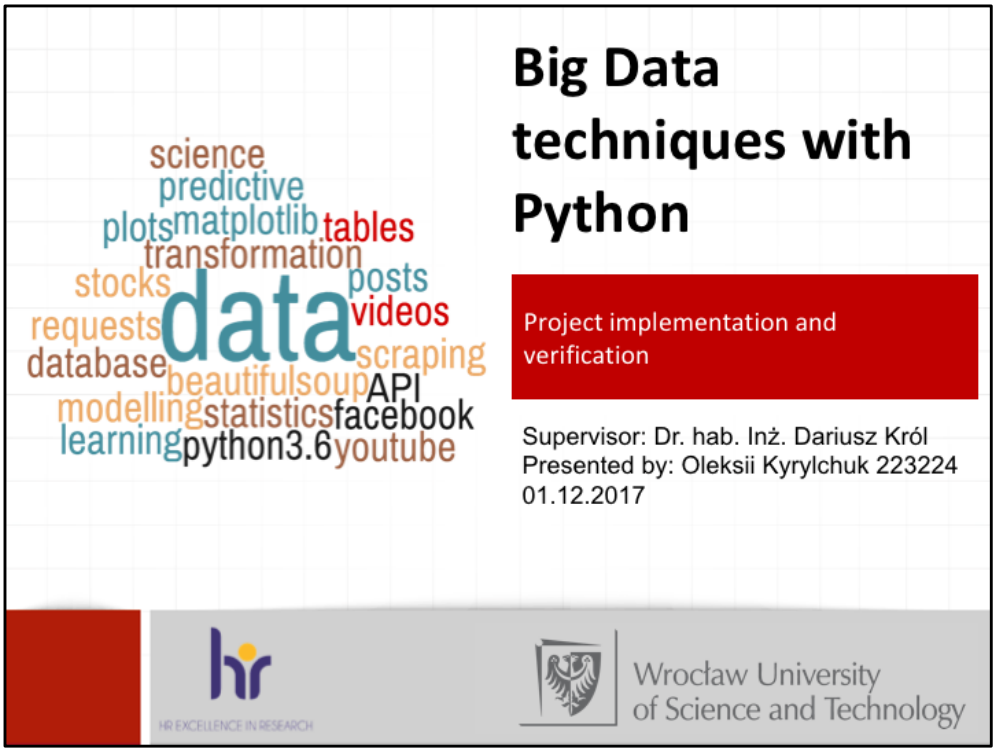
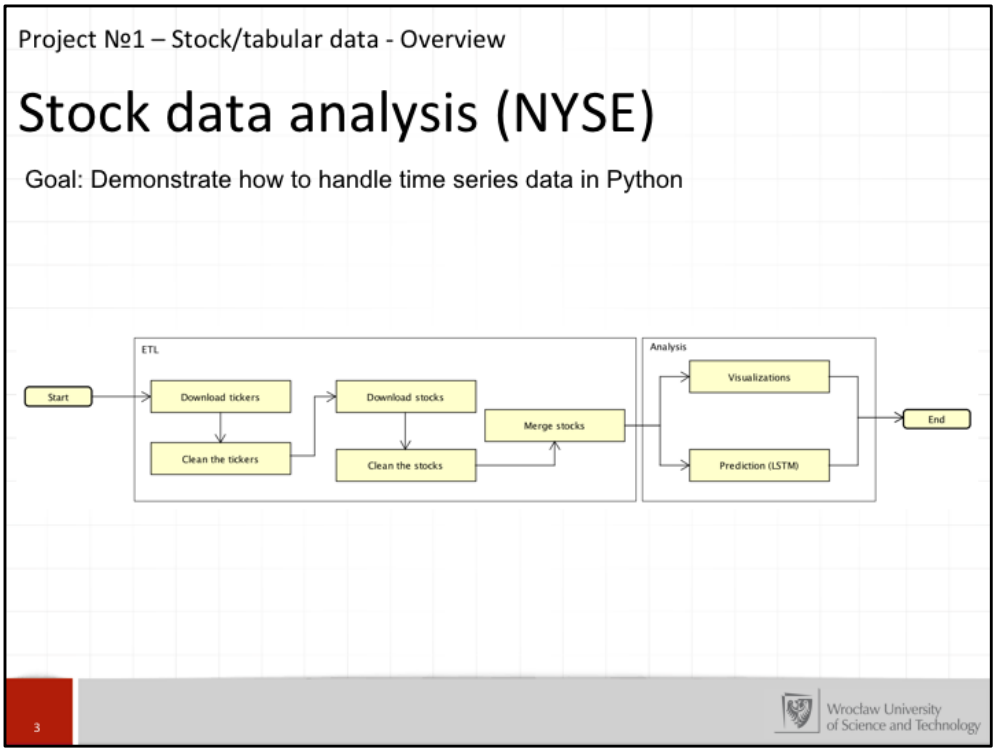


Intro

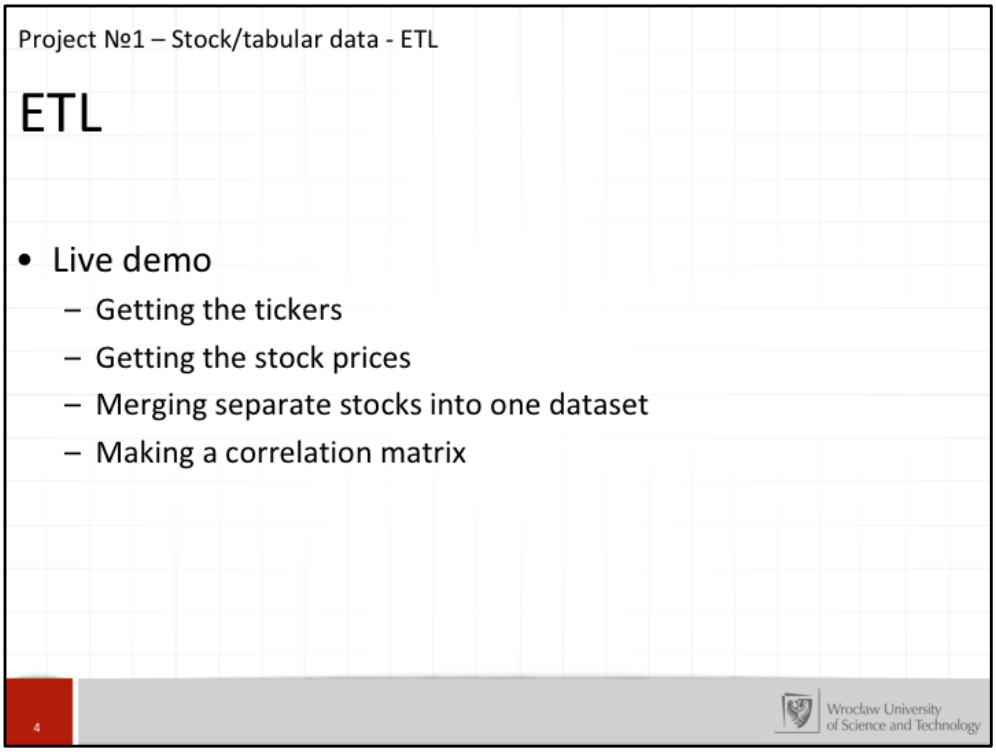
Hello, my name is Oleksii Kyrylchuk and my topic is “Big Data techniques with Python”. My supervisor is Dr. hab. Inż. Dariusz Król.

Today I’ll show you what I’ve managed to do in my project(s) so far, and that includes both different plots as well as just regular etl code.

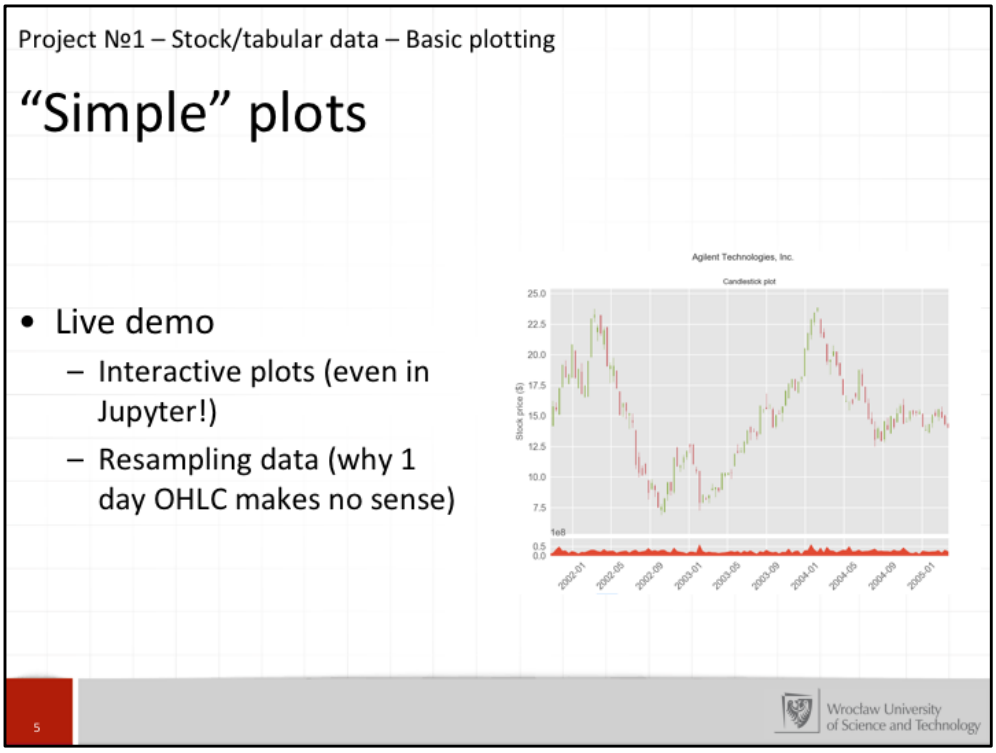
Project №1 –
Stock/tabular data



The goal of my first project was to analyze the stock data, so let's take a look at different parts of its implementation.



Firstly, I had to get the data from some resource, which in my case was Yahoo auctions. As I'll explain in the demo, I get individual stock data from them and then clean/merge it after I finish downloading everything. *Live demo*

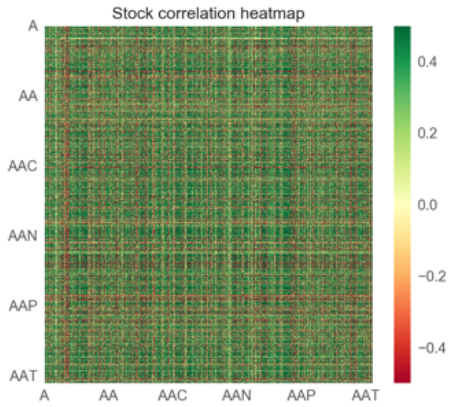


The first type of plots that I made in this project are the relatively simple plots that you see everywhere – and being simple doesn’t mean that you don’t have to learn how to make them or anything like that, they’re still pretty useful. In the live demo, you’ll be able to see how to make interactive plots possible even when you’re using a Jupyter notebook, and learn a basis of OHLC and moving average plots. *Live demo*

Project №1 – Stock/tabular data – More complex plotting

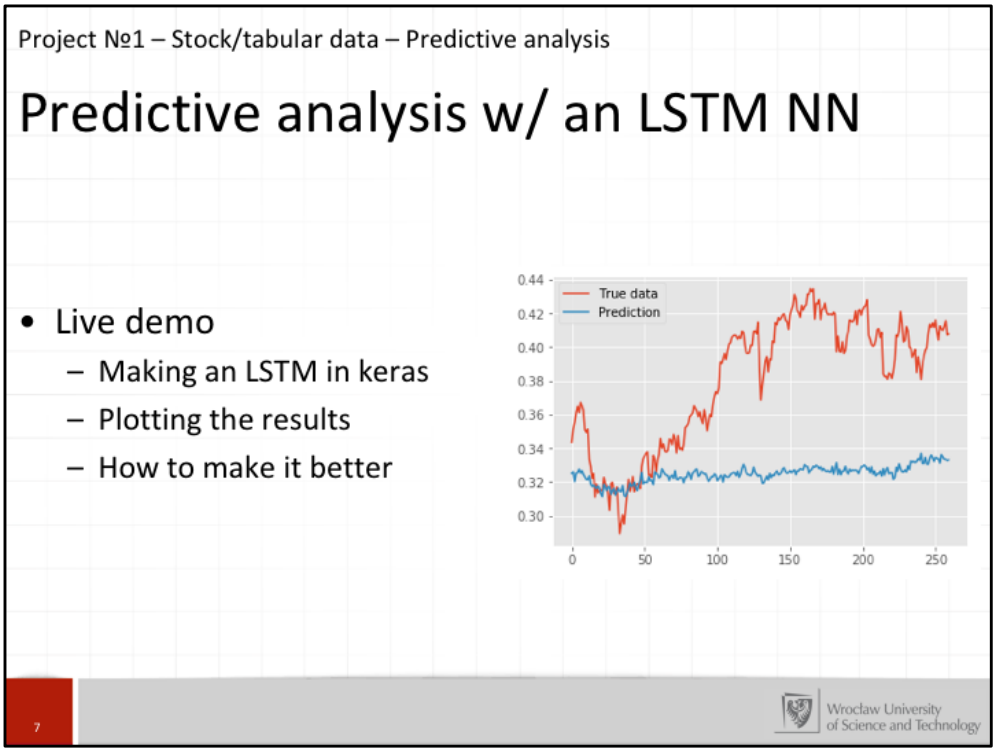
Correlation matrix

- Live demo
  - How to make this plot make more sense (in practical terms)
  - Fixing the ticker labels

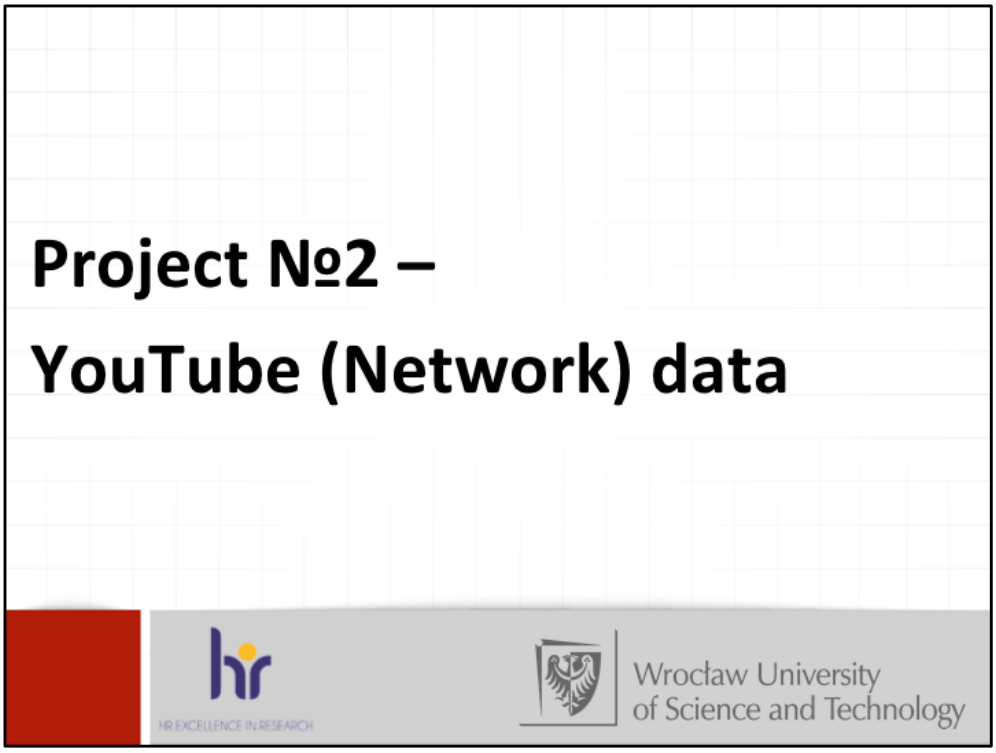


Wrocław University of Science and Technology

In here, I'll show you how I approached more complex plots, which required overriding some matplotlib classes and also show you how to make something like this more comprehensible. *Live demo*

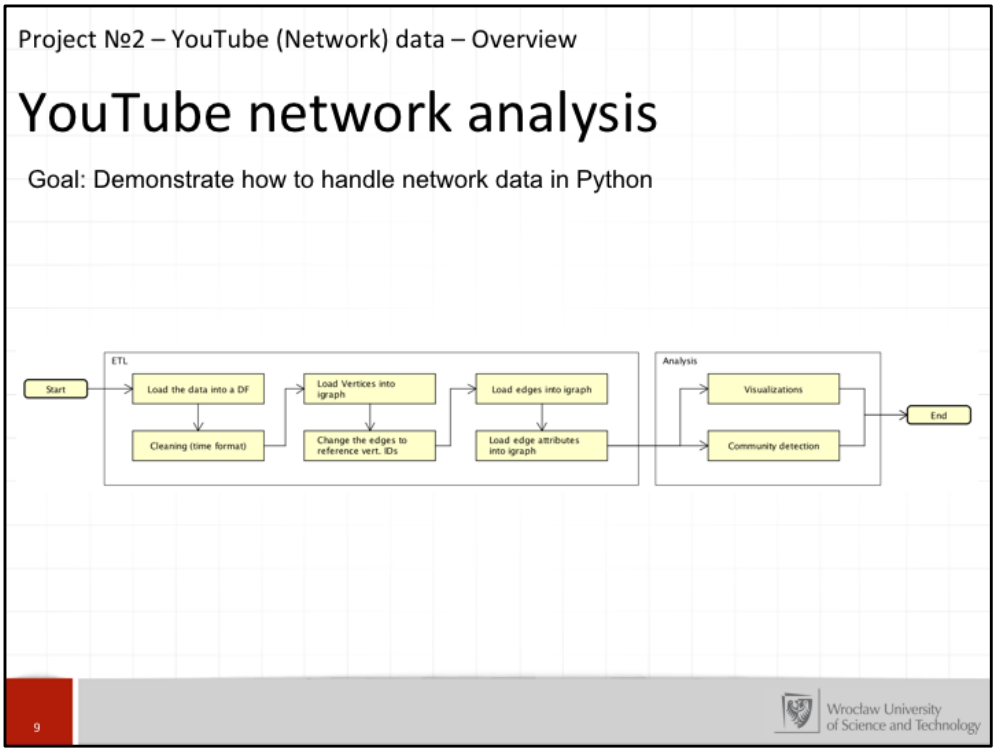


Machine learning is also a big part of using data of any kind, and in my project I've created a LSTM using keras to try and predict a data from a single stock by using the entirety of a stock market a month ago as an input. You'll be able to see the model structure and more prediction plots in the live demo. *Live demo*

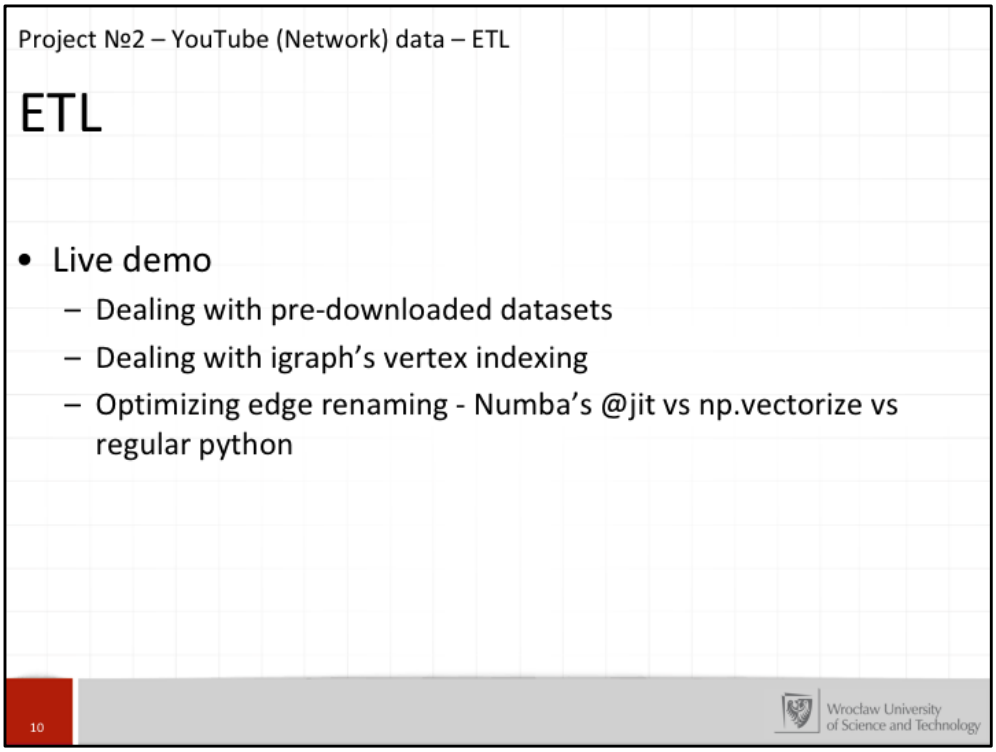


As I've clarified with my supervisor, my goal is to write a manual/guide on how to analyze data with Python – that's why I have 3 different subprojects, each working with a different data type.

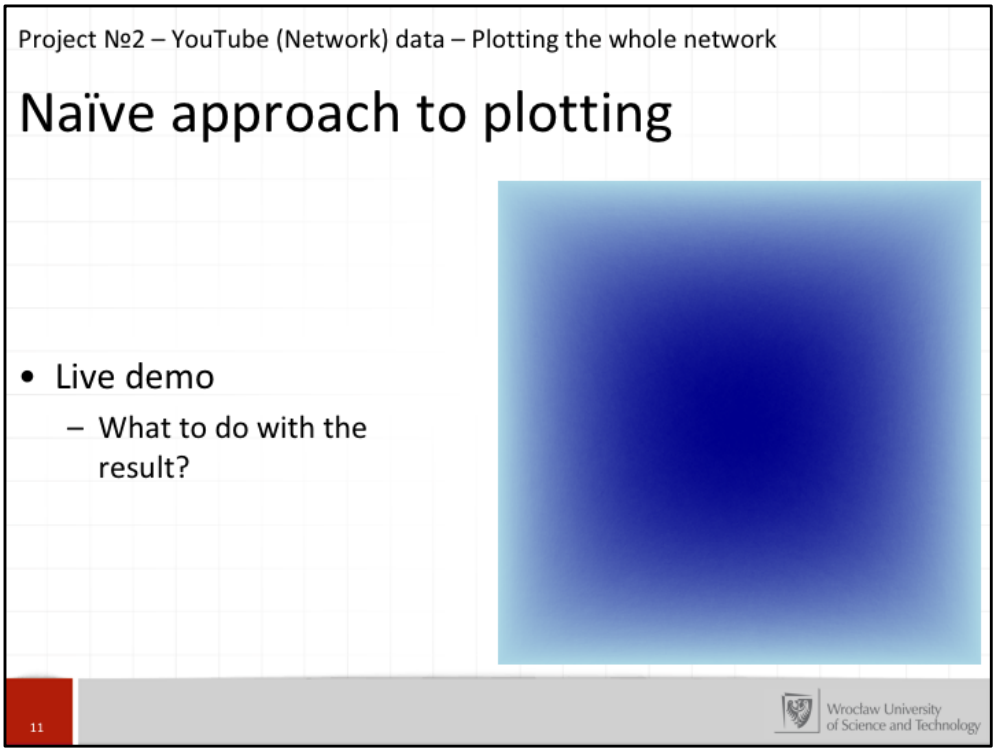
So in this section I'll introduce you to the experiments and their current structures.



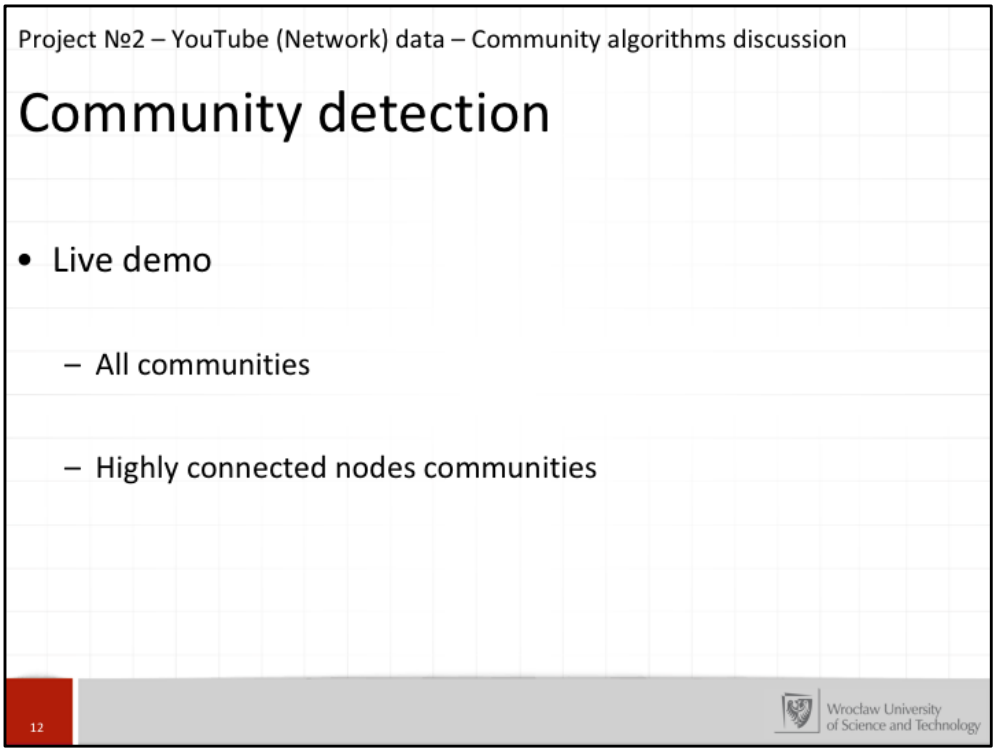
My second project covered analyzing network data, and I chose youtube social network as my data source.



ETL was much simpler in this case, since I've used the already prepared dataset published on KONECT, but I still had to change the edge naming to fit igraph's standards and do something with the date being represented by seconds.



In the first plot, I tried to be naïve and just plot the whole network, which (as you can see here) didn't go too well. That's why I decided to split my network into different communities before trying to plot something again. *Live demo*

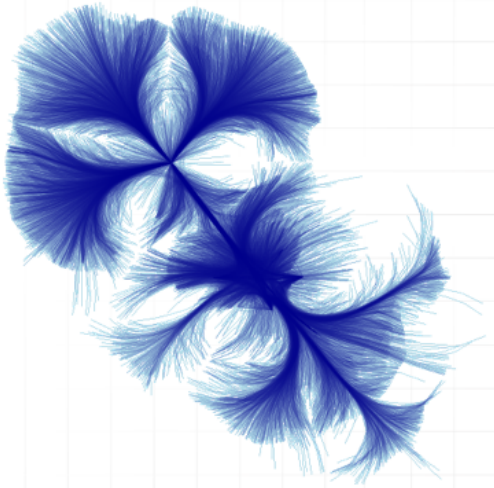


I've decided to do two different community splits to showcase two different plotting libraries – datashader and igraph itself. For the first one I've used two different community detection algorithms ot find communities and communities of communities in the graph, while for the second one I've firstly selected high-connectivity nodes and then ran a community detection algorithm on them.

Project №2 – YouTube (Network) data – Datashader live demo

Datashader plotting – All communities

- Live demo
  - Vertex positioning
  - Edge bundling
  - Coloring communities



13

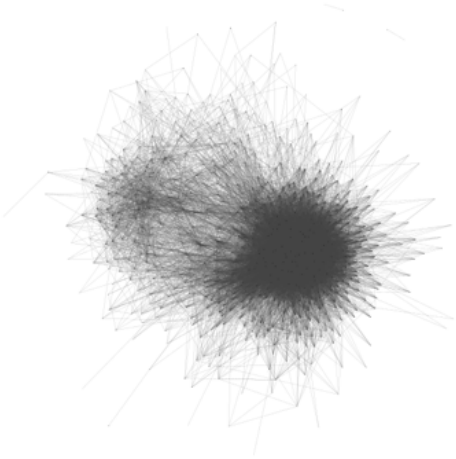
Wrocław University of Science and Technology

And now I can show you the datashader part of my implementation, and discuss what you can do to make your plots look nicer by using it. *Live demo*

Project №2 – YouTube (Network) data – igraph live demo

igraph plotting (Highly conn. nodes)

- Live demo
  - Making sense of this black hairball
  - Giving vertices weight with Google's PageRank
  - Plotting communities (weighted and unweighted)



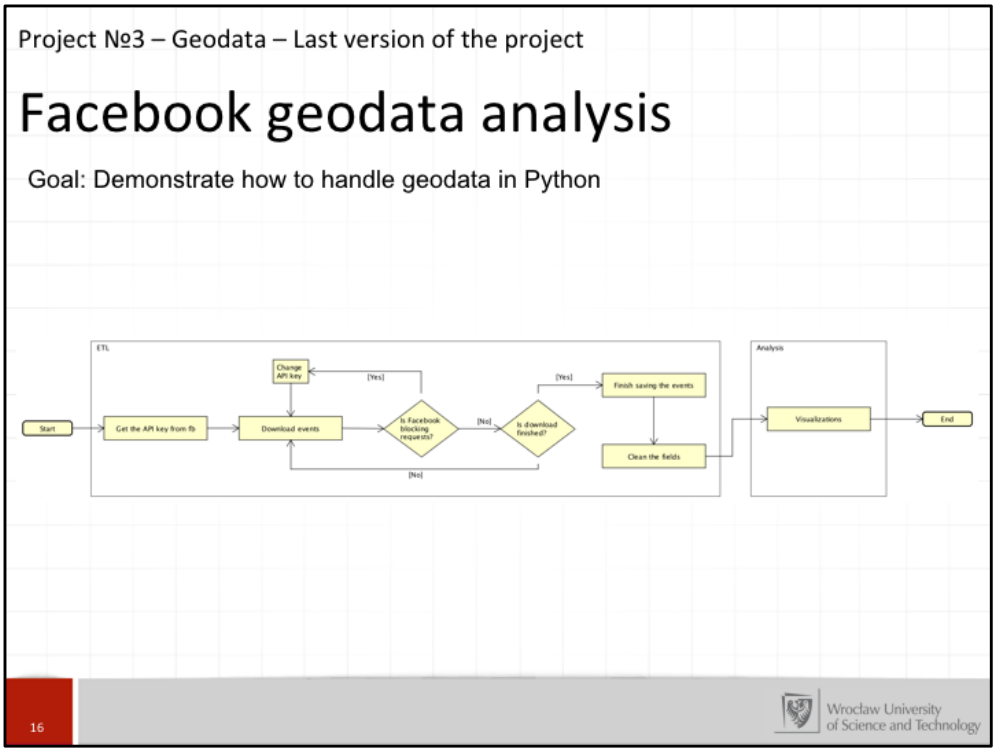
Wrocław University of Science and Technology

And now I'll show you how I've used igraph to make what you see here into something that makes more sense. *Live demo*

Project №3 – Geodata

HR EXCELLENCE IN RESEARCH

Wrocław University of Science and Technology

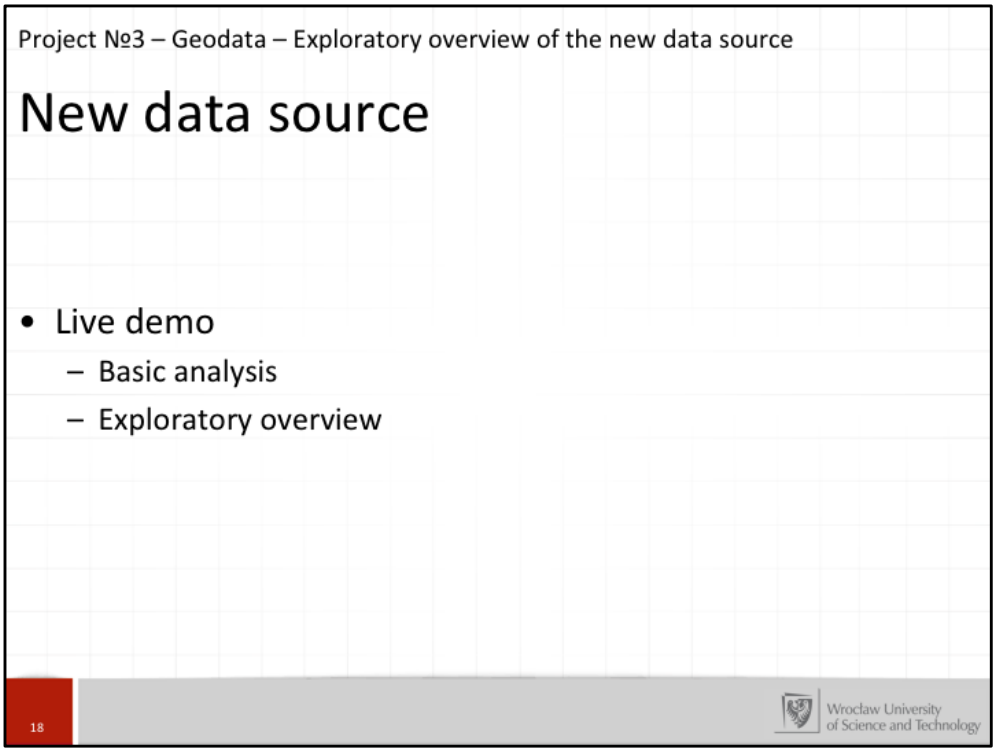


My third project was connected to facebook data for a long time, but we with my supervisor recently discovered that Facebook has a pretty strict no-open-data policy, which means that I shouldn't be able to publish the data that I gathered using their API. (Which is a bit strange, since there are still some publications based on data scraped from facebook's api, but we decidede to not take any risks)

Problems

- So, it turns out that using their data isn't really welcomed by Facebook…
- What can we do? Use other data source!

Some data, that 100% isn't taken from any company and is randomly generated



In this section, I'll show you some basic plotting that I made using my new geodata sources, one for country-level data and one for a more granular level of data. *Live demo*

Thank you for your attention!