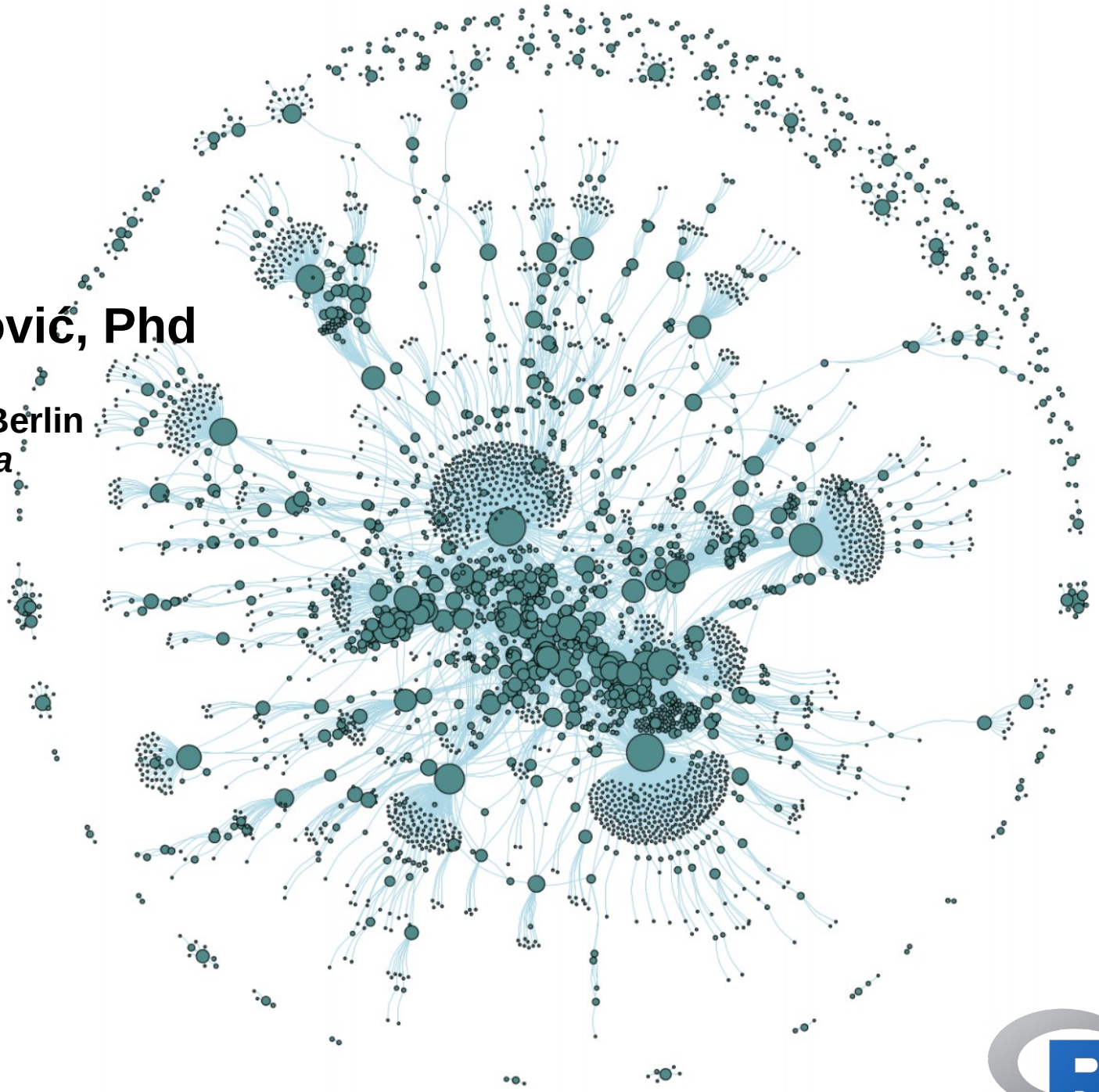
Semantic Web Technologies and Wikidata from R

Goran S. Milovanović, Phd

Wikimedia Deutschland, Berlin
Data Scientist for Wikidata
DataKolektiv, Belgrade
Owner



DATAKOLEKTIV

WIKIMEDIA DEUTSCHLAND

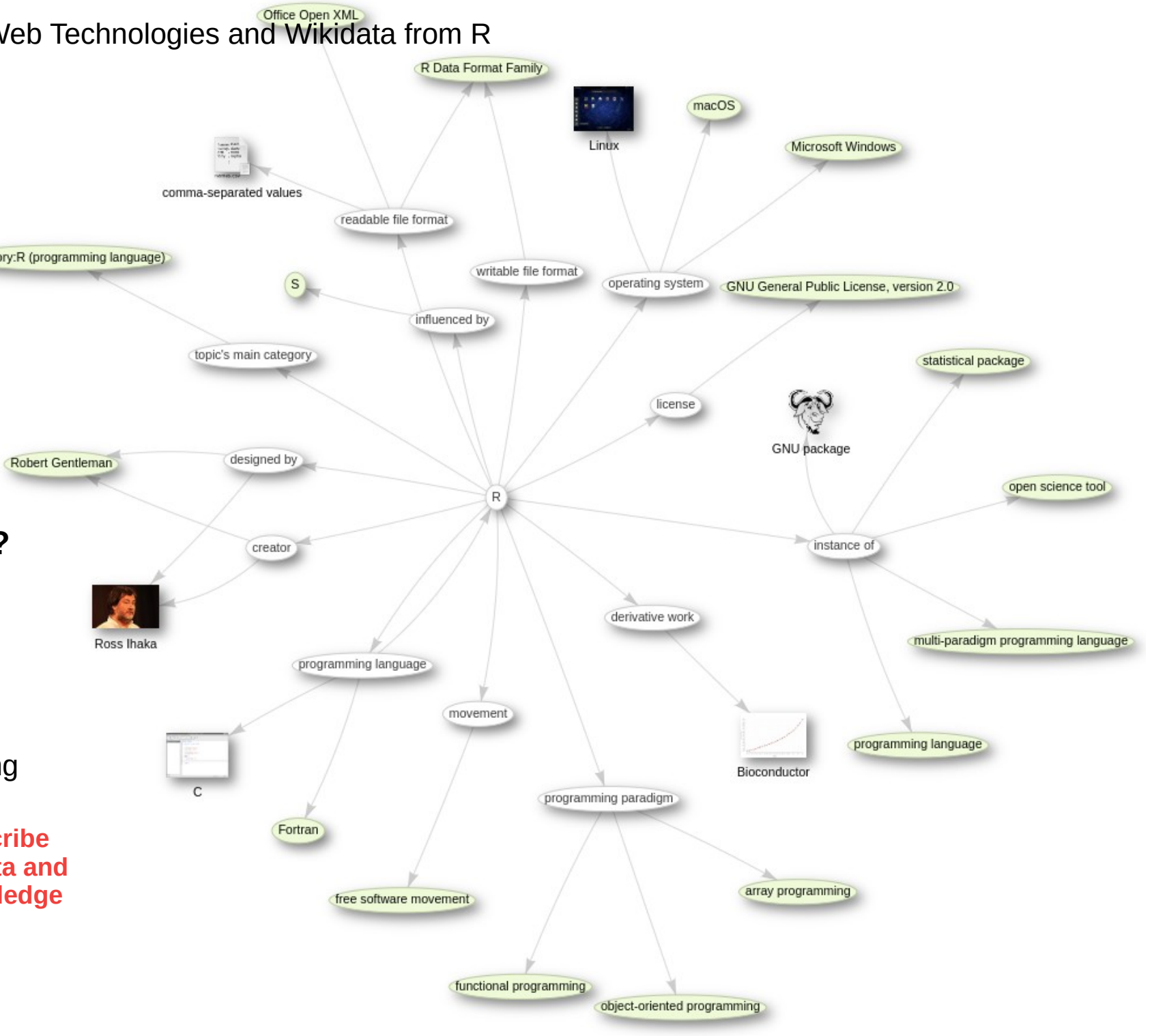
WIKIDATA

R (**Q206904**)

Q206904 (R) is
P31 (instance of)**?**

Q206904 (R) is
P31 (instance of)
Q9143 (programming language).

(triplet: a unit to describe knowledge in Wikidata and other semantic knowledge bases)



How to access

Method 1: SPARQL query against the Wikidata Query Service (WDQS)
https://query.wikidata.org/

Examples

All programming languages:

```
SELECT ?item WHERE {
  ?item wdt:P31 wd:Q9143 .
}
```

1418 results

All functional programming languages:

```
SELECT ?item WHERE {
  ?item wdt:P31 wd:Q9143 .
  ?item wdt:P3966 wd:Q193076 .
 }
```

73 results

R Notebook to learn from:

RWikidata_RLadies20190911.Rmd

GitHub: DataKolektiv's R-Ladies_Belgrade_20190911 repo:
https://github.com/datakolektiv/R-Ladies_Belgrade_20190911

How to access

Method 2: Wikidata MediaWiki API
https://www.mediawiki.org/wiki/API:Presenting_Wikidata_knowledge

Examples

```
https://www.wikidata.org/w/api.php?action=wbgetentities
&ids=Q180736&props=labels&languages=en&sitefilter=wikidataw
iki&format=json
```

R Notebook to learn from:

RWikidata_RLadies20190911.Rmd

GitHub: DataKolektiv's R-Ladies_Belgrade_20190911 repo:
https://github.com/datakolektiv/R-Ladies_Belgrade_20190911

How to access

Method 3: {WikidataR} R Package

https://cran.r-project.org/web/packages/WikidataR/vignettes/Introduction.html

Examples

```
# - Retrieve the Wikidata item: Milano (Q490)
item <- get_item(id = 490)

# - retrieve all claims for Q490
claims <- names(item[[1]]$claims)
head(claims, 20)

"P2924" "P373"  "P1225" "P1082" "P1667" "P625"  "P910"
"P3365" "P349"  "P268"  "P1791" "P242"  "P1036" "P1334"
"P227"  "P2046" "P6"
"P1792" "P1448" "P395"

# What is P2924?
# UseWikidataR::get_property()

prop <- get_property(id = 'P2924')
prop[[1]]$labels$en$value

[1] "Great Russian Encyclopedia Online ID"
```

R Notebook to learn from:

A_WikidataFromR.nb.html

GitHub: DataKolektiv's MilanoR2019 Repository
https://github.com/datakolektiv

How to access

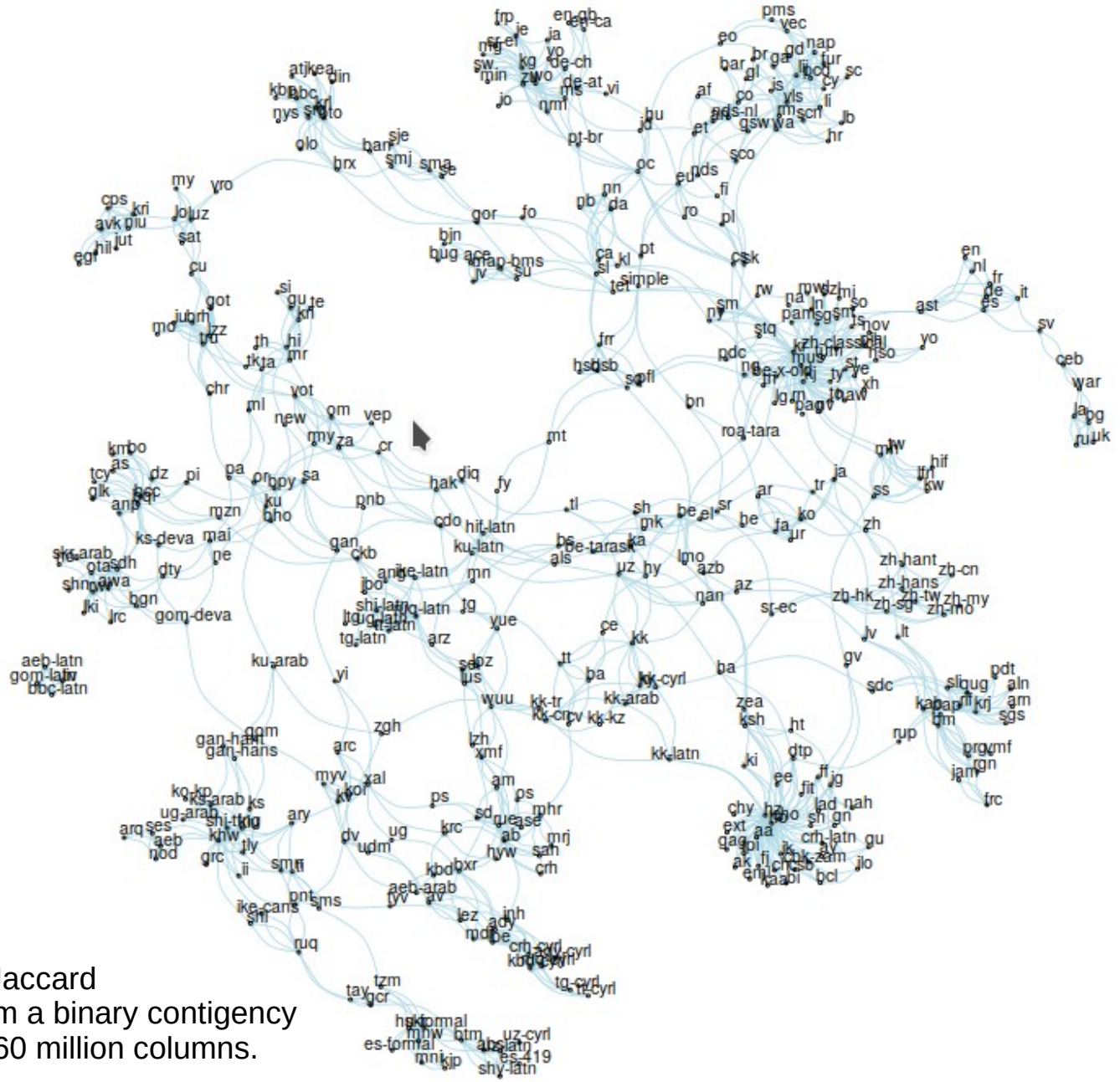
Method 4: processing the Wikidata JSON dump in R

https://github.com/datakolektiv/R-Ladies_Belgrade_20190911

```
# - read one line from the dump
  f <- readLines(con = con,
                 n = 1,
                 ok = FALSE,
                 warn = TRUE,
                 encoding = "unknown",
                 skipNul = FALSE)
  # - if the line is empty: break (EOF)
  if (length(f) == 0) {
    break
    # - else: parse JSON
  } else {
    # - parse w. rjson::fromJSON, remove "," at the end of the line;
    # - defensive:
    fjson <- tryCatch({
      rjson::fromJSON(gsub(",$", "", f),
                      method = "C",
                      unexpected.escape = "skip",
                      simplify = FALSE)
      },
      error = function(condition) {
        FALSE
      })
    # - check if the JSON was parsed correctly
    if (class(fjson) == "logical") {
      next
    }
    # - if fjson$labels$en$value is not null: process and write data
    if (!is.null(fjson$labels$en$value)) {
      writeLines(paste0('"', fjson$id, '"', ",", '"',fjson$labels$en$value, '"'), conOut)
    }
```

Semantic Web Technologies and Wikidata from R

How to access



This graph is derived from a Jaccard distance metric computed from a binary contigency matrix of approx. 440 rows x 60 million columns.

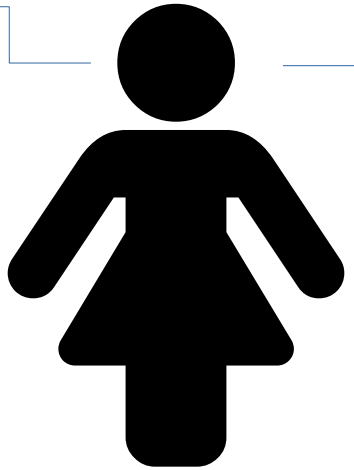
In base R + {Matrix}

Direct market value of open data in EU from 2016 to 2020: estimated EUR 325 billion
Predicted number of Open Data jobs in Europe by 2020: 100,000 (35% increase)

[Source: GovLab, http://thegovlab.org/open-data-index-2018-edition/]

… “they” have acess to expensive Big Data technologies, I can’t afford such an investment.

… “they” know Big Data, I will never make it.



I still need to take a tons of courses and I already barely have any time for myself, I will not make it.

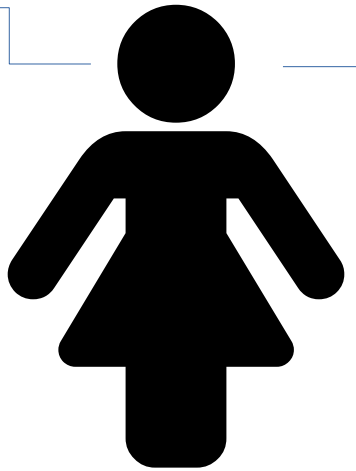
Direct market value of open data in EU from 2016 to 2020: estimated EUR 325 billion
Predicted number of Open Data jobs in Europe by 2020: 100,000 (35% increase)

[Source: GovLab, http://thegovlab.org/open-data-index-2018-edition/]

I wanted to show you that it can be done on your laptop.

… “they” have acess to expensive Big Data technologies, I can’t afford such an investment.

… “they” know Big Data, I will never make it.



I still need to take a tons of courses and I already barely have any time for myself, I will not make it.

Direct market value of open data in EU from 2016 to 2020: estimated EUR 325 billion
Predicted number of Open Data jobs in Europe by 2020: 100,000 (35% increase)

[Source: GovLab, http://thegovlab.org/open-data-index-2018-edition/]

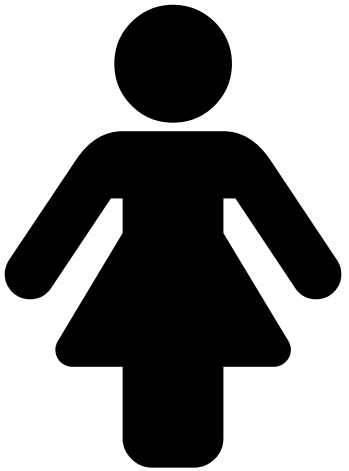
I wanted to show you that it can be done on your laptop.

I run one i7, 4 physical/8 logical cores + 32Gb RAM and .5Tb SSD remotely.

It costs me EUR 50 monthly.

Imagine what a good R developer can do there.

… Where will I find the money to invest in the infrastructure..?



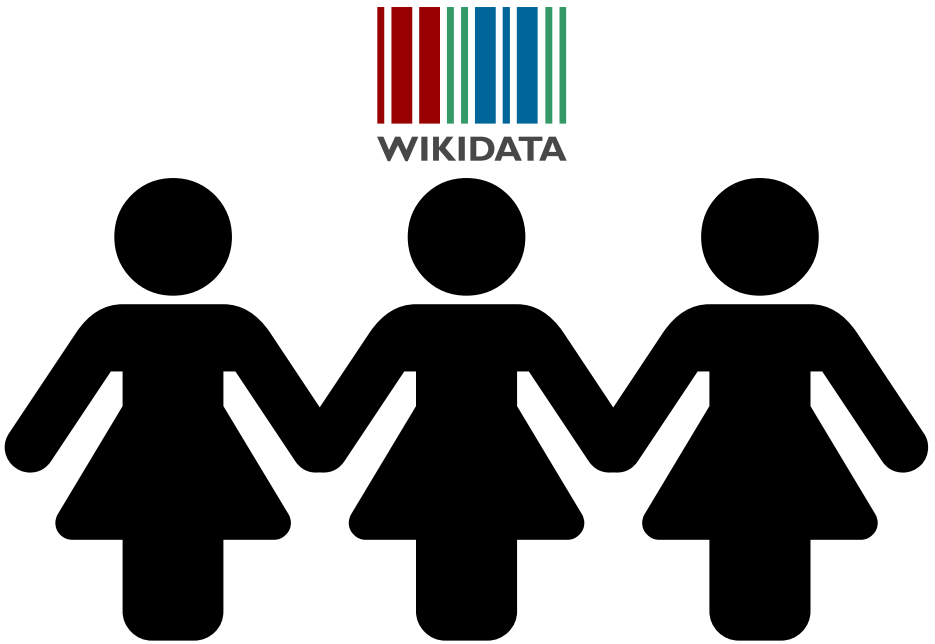
Direct market value of open data in EU from 2016 to 2020: estimated EUR 325 billion
Predicted number of Open Data jobs in Europe by 2020: 100,000 (35% increase)

[Source: GovLab, http://thegovlab.org/open-data-index-2018-edition/]

I wanted to show you that it can be done on your laptop.

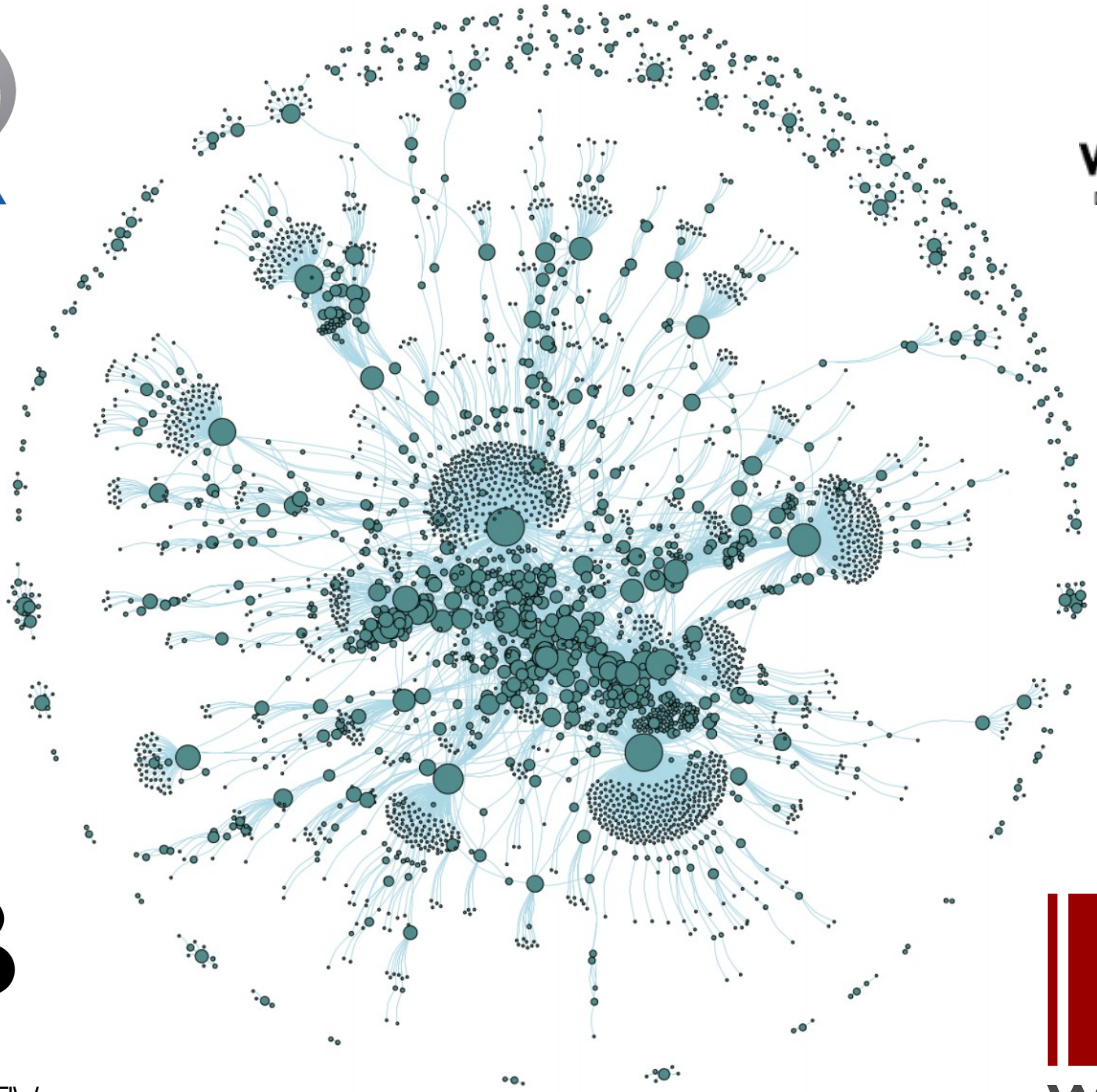
Start prototyping now.

For approx. EUR 150 monthly (and probably less) you will find a remote server w. 40 cores, 128Gb of RAM, and some 2Tb of SSD or more, and then there's no end to what you can do.



Direct market value of open data in EU from 2016 to 2020: estimated EUR 325 billion
Predicted number of Open Data jobs in Europe by 2020: 100,000 (35% increase)

[Source: GovLab, http://thegovlab.org/open-data-index-2018-edition/]



WIKIMEDIA
DEUTSCHLAND

DATAKOLEKTIV

WIKIDATA