
|---|---|---|---|
| **10040** | 0.232558 | 0.071071 | 0.400 |
| **13449** | 0.223699 | 0.099099 | 0.425 |
| **14310** | 0.252492 | 0.083083 | 0.400 |
| **15541** | 0.250277 | 0.119119 | 0.425 |
| **14892** | 0.265781 | 0.042543 | 0.425 |
| **0** | 0.241417 | 0.074074 | 0.425 |

**W = 0.1**

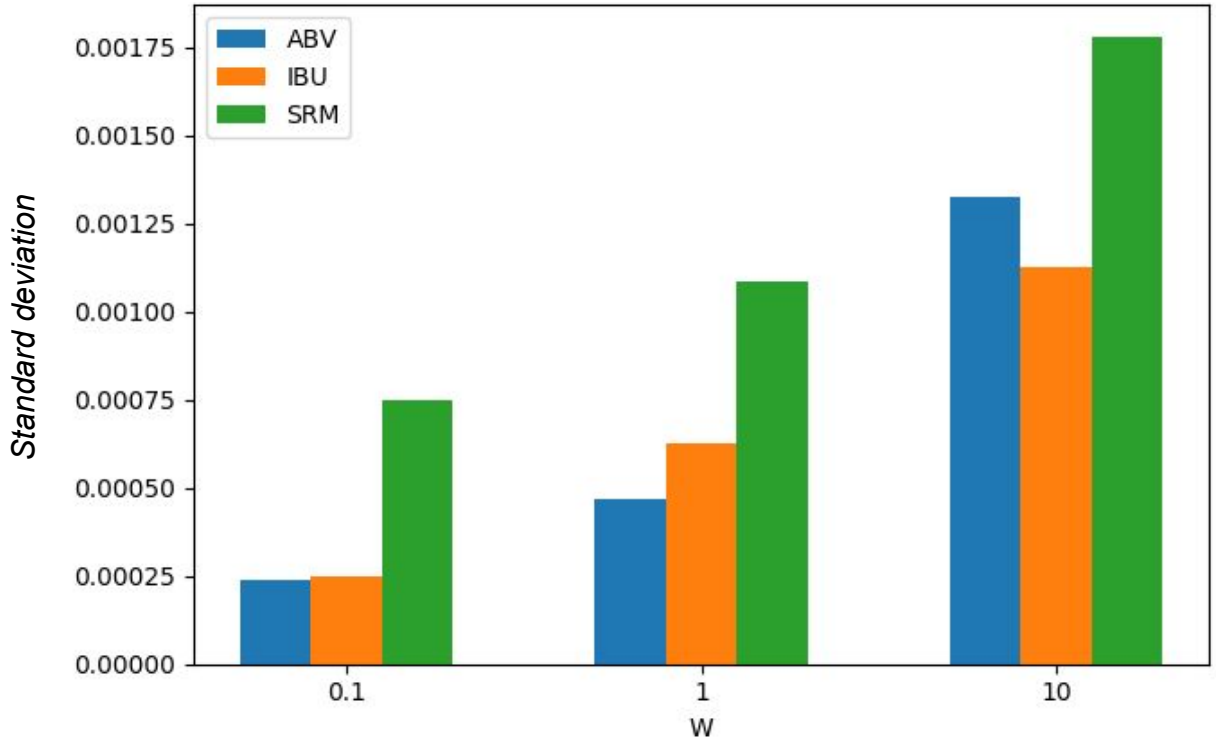
| | abv | ibu | srmId |
|---|---|---|---|
| **15541** | 0.250277 | 0.119119 | 0.425 |
| **15533** | 0.281285 | 0.029029 | 0.425 |
| **303** | 0.181617 | 0.041041 | 0.400 |
| **8988** | 0.208195 | 0.019019 | 0.475 |
| **12201** | 0.224585 | 0.101101 | 0.375 |
| **0** | 0.241417 | 0.074074 | 0.425 |

**W = 1**

| | abv | ibu | srmId |
|---|---|---|---|
| **15541** | 0.250277 | 0.119119 | 0.425 |
| **10040** | 0.232558 | 0.071071 | 0.400 |
| **13449** | 0.223699 | 0.099099 | 0.425 |
| **15533** | 0.281285 | 0.029029 | 0.425 |
| **12201** | 0.224585 | 0.101101 | 0.375 |
| **0** | 0.241417 | 0.074074 | 0.425 |

**W = 10**

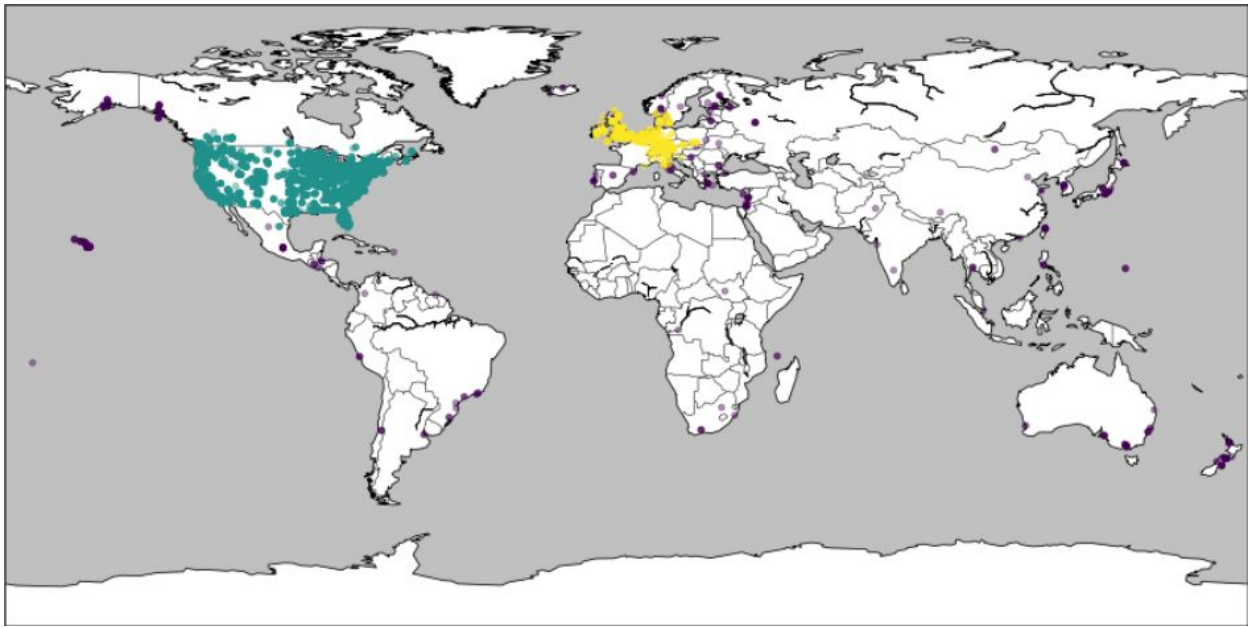
# Distance ponderation Comparison



# Location Clustering

Clustering with DBSCAN:

1) Locations (longitude, latitude)

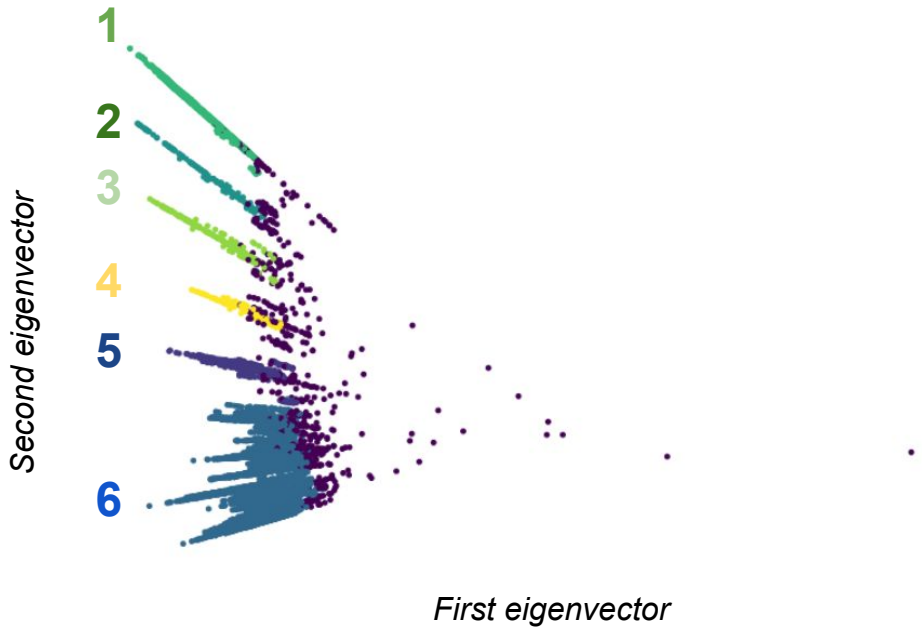


*Clustering of locations with DBSCAN and Mercator projection*

# Features Clustering

Clustering with DBSCAN:

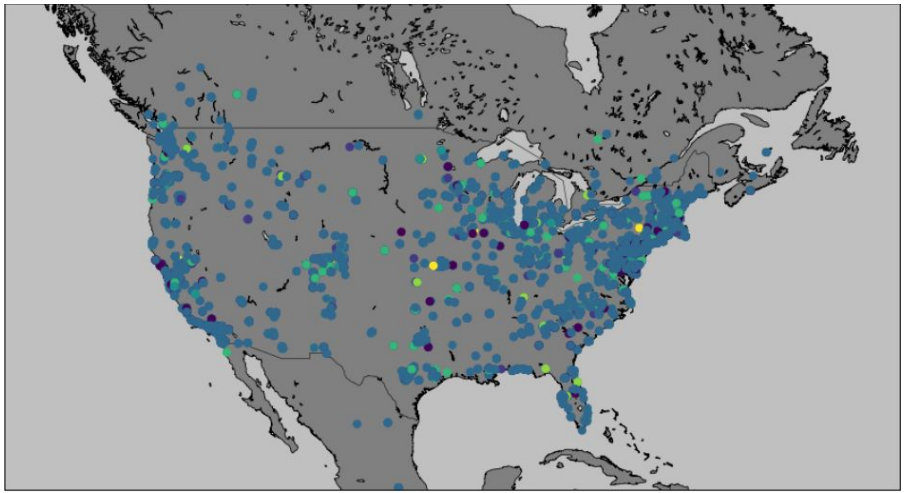
1) Locations (longitude, latitude)
2) Features (IBU, ABV, SRM)



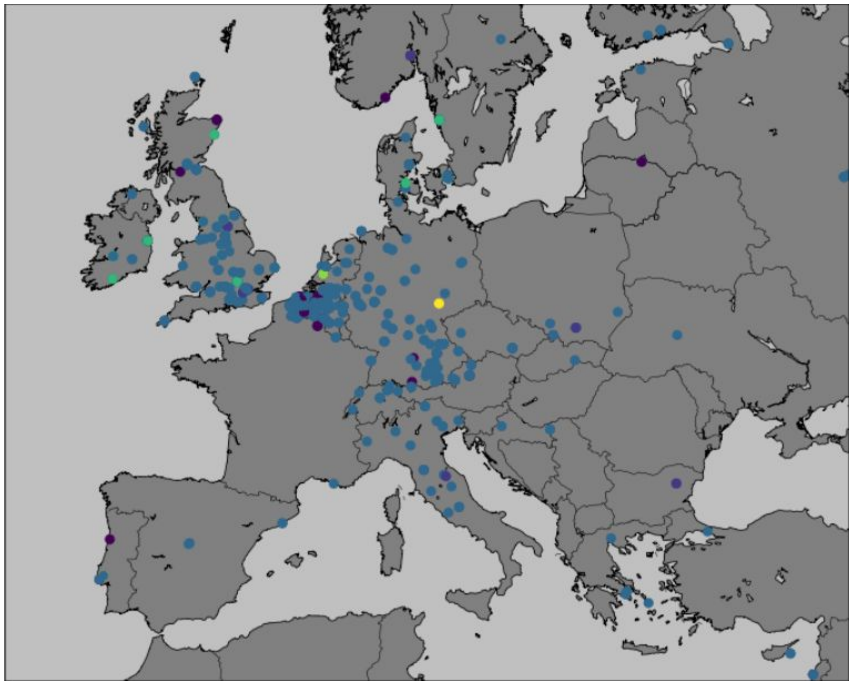

*Clustering of features with DBSCAN and projection on the two first eigenvectors*

# Are Types of Beer Related to a Specific Region?

Result of the clustering by region



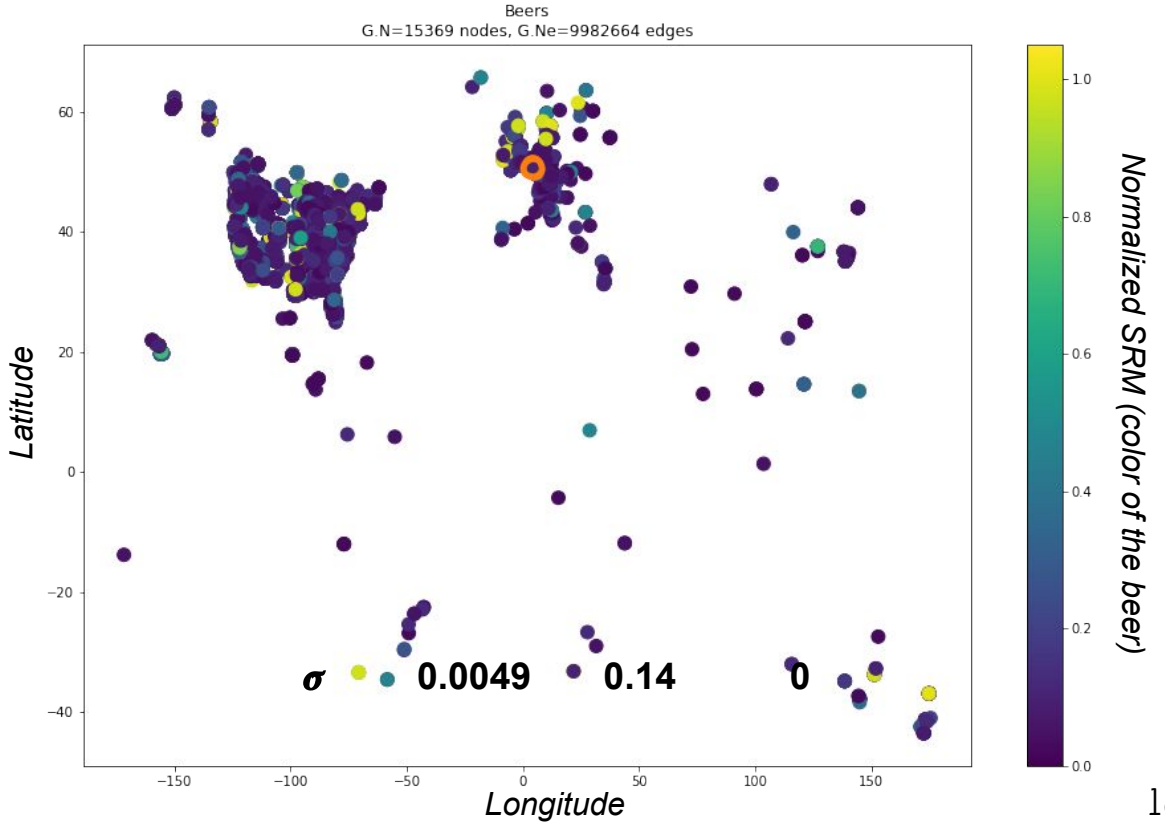
*Results of the clustering in USA*



*Results of the clustering in Europe*

# Demo


| ABV | IBU | SRM |
|---|---|---|
| 6.6% | 28 | 4 |



# Demo


| ABV | IBU | SRM |
|---|---|---|
| 6.6% | 28 | 4 |

| | abv | ibu | srmId |
|---|---|---|---|
| **14616** | 0.139535 | 0.033033 | 0.075 |
| **12676** | 0.141750 | 0.019019 | 0.075 |
| **181** | 0.141750 | 0.019019 | 0.075 |
| **14824** | 0.148394 | 0.019019 | 0.075 |
| **14120** | 0.150609 | 0.019019 | 0.075 |
| **15094** | 0.143965 | 0.027027 | 0.075 |
| **$\sigma$** | **0.0049** | **0.14** | **0** |

# Demo


| ABV | IBU | SRM |
|---|---|---|
| 6.6% | 28 | 4 |


| ABV | IBU | SRM |
|---|---|---|
| 6.8% | 20 | 4 |


| ABV | IBU | SRM |
|---|---|---|
| 6.5% | 20 | 4 |

# Conclusion

- Purpose of the project
- Dataset and beers features (IBU, ABV, SRM and Location)
- Data collection/cleaning and visualization
- Distances computation/Weights matrix
- Relation analysis between the beer’s type and location
- Demo → Beer Suggestion


Thank you for your attention