Project Requirements

Goal: discuss the BDV course project requirements, with emphasis on creating original, scientifically valuable insights from big data using visualizations.

For the BDV project, you should start from data that is potentially *insightful*, and *exciting* to present.

The **goal** is to craft a complete "story" centered around visualizations: what does the data "show" us, that was otherwise not (so) obvious.

Datasets: I will provide several suggestions in class. Additionally, you are encouraged to seek out your own. The data may represent real-world or synthetically generated (e.g., simulation) measurements. Of course, the real-world data does not have to be collected by you, so feel free to scan public repositories. I encourage you to use data that can be modeled as a complex network (even though not mandatory), and to apply some of the graph measures and computational tools introduced in class.

Teams: Students are highly encouraged to work in teams of two (2). Both students must be able to present the final project and answer questions.

Try to include concepts like averages, standard deviation, correlations, normal/power-law distributions, communities, but most important, different types of plots in practice.

Data acquisition

Many online data sources will have an API which allows querying and downloading the data in a targeted way. Other sources will provide raw data (e.g., csv) that require processing, either manually or through a program you will write.

Example data: Kaggle, Mendeley Data, Dataverse, Goodreads, IMDB, Global comics database (http://www.comics.org/), scientific publication database (dev.mendeley.com), Google Trends, NOAA climatic data, US Census Bureau, European Union Open Data Portal, Data.gov.uk, Canada Open Data, The CIA World Factbook, Healthdata.gov, UNICEF statistics, Amazon Web Services public datasets, Facebook Graph, Data Market (economics), Gapminder, Google Finance (stock markets), world events on GDELT, sports, medical data, HumanProgress etc. A long list of repositories is available on Forbes. etc.

(Optional) **Network modelling**

If a network model is appropriate for your data, be careful that most datasets will admit more than one representation as a network. Some representations will be more or less informative than others. Figuring out the "network" that is buried in your data is part of your project.

Remember graph != network. A graph is a mathematical model of relationships $G=(N, E)$. A network is a model of a real-world complex system, under the form of a graph.

Documentation & Visualizations

All findings must be structured in a documentation. I encourage the usage of Overleaf (online Latex editor) for the creation of a **max 10-page** document (min 5 pages).

The structure of the document should resemble a scientific paper (e.g., IEEE double column [Word], [Overleaf]):

- Title page and Introduction (~1 page): title, authors, affiliation, introduction (do not leave a blank page and start on the second page). What is the context of the project? What is the goal of the project?
- Data (1-2 pages): about the data (acquisition, format, meaning, processing).
- Results (2-3 pages): include at least 3 figures representing insights into your data (e.g., line, bar, box, scatter plot, histogram; or more advanced plots using Folium, waffle, or word clouds). Provide captions for your figures/tables and reference them in your documentation (e.g., Figure 1 depicts …). Do not interpret the results in this section, simply present them in detail.
- Discussion and Conclusions (1-2 pages): discuss what the results mean. What is the story and the novelty of your finding? Bonus: include a comparison with some state-of-the-art studies on the same topic.
- References: provide a list of several references supporting your documentation (e.g., data source, paper where the data or used methodology was first presented, similar studies, tools used). For Latex users, use BibTex format for citations from Google Scholar (Search topic → Cite paper → BibTex → Copy-paste to your *.bib file, then \cite{} in paper).

Deliverables (W14)

The project documentation (5-10 pages) in electronic format. Should resemble a paper, like an IEEE one or two column conference paper, Springer CS proceedings paper, Elsevier journal templates, Scientific Reports template.

Informal, online hands-on presentation of your workspace with me and your colleagues, during W14/session. No PowerPoint needed; just talk, and show the results in real-time, e.g., in Gephi, Jupyter notebook, R Studio, Excel. I will read the documentation privately, so the focus will not be the content of the documentation, but rather a hands-on dialogue focusing on your work.

The code/script and the link to the dataset(s) included separately in a text file to be uploaded on the Virtual Campus.

Project Grading (1 + 9 points)

- Gather & use data in an insightful and correct manner – 1p
- Use appropriate methods & tools to analyze the data – 1p
- Extract insights from your data under the form of visualizations – 3p
- Documentation (content + formatting, structuring, quality of writing) – 2p
- Presentation & task assignment in teams – 2p