Group Assignment

Joris Vincent & Teun van Gils*

January 19, 2026

Welcome to one of the main components of this course: the group project! You will be designing, planning and executing your own research project on a large data set, working in groups of (mostly) 4. We expect you to work on this project together, at all regular course times (Mon-Fri 10:00-18:00). Attendance is still mandatory.

1 The assignment

For this assignment, there is no topic, no research question, no limits (except time and ability). We will give you a very large data set, the rest is up to you – all we require is that you use at least the given data set, and that the project keeps you and your group busy the whole week (but not outside of class hours).

The focus of this project is not on programming per se, but on how you use the tools to answer your research question and how you report on your findings. You should aim for scientifically proper results, and understand how the limitations of the data (spoiler – there are many) limit the conclusions that you draw.

The actual process will be split up into three phases:

- Specification and operationalizing (exploration)
- Implementation and interpretation
- Presenting: conclusion and extrapolation

By this, we simply mean that we would first like you to focus on exploring the dataset, building a pipeline to process the data, and coming up with some preliminary findings that you can report on.

1.1 Phase one: specifying and operationalizing your Research

On Monday morning, we will form groups, and you will start exploring the structure of the data, brainstorming about possible directions. Importantly, you need to also *come up with a team name.*

We will try to check in with each group to discuss what you have been thinking about, and give you some pointers, where necessary. During the rest of the day, you will start actually exploring the data and building the pipeline required to extract the information you are interested in. This also includes setting up some of the project flow, like starting to write the LaTeX report, setting up a Git repository, and dividing tasks.

*prev. Lucie Kattenbroek & Joska de Langen

On Tuesday, you will continue working on this pipeline, and are expected to start seeing some preliminary results. In the afternoon, you can then shift your focus to presenting these findings, and coming up with a detailed proposal for the remainder of the week.

You are expected to **hand in an intermediate"exploratory" lab report and proposal, 1-1.5 page, by 18:00 on Tuesday**. This report is kind of in-between a narrative-focused scientific research report, and a methodology-focused lab notebook.

It should be structured somewhat like a research report with an Introduction section and a Method section, and ideally at least some Results that you can Discuss. Content-wise, it should provide at least:

- your Research Question (RQ)
- Introduction providing Context to this RQ
- initial sketch of your Methodology including
  - Operationalization of your RQ: which variables in the dataset are relevant for your RQ?
  - Data Question (DQ), translating the RQ into those variables
  - What explorations did you try, that worked (in terms of data operations)?
  - what explorations did you try, that didn't work? *Note: this normally wouldn't be part of a Research Report, but for this Proposal, we want a more in-depth lab-reporting.*
- Initial, lab-report style Results of exploration of those variables, to assess the feasibility (enough observations, range of the variables, etc.)
  - What does the dataset need to have, to allow you to answer your DQ(s)?
  - What are the initial results of your explorations? Do they help you assess that feasiblity?
- Discussion of those preliminary results
  - Is/are your DQ(s) answerable with this dataset?
  - What limitations are you already aware of? Are those `fatal`?
  - What are your next steps (in terms of data operations)?
  - What is your time-planning and division of labor for this? Is this feasible?

We will provide feedback on this report at the start of the next day. This report will also be taken into account when determining your group project grade.

1.2 Phase two: answering a question

On Wednesday and Thursday, you will work on *implementing and interpreting* the analysis.

This requires iteratively building on your pipeline, including coming up with the optimal visualizations to present your story. As most of your grade will be determined by this final report, you will likely want to spend a considerable amount of time on detailing your approach, hypotheses and findings, but also on problems and limitations of your approach.

You should focus on answering your RQ as well as possible, rather than branching out more: this means cleaning data over adding new datasets, controlling for existing variables rather than adding new ones, and working on your final paper to justify your choices and explaining the

limitations. Perhaps the only exception would be when, while writing, you realize that there are some obvious limitations that you could address by expanding your approach. This ensures that all work you do during this time actively contributes to the quality of the final results.

Ideally, you will be done analysis results by early afternoon on Thursday.

1.3 Phase three: presenting a narrative

Thursday afternoon and Friday, you will work on the *presentation* of your results. This will take two forms:

- Presentation to the whole class on Friday morning. (everyone's slides are due before start of class, before any of the presentations have started!).
- Final research report, due by 18:00 on Friday.

On Friday morning, you will give a presentation to the whole class, on your research results. This presentation should be about 12 minutes long, we'll cut you off at 16 minutes. Following your presentation, there will be approx. 10 minutes of Q & A from everyone, as well as feedback. All members of the group should present. Remember: present the *story* of your results, not a story about you and your process.

After the presentations, you will have the rest of Friday (i.e., the afternoon) to finalize your research report. We and your classmates will have give some feedback on your presentations, focusing on interpretation, and possibly visualisation, of your results. You should incorporate that feedback in your final report as well. This final report should be structured and styled like a scientific research report; We'll look for a structured introduction section, and complete and detailed methods section, a results section with appropriate and effective visualisations and descriptions of patterns of results, and a thorough discussion section intepreting these patterns and expanding on the implications and limitations.

2 Project scope

While designing your project, your main limitation will be time (as is often the case in life). We strongly recommend choosing a project that is scalable (something that starts out simple but can become as complex as time allows). This way you prevent running out of time, or out of things to do.

Considering you will want to build up the project in complexity, you have to make sure your code is adapted for that. A few tips to do so:

- Write and comment your code so that you group members can follow what's going on without asking you. This makes the process faster and makes it easier for code to be reused.
- Write your code to be flexible. For instance, write a function for something you know you will want to do frequently.
- Share your code around through a medium like Git.
- Check early on whether what you want to do is possible by going through the whole data analysis process at least once: browse with CLI, process with Python, visualise with R. Only after you have confirmed you can do all that with your data, expand your horizon to bigger issues.

If we have serious concerns about your project, we will tell you as soon as we can.

2.1 Statistics

You are welcome to use a variety of statistical methods in your paper, and we will think along with how to execute the statistics in Python or R. However, since this isn't a stats course, you are not required to use any more advanced statistics and we in general won't grade on it either. Be weary that if you end up including statistical methods in your report and use those for your conclusions, that your scientific reasoning should still be sound. Since we also grade on the scientific soundness of your paper, including incorrectly used or interpreted statistics could mean that we have to deduct points for this.

2.2 Quantity

The final product is more about showing your understanding of the data analysis process and less about how much you managed to do. Doing more can never make up for a badly written final report. Quality trumps quantity or complexity. Of course we do want to see a minimum of both, but since what you can realistically manage to get done depends on many different factors, some of which outside of your control, we can't be more explicit on the minimum requirements. Essentially: don't worry too much about whether your project is too simple to obtain a good grade. If you are worried about how much your group is producing or if your question is complex enough, you can always discuss these concerns with us.

What we expect from the final report, is that you address a single research question effectively – that means you can explain why it is an interesting question, why the approach you took makes sense, why the results you obtained can actually answer your question (or, if not, why?), and what the general limitations of your approach are. Anything more than this will only improve your grade if these requirements are fully met. Just like more questions, more results, or more code don't necessarily add to the quality of a project, this also applies to your report: write what you need to for a coherent and complete report, but not more than that. Again, reach out to us if you have questions about the specifics.

2.3 External data sets

We are not requiring you to include external sets. If you can think of an interesting research project that keeps you busy all week using just the DBpedia data set and you report on it well, that is perfectly fine. However, in many cases, we have seen that projects can be made more interesting when an external set is included. We will be around all week to help you process the dataset, but, since we probably also won't have worked with it before, we can't guarantee that we can solve any issues any faster than you.

3 Our role

Because you are delving into a project yourself that we have not completely prepared for you, you are bound to run into problems that you, and we, haven't seen before. This is great, and one of the beautiful parts of programming: pick your project, you'll learn along the way. If you run into any issues, we will happily brainstorm along. If you aren't sure if something is possible, do ask, and we'll delve into it with you. Also make sure to ask questions when you are stuck on a problem for more than 15 minutes; the point of the project is definitely not for you to be stuck. We will happily join you in your projects whenever you're stuck!

The point of this week is to apply the tools you've been introduced to along the way, as opposed to the first two weeks which focused on learning to use these tools. We want your

results to be interesting and informative as much as you do, so we're happy to help you along the way. Our rule of thumb for helping you is as follows: if we think you can figure it out by yourself quite easily, we will tell you that or give you some pointers on how to get there. If we think something you are trying to do is beyond what we have taught you, or beyond the scope of the project, we might either advice you to take a different (simpler) approach, or we will at least sit down and help you through the trickiest parts or – in very exceptional circumstances – might do it for you.

3.1 Feedback conversations

We will try to sit down with you most days of the week:

- Monday afternoon to discuss the general direction,
- Tuesday afternoon to look over initial results and discuss proposals
- Wednesday morning to discuss your proposal and what you will be doing next,
- Thursday to guide final analyses
- Friday after the presentations to guide final steps in your final report.

We have these conversations to give us an idea of where you are at, and for us to point you in the right direction if you are stuck and to prevent you from descending into the depths of programming hell. We may have other intermediate conversations as well, if deemed useful or necessary.

4 Grading

The group project will account for 40% of your course grade. We want you to write your report in LaTeX and hand in your code through Git. The report should be a proper scientific report: it needs to describe your project and its relevance, include a methods section and a discussion, be properly cited and typeset, et cetera.

4.1 Requirements

Your grade will be based on the following components:

- a report with a proposal, handed in by Tuesday 18:00,
- a presentation (12-16min, with 10min Q&A), held on Friday morning (slides due at 09:30)
- and the final paper and code, as handed in by Friday 18:00.

We may also take other observations that we have made over the course of the project, in our meetings or other interactions with you, into account, although the majority of your group project grade will be determined by the aforementioned elements.

When grading, we will consider (at least) the following points:

- Did you use the right tools?
- Did you manage your time well?
- Do you clearly describe your findings?

- Do you effectively communicate the meaning and implications of your findings?
- Are your LaTeX documents logically structured and well-formatted?
- Did you use GIT well? (e.g. multiple commits by different people, with sensible commit messages)
- Is your code well-formatted and documented?
- Do your reports and presentation look polished?

4.1.1 Initial report

More specifically, for the first report and proposal:

- Do you clearly describe the different things you have tried, and why you believe they are or are not feasible?
- Did you come up with a sensible research question, given your exploratory findings?
- Did you come up with a clear and actionable plan to answer this research question?

4.1.2 Final report and presentation

And for the final research report:

- Do you clearly describe your research question and hypotheses?
- Do you clearly explain your approach to answering your research question?
- Are your graphs informative and well-chosen?
- Does your interpretation of the results make sense?
- Do you understand the limitations of your results?
- Does your final report conform to (a common) standard academic style?

4.2 Some useful hints

Some good rules of thumb when writing:

- Make sure you answer the (same) RQ you ask in your introduction (even if the answer is that you could not answer the question, in the end);
- Make sure you introduce all research questions you answer (don't just associatively answer related questions, but make sure to edit your introduction appropriately to make your report into a coherent whole);
- Preferably, you should only try to make a single point per figure, meaning the figure provides a clear answer to a specific question that is relevant to your research;
- Don't just describe what is shown in a figure, but also what this means for your research question;

- Describe your results in your results section, including a basic interpretation of your results, e.g., that a certain trend exists, or that some result clearly answers one aspect of your research question;
- The results section should not be dependent on the placement of the figures, i.e., the text should be readable even if you skip all the figures;
- The discussion section should not introduce any new results: it is meant for discussing limitations and implications of your findings, and embedding them into a broader context (or, optionally, existing literature).

If you have any questions about properly structuring your reports, just reach out to us!

5 The dataset

5.1 The primary set: DBpedia

The DBpedia project takes Wikipedia pages and puts them in usable data sets. We will work with The People of Wikipedia: a data set that contains a large number of pages about individuals. You can find the ontology here. The data set is pretty huge: even separated into different files per year, some recent files are way too long to scroll through manually.

As hard as DBpedia tried - and they got pretty far - the data is not perfect. You can imagine that it can be quite difficult to automate reading a Wikipedia page and putting it into a data entry, considering the large variety of formats Wikipedia pages have, the vastly different type of information presented, and the absence of an index of the pages. So, be warned, some data are incorrect and many (most) fields are empty. Regardless, the data that are left over have a lot of potential for interesting analysis.

5.2 Entries

Each observation (i.e., metadata of Wikipedia entry) is stored as a dictionary, containing a variety of fields. The exact fields that an entry has, depends on the article. For instance, actors and directors might have an entry for `imbdId`, but Olympic athletes would probably not.

An example, the entry for Tom Cruise:


```
{"title": "Tom_Cruise",
 "ontology/activeYearsStartYear": "1980",
 "ontology/spouse_label": ["Nicole Kidman", "Mimi Rogers", "Katie Holmes"],
 "http://www.w3.org/1999/02/22-rdf-syntax-ns#type_label": ["owl#Thing", "person", "Q5", "Person",
 "ontology/birthName": "Thomas Cruise Mapother IV",
 "ontology/birthDate": "1962-07-03",
 "ontology/religion": "http://dbpedia.org/resource/Scientology",
 "ontology/spouse": ["http://dbpedia.org/resource/Nicole_Kidman", "http://dbpedia.org/resource/Mi
 "ontology/birthYear": "1962",
 "http://www.w3.org/1999/02/22-rdf-syntax-ns#type": ["http://www.w3.org/2002/07/owl#Thing", "http
 "ontology/occupation_label": "Tom_Cruise__1",
 "ontology/birthPlace": "http://dbpedia.org/resource/Syracuse,_New_York",
 "http://www.w3.org/2000/01/rdf-schema#label": "Tom Cruise",
 "ontology/occupation": "http://dbpedia.org/resource/Tom_Cruise__1",
 "ontology/religion_label": "Scientology",
 "ontology/birthPlace_label": "Syracuse New York"},
```


A couple of useful fields that most entries should have:

- `title`: the title of the entry/article
- `http://www.w3.org/1999/02/22-rdf-syntax-ns#type_label`: the label(s) that this entry got assigned by the DBPedia curators. There is a whole syntax and tree of labels to this, which you can find at that URL in the key. This field contains a list of strings, some of which might be useful (e.g., `actor`, `fictional character`).

5.3 Files

The data is provided to you in JSON files. Each observation – entry in the data set, i.e., individual article – is on a separate line. So, the entire data set would be about 1.5 million lines. If this

was in a single JSON file, it would be approximately 2 Gb. To load that into your RAM could be problematic.

Instead, we have split the data set into 26 separate JSON file, split, *very roughly*, alphabetically by first letter of `title`. Thus, `A_people.JSON` would likely contain an entry for Aaron Burr, but perhaps also the Archbishop of Canterbury. There's some more weirder sorting problems going on too – all to say, don't rely on the filename to be super helpful in filtering data.

5.4 Secondary data sets

The potential of the DBpedia set becomes even clearer when you imagine combining it with an external data set. For this assignment, we strongly recommend thinking about what kind of external data you think would be cool/interesting to include. However, if you decide to do this, make sure of the following two things: 1) still use the DBpedia dataset in a non-trivial way, don't just sidestep problems you encounter by using an external dataset instead; 2) don't be overambitious, but start small, either with just the DBpedia dataset, or possibly with both datasets in parallel, and gradually build out your question. This gets to a point we will reiterate a lot: it's better to do very little, but do it very well, than to be halfway through something very complex and ambitious. Realize that we don't grade on the complexity of the question, but rather on your approach to answering it and the quality of the report.

Finding the data is a non-trivial task. Not everything can be found everywhere. It is likely that combining the two data sets means having to process either set to suit the other. Exactly how to do that and executing it is an interesting exercise. Regardless, your project is likely to be much more interesting if you combine two data sets. If you decide to go for it, we will be around all week to help you out along the way.

6 Recommended approach

In the first phase of the project, your main goal is to explore the dataset, and come up with an *answerable* research question. This means you will have to figure out fast whether the dataset *can* answer your question. The preliminary research proposal should convince us (and you) that you have a research question that you can reasonably answer with the data at hand and in the timeframe. This is not trivial, given the structure and messiness of the data, so how should you go about this?

1. BEFORE you open the dataset itself:
   (a) FAMLIARIZE yourself with the variables in the dataset. Take time to explore the "headers": what variables are (in principle) available?
   (b) BRAINSTORM some possible research *directions* or topics; not yet actual research questions, just generally what variables seem interesting?
   (c) TARGET a specific POPULATION. A good way to narrow down the research topic is to think about a specific population yo may want to study, e.g., musicians, fictional characters, athletes, politicians, actors, etc. This is not to say that you *should* pick a specific population – your target population may well be "people" (as sampled by Wikipedia...) – but then you need to look at some variables that are relevant for a broad population.
   (d) ESTIMATE how many observations actually have data for the variables you are interested in! There are lots of variables in this dataset – but many of them are present naturally only for a (small) subset of the observations. For example: "height" is often recorded for athletes, but not for politicians....
   (e) PLAN how you would subset for your target population. You will have to determine for each observation if it belongs to your target population or not. This may not be straightforward, as there is no single variable that indicates this directly. What fields/variables in the dataset seem like they may contain this information? It may be more than one variable, or some combination of variables as well.
2. PARTLY explore the dataset.
   (a) The full dataset (all JSON files) is quite large. It is tempting to just read all files into Python and combine them into one big dataset – this will probably crash. It is too much data to fit into the working memory of your computer. So, instead, you will have to do some initial filtering/subsetting of the data, per file. Then, you can combine the subsetted data from different files into one manageable dataset.
   (b) REFINE (somewhat) exploration for one file. Focus on writing the code to do this filtering/subsetting for one file first, and when you're happy with the result for that one file, only then worry about applying it to all files.
   (c) EVALUATE feasibility first based on part of the dataset. Can you actually extract your target population reasonably well? Do you have enough observations with data on the variables you are interested in? Obviously if you focus on only one file first, you'll have to extrapolate from one file to the others This should be reasonable, the sampling between files is pretty evenly distributed. Thus, if you find that one file only gives you a few relevant observations, you can reasonably expect that the other files will be similar...

(d) DON'T OVERDO the initial filtering and subsetting. To assess feasibility, it probably doesn't have to be perfect. You just need a rough idea of how many observations you can expect to have for your target population, and how many of those have data for the variables you are interested in. There will likely be some more complicated, refined filtering later on, but make a rought estimate how much that will affect the number of observations.

(e) DO SPOT CHECK of your initial filtering. Think of some people that you expect to be in your target population (e.g., 'Barack Obama' for politicians), and some that you expect not to be in your target population (e.g., 'Barack Obama' for artists). Check whether your filtering/subsetting code actually includes/excludes these people correctly. Obviously, don't blindly rely on this either: just because you can find some people, doesn't mean you can find your whole target population correctly.

3. DEFINE your research question.

(a) DEFINE your target population. Have a sense of how many observations you can expect to have for your target population in the full dataset.

(b) DEFINE your variables. What fields will you be looking at? Does (almost) everyone in your target population have data for these variables?

(c) FORMULATE your research question. What do the variables you selected allow you to investigate? What are the more abstract concepts that they represent?

(d) TRANSLATE your research question into a set of data questions. For the variables you selected, what is the specific questions in the terms of those variables? Start with the simplest data questions (overall group summary, simple comparisons), before thinking about more complicated data questions (subgroups, interactions, additional variables).

(e) EVALUATE feasibility again, now for each data question. Based on your initial exploration, do you think you can reasonably answer each DQ? Do you have enough data? If you cannot answer some specific DQ, is that one critical to your RQ? Is there an alternative DQ that you can answer instead? Are there additional variables you could include to help answer your DQs? Are there additional variables missing that would massively influence your ability to answer your DQs?

4. ITERATIVELY DEVELOP

(a) PLAN the first data question. What is needed to answer it? What is the quickest filtering/subsetting that youn can do? What visualizations would be helpful to answer it? What data operations are additionally necesary to answer it?

(b) IMPLEMENT the plan for the first data question. Get to a quick and dirty visualization / result as fast as possible. An overview and initial answer is the priority, not perfection.

(c) EVALUATE the result for the first data question. Was it feasible? Does it make sense? Does it help you answer your overall RQ?

(d) Only then, move on to additional data questions.