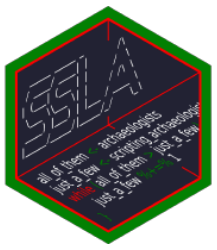
W4 CAA Scripting Languages Hackathon I –



Can you code this?

Clemens Schmid[1], **Martin Hinz**[2], **Carolin Tietze**[3]

[1] Römisch-Germanisches Zentralmuseum Leibniz-Forschungsinstitut für Archäologie: Mainz, Rheinland-Pfalz
[2] Institut für Archäologisches Wissenschaften, Universität Bern
[3] Institut für Klassische Altertumskunde, Christian-Albrechts-Universität zu Kiel

clemens@nevrome.de, martin.hinz@iaw.unibe.ch, ctietze1991@gmail.com

---

1. General Information

- A **repository** with all information and data for this workshop is available at `https://github.com/sslarch/caa2019_hackathon`
- 2-4 **groups** are formed on site according to framework preference and skill levels (*Unconference style*).
- You have **5** hours to work on all tasks, breaks can be taken as you wish. It's not necessary to complete all possible tasks. Work as far as you can go. Most tasks are independent of each other, so you can skip boring ones. This is **NOT** an exam.
- All results must be submitted in one **reproducible report** with all code and plots. This can be rendered from IPython Notebook, Rmarkdown, Latex, etc. or compiled manually. Ideally the report is submitted as a Pull Request to the hackathon repository on github. The file(s) should be added to the `reports` directory.
- The organizers of this workshop are available for questions and advice. They are able to assist you with problems as far as they are familiar with your toolset.

---

2. Dataset

The data for this excercise — `Michelsberg` — are taken from the R package `archdata` (Carlson/Roth 2018). It's a features by types table of abundance data on vessel types in archaeological features of the Younger Neolithic Michelsberg Culture from Belgium, France and Germany by Birgit Höhn (2002). The 109 observations/lines represent individual features (pits, ditches, etc.) from sites. For each feature we have information about 42 variables/columns. These include identifiers, classified pottery type counts, phase attribution and site coordinates.

From ?archdata::Michelsberg:

- id: Unique identifier (categorical, integer)
- site_name: Name of site (categorical, character)
- catalogue_nr: Number in catalogue of Höhn (2002) (categorical, integer)
- feature_nr: Number of the archaeological feature (categorical, numeric)
- to3: Pot/vessel type 3 count
- f4: Bottle type 4 count
- b2: Beaker type 2 count
- to2: Pot/vessel type 2 count
- b3: Beaker type 3 count
- b7: Beaker type 7 count
- kw5: Carinated bowl type 5 count
- vg1: Storage vessel type 1 count
- vg2: Storage vessel type 2 count
- t4a: Tulip beaker type 4a count

- kw2: Carinated bowl type 2 count
- kw4: Carinated bowl type 4 count
- b5: Beaker type 5 count
- t3b: Tulip beaker type 3b count
- f3: Bottle type 3 count
- kw3: Carinated bowl type 3 count
- kw1: Carinated bowl type 1 count
- b6: Beaker type 6 count
- to1: Pot/vessel type 1 count
- b1: Beaker type 1 count
- t3a: Tulip beaker type 3a count
- vg4: Storage vessel type 4 count
- ks2: Conical bowl type 2 count
- ks1: Conical bowl type 1 count
- t2b: Tulip beaker type 2b count
- f2: Bottle type 2 count
- bs3: Globular bowl type 3 count
- t2a: Tulip beaker type 2a count
- bs2: Globular bowl type 2 count
- b4: Beaker type 4 count
- bs1: Globular bowl type 1 count
- f1: Bottle type 1 count
- t1b: Tulip beaker type 1b count
- vg3: Storage vessel type 3 count
- t1a: Tulip beaker type 1a count
- mbk_phase: MBK phase according to Lüning (1967) as an ordered factor with levels $<$ I/II $<$ II $<$ II/III $<$ III $<$ III-V $<$ III/IV $<$ IV $<$ IV/V $<$ Munz $<$ V
- x_utm32n: x coordinate in m; projection UTM WGS 84, zone 32 nord
- y_utm32n: y coordinate in m; projection UTM WGS 84, zone 32 nord

Höhn (2002) recorded pottery vessel shapes from 108 archaeological features (pits, ditches etc.) from 69 sites of the Central European Younger Neolithic Michelsberg Culture (MBK; 4350-3500 BC) following Lüning's (1967) typology. Her correspondence analysis of the abundance data (columns 5 to 39) exhibits a pronounced Guttman effect or arch, suggesting the data set is structured by a time gradient. Recently Mischka et al. (2015) projected an 109th Michelsberg assemblage, Flintbek LA48, a pit with Michelsberg pottery from a North German site of the Funnel Beaker Culture (TRB), as a supplementary row into the existing chronology thereby connecting the relative chronologies of TRB and MBK. The data frame contains as attributes the references for the data, a typological key and the map projection. Note that ambiguous fragments of conical bowls (ks1 and ks2) are assigned as 0.5 to each of the two types resulting also in positive entries suitable to analysis by CA.

Höhn, B. 2002. Die Michelsberger Kultur in der Wetterau. Universitätsforschungen zur prähistorischen Archäologie 87. Bonn: Habelt.
Mischka, D., Roth, G. and K. Struckmeyer 2015. Michelsberg and Oxie in contact next to the Baltic Sea. In: Neolithic Diversities. Perspectives from a conference in Lund, Sweden. Acta Archaeologica Lundensia Ser. 8, No. 65, edited by Kr. Brink et al., pp 241–250.
Lüning, J. 1967. Die Michelsberger Kultur: Ihre Funde in zeitlicher und räumlicher Gliederung. Berichte der Römisch-Germanischen Kommission 48, 1-350.

3. Tasks

All these tasks focus on the technical implementation of data analysis, not the interpretation of the results in the context of archaeological research questions. But feel free to observe and document interesting patterns.

1. Create a table with information about the sites (`site_name`) and the amount of features (`feature_name`) documented per sites. Based on this table: Who many sites are represented in the table by more than one feature?
2. Create a table with information about the total material occurrence: How many individual objects of `to3`, `f4`, ..., `t1a` do we have in total across all sites? The table should not just contain the shortend type ids (e.g. `to3`), but also a human readable type name (eg. `Pot/vessel type 3`). Plot an ordered bar chart based on this table.
3. How much artefacts are documented per MBK phase (`mbk_phases`) in total? Visualise the grouped counts with a time series plot. `mbk_phases` is an ordinally scaled time variable. Reduce your data selection to the 6 sites with the most features (cf. Task 1) and draw the same time series for each of these variables (ideally in a plot matrix).
4. Plot a spatial map of sites based on the coordinate information. The map can be interactive. Add a meaningful background layer. The size and colour of the site markers should reflect the occurence of `Carinated bowl type` (`kw3`) at the site.
5. Run a Correspondence Analysis (CA) of features and pottery type variables (`to3`, `f4`, ..., `t1a`). Remove very rare types and features with almost no finds beforehand. Prepare two result scatter plots: 2D: Dimension1 Dimension2 and point colour according to the MBK phase (`mbk_phases`). 3D: Dimension1 Dimension2 Dimension3.
6. Run a Correspondence Analysis (CA) of sites (!) and pottery type variables. Use only binary presence-absence information for the CA, not object count. Plot the resulting rank order along Dimension1 or along a fitted principal curve on a spatial map.
7. Calculate cultural distance between features (Euclidian distance, Chi-square distance, ...) based on the pottery type variables and visualize the distance network.