# Lecture 03: Distributions and patterns
# Exercise description

## Goals

Exploration of distributions, identifying and interpreting patterns in distributions.

- Exploration of temporal and spatial distributions of discrete entities using time histograms and maps.
- Exploration of a joint distribution of multiple attributes over a set of entities using projection and partition-based clustering.

## Data

### Original data

The original data are geolocated Twitter messages (i.e., the records include geographic coordinates) from the territory of the UK and Ireland and the time period from March 27 till April 2, 2015. The tweets are supposedly related to a storm that happened in this period. From a large set of tweets available for this territory and time period, only those containing storm-relevant keywords were selected.

### Data pre-processing

We created the following list of keywords: storm, wind, rain, snow, hail, flood, road, collapse, injury, disrupt, and forecast. The keywords with their synonyms are contained in file keywords.txt. Using the keywords and synonyms, we generated 11 attributes with values 1 and 0 denoting the presence and absence of the respective keywords in the message texts or tags. The tool that we used also generated an attribute 'features' with text values composed of the keywords that were present in the messages or tags. The values of this attribute where taken as 'names' of the tweets, to be used as labels in visual displays.

Based on these 11 attributes, we computed further attributes:

- 'bad weather (sum)': sum of the values of the attributes referring to bad weather conditions, i.e., storm, wind, rain, snow, and hail;
- 'bad weather': 1 if the tweet contains any keyword referring to bad weather conditions and 0 otherwise;
- 'consequence (sum)': sum of the values of the attributes referring to storm consequences, i.e., flood, collapse, injury, disrupt;
- 'consequence': 1 if the tweet contains any keyword referring to storm consequences and 0 otherwise.

## Draft Python scripts

We provide you with several draft scripts in Python (as several Jupyter notebooks). The scripts load the data and apply some operations to them in accord with the exercise goals and tasks.  You use these scripts as a starting point, but you are expected to adapt them to the tasks at hand along the analysis process. You may use your knowledge of Python to try alternative approaches.

Script contents:

- 2019-03.ipynb: Data loading, transformation of strings specifying dates and times to the type datetime, study of temporal and spatial distribution of tweets.

- 2019-03cont.ipynb: Focus on keywords, their distribution and co-occurrence, temporal and spatial patterns of them.

## Tasks

- Explore, characterize, and interpret the temporal distribution of the tweets. What does it tell about the storm development?
    - Use a time histogram
- Explore and characterize the spatial distribution of the tweets.
    - Use a map
- Explore and characterize the joint distribution of the different keywords. Which combinations of weather phenomena and/or events that happened during the storm occur frequently in the tweets? Which are rare? Which keywords tend to occur more in isolation from others and which in combinations with others?
    - Use a projection
- Explore how the spatial distribution and the keyword distribution changed from day to day
    - Select data subsets from the different days and represent them on the map and in the projection plot.

We suggest you to note your findings as comments in the notebooks you are using.

## Training of method use

- In applying a time histogram, investigate the impact of changing the bin width.
- In applying projection, investigate the impact of using additional attributes: first apply the projection algorithm to the keyword attributes only and then add the attributes ‘bad weather’ and ‘consequence’.
- Use projection in combination with partition-based clustering (k-means) applied to the same attributes as the projection. Observe how the clustering algorithm groups the tweets. Note consistencies and inconsistencies between the arrangement of the tweets in the artificial space (relative distances and visual groupings) and the groupings produced by the clustering method.
    - Run k-means and paint the points in the projection plot according to the cluster membership.
    - Extend the script to select data subsets according to cluster membership; look at the distribution of the members of each cluster in the projection plot.
- Run the clustering methods with setting different numbers of clusters (values of the clustering parameters). Observe the effects on the object grouping.

We suggest you to share your notebooks with the changes and notes you have made in the moodle forum.

## Additional material: demo of use of interactive tools

We provide you a demo (4 minutes long) showing how the operations required for fulfilling the exercise can be performed using interactive tools of V-Analytics. You can look at the demo before starting to do the exercise for getting more familiar with the data and the methods applied to them. If you wish, you can later try to redo the exercise using V-Analytics.

We wish you a successful and fruitful fulfilment of the exercise.