# Assignment 1

Welcome to the visualization course! To understand the difficulties and challenges that visualization researchers face today, and the visual design process, we are going to build a visualization tool during the course. Generally, there are various application fields that generate different kinds of data sets, typically recorded over longer time periods. In the lecture, you will see different data types that can be primitive or more complex, but also static or dynamic. Moreover, the data sets can be combinations of several data types. However, we will focus on a specific scenario which is the visual analysis of high-dimensional tabular data.

**Assignments will NOT be evaluated, they are meant for you to have a guideline towards the final project result**

## Exercise 1 – Data Set

You have to choose one of the following two data sets for your project:

### Airbnb Open Data New York

https://www.kaggle.com/data sets/arianazmoudeh/airbnbopendata

**Data Description**

Airbnb is an American corporation that offers an online marketplace for renting and booking accommodations owned by individuals, hotels and investors. The data set has multiple aspects that can be explored.

The data set provided here contains listing activity in New York city (US). The following Airbnb activity is included:

- Listings, including full descriptions and average review score
- Reviews, including unique id for each reviewer and detailed comments
- Calendar, including listing id and the price and availability for that day

You can find the data set in the section 'Files' airbnb_data in canvas.

You can access the data through the csv file named "airbnb_open_data.csv". The Jupyter notebook named "airbnb_open_data.ipynb" can be used for an initial exploration of the data.

A dictionary is made available that describes each attribute, we provide the file Airbnb Open Data Dictionary.xlxs. You can also find it here:
https://docs.google.com/spreadsheets/d/1b_dvmyhb_kAJhUmv81rAxl4KcXn0Pymz/edit#gid=1967362979

### The Insurance Company (TIC) Benchmark

https://www.kaggle.com/data sets/uciml/caravan-insurance-challenge

**Data Description**

# JBI100 Visualization

**Anna Vilanova**

This data set is supplied by the Dutch data mining company Sentient Machine Research and once used in the data mining competition CoIL Challenge 2000. This data set contains information on customers of an insurance company. Although it is well exploited in machine learning areas for prediction tasks, our goal is to analyze the data from various aspects which are suitable for a visualization solution. The data set has multiple aspects that can be explored.

You can find the data set in the 'Files' tic_data section on canvas.

- The data: 'tic_data.csv'
- A description of all the columns: 'TicDataDescr.txt', Column1 in the data corresponds to 1 MOSTYPE for example. You should rename your columns to make them meaningful.
- A small python script to load the data in a data frame: 'load_data.py'

You are welcome to enhance these data with other data sets from other sources to achieve interesting and meaningful analysis. However, they are considered extra and should not be an alternative to choosing one of the proposed data sets

Our goal is to design a visualization tool for high-dimensional data to achieve specific goals/tasks.

One important choice is to identify the goal/users that will be the focus of your visualization design. Notice that not all possible goals are suitable for a visualization solution.

One of the first jobs is to get familiar with the data set and the domain. Overseeing the structure, size, potential, and challenges of a given data set and the domain is one of the key problems in visualization.

(a) What is the information you can obtain from the data set/ data sets?

(b) What are the attributes in the data and what is their meaning?

(c) Write a small parsing function that can read the data position (column, row) from the file format you selected.

(d) Write another function that outputs the distribution of the attributes, and counts the frequencies of the different values.

(e) Try to describe the data set in just a few sentences. How is the data provided? Which kind of attributes are contained in the data set? How large is the data set in terms of the number of those elements (listings, reviews, vehicles, geographic regions and locations, extra records, and so on)?

(f) Analyze the errors and missing values. Write a function to count how many missing values per attribute and per entry you have. Analyze *what are the most relevant missing values that might hinder the analysis according to you.*

# JBI100 Visualization

**Anna Vilanova**

From this exercise on, we give some initial steps and you should start writing the corresponding sections of the interim report. Please, read the final report information provided in canvas. The assignments give a guideline, notice that extra points to what is mentioned in the assignments are needed to have all aspects of the interim report covered.

## Exercise 1 – Goal - Data (Domain specific)

We will be following the nested model presented in the lectures for the visualization design process. So the first step is to understand the domain situation and formulate the goal of the visualization. You need to identify by yourself what user and goal you want to work on. We are in a visualization course so the main goal should suit a visualization solution.

(Introduction) Describe what you envision will be the general overall goal and users of the visualization tool. The goal is meant to be from the perspective of the user, which will be the goal/question from the user's perspective to use the visualization tool. Think about the different goals of visualization presented in class and the high-level actions. Define for which users your tool is meant, and which overall goal. The reason why this goal is suitable for the available data and why this is a goal where a visualization tool is the right means to solve it (e.g., visualization vs. an automatic solution).

## Exercise 2 – Data (What) Domain specific

(a) Write in section *What (Data)* the description of the data. You can base it on the analysis you have done in exercise 1. What are the general properties of the data you want to use?

(b) Most of the data sets contain noise, missing data values, and relations, or measurement errors. The data of this course is no exception. In exercise 1, you already looked at the missing values. How will you handle missing data values or measurement errors? Think of multiple ways and their pros and cons.

(c) (Data (What)) Choose one of the methods and implement it for the data set. Describe it in the section and mention what is the effect on the data.

## Exercise 3 – Data (What) Abstraction

Once the goal, and data are understood from the domain point of view. We enter into the abstraction phase such that we can identify the most adequate designs later on.

Make the data abstraction according to the "what" in Munzner's framework. Present it in a summarized version you do not need to present it for each individual attribute. Build a table with the general overview.