

Telecommunications Datathon

Introduction

The Energic COST action aims to demonstrate the potential of Volunteered Geographic Information such that data, generated by a wide range of participants ranging from authoritative bodies across scientists to individual citizens, can be used to provide information relevant to scientific, societal and policy in a European context. The objective of activities within the Energic Datathon is to allow anyone to participate in our activities, and demonstrate the potential of transformations of data to knowledge. To lower the barrier to entry for participants each Datathon task provides:

- a description of the underlying motivation for the task
- sets out an initial set of questions that might be explored through the data
- gives access to some prepared data and suggests potential additional sources, and
- suggests potential tools and methods which might be used in the task.

However, these guidelines are only intended to give a starting point to the activity, and we encourage you to be as creative as possible. Entries to the datathon will be judged by a panel of Energic members, and the best will be invited to present their results at the Energic closing meeting in London.

Specific Introduction Including Overarching Objectives

The telecommunication data can be a very rich piece of information for sensing the cities. Telecom Italy made available as open data an interesting dataset covering two months (November/December 2013) and area of Milan. The data are available here <https://dandelion.eu/datamine/open-big-data/> . They correspond to a big amount of data, therefore it is suggested to use big data technology for their managing.

In this datathon we propose to analyse the data considering other data sources either VGI or authoritative.

Available Datasets And Additional Potential Data Sources

VGI

- OSM
- DBPedia
- Geonames
- Twitter
- Facebook
- Flickr
- etc.

Authoritative Data

- Telecom Italia Big Data Challenge (<https://dandelion.eu/datamine/open-big-data/>) : Weather Station Data, Air Quality Data, Telecommunications Data (Received SMS, Sent SMS, Incoming Calls, Outgoing Calls, Internet traffic), geolocalized tweets.
- Municipal data (<http://www.geoportale.regione.lombardia.it/download-dati>)
- Land coverage (<http://www.globallandcover.com>)
- etc.

Potential Questions to Be Asked of the Data

Potential questions include, but are not limited to, analysing spatial and temporal distribution of population, taking into account the day of the week (working or non working day), weather condition or other available information, like, for instance, the land use. Or considering the communication from and to foreign countries (frequency, the most popular areas, the most connected countries and so on). Or evaluating this data with respect to the VGI available information (like for instance detecting from telecommunication data and VGI the most popular locations in the city).

Part I:

- read the Telecom data (T) and, in case, put them in an appropriate DBMS.
- find other data, user generated content, volunteer and involunteer, (UG) and authoritative (A), with the same spatial and temporal coverage (Milan city and November/December 2013); some data are already provided on the website of the Telecom Challenge.
- Eventually put UG and A in the same DBMS
- Clean the data, detecting possible outliers

Part II:

- Verify which are the questions you want to answer
- Research the relevant scientific literature related to the previous point
- Plan your research (methodology and output that you want to obtain)

Part III:

- process the data using the proposed methodology and suitable tools (GIS, geospatial libraries, statistical libraries, etc)

Part IV

- validate your results
- produce relevant maps, tables, graphs, etc and a short report.

Possible Methods and Tools

Data cleaning and aggregation

Basic spatial analysis with GIS, geospatial libraries, statistical libraries, etc.

Reporting your results

You should prepare a report of your results which explains briefly:

- The data and methods you used (and provides links to these such that your work can be reproduced)
- Interprets your results, concentrating on what you learnt through the datathon and linking to the questions set out above
- Emphasises challenges in carrying out the datathon
- Illustrates the originality and novelty of your approach
- References any external sources you used to help you complete the task
- A 2 minute video pitch presenting your report

Your report should be prepared as a self-contained set of HTML pages which can be accessed by the judges and uploaded to the Energic website after the challenge. All content on the website should be licensed CC-BY-SA (where you use data sources covered by other licenses you should provide tools and access to these and make clear any limitations in their use).

Judging criteria

A panel of Energic members will judge the quality of entries to the Datathon and select the best examples for presentation at the final Energic meeting in London. The following criteria will be used in judging entries:

- Overall quality of the entry to the datathon
- Originality and novelty of the approach taken
- Quality of the description of the data and tools used, especially with respect to reproducibility
- Soundness of the approach taken
- Potential scientific, societal and policy impacts of the results
- Quality and engagement in the video pitch

Information for organisers

The Energic Datathon is open to anyone. However, it will be most fun, and probably also most productive for small groups (typically 3-4 people). The tasks have been designed such that they can be carried out by groups with different levels of skills, ranging from basic spatial analysis using standard GIS to creation of more complex workflows using programming skills. We estimate that typical time investment for a Datathon task should be of the order of 12 hours - however, it is of course up to participants how much or how little time you invest. The only hard rule is our deadline for submissions of **31.07.2016**.

There is no need to register for the Datathon, just submit your report to ross.purves@geo.uzh.ch by the deadline. However, we'd like to know that you're taking part, so feel free to drop us a mail telling us who you are, how many of you are participating in which challenges, and whether or not others are welcome to join you. Please Tweet about the event using the HashTags #Energic and #Datathon.

Some useful information about running datathons events can be found at:

- <https://hackathon.guide/>
- <http://guide.mlh.io/>

Contact information

For this particular datathon:

Maria Brovelli maria.brovelli@polimi.it

For the Energic Datathon challenge:

ross.purves@geo.uzh.ch

f.o.ostermann@utwente.nl

r.l.g.lemmens@utwente.nl