

Open Data Projets @ SAPO Labs

Jorge Teixeira

SAPO LABS

- Research & Development Laboratories
- Potentiate the approach between industry and academy
- Promote and articulate R&D projects with national and international academic and research institutes
- Position itself as a R&D partner

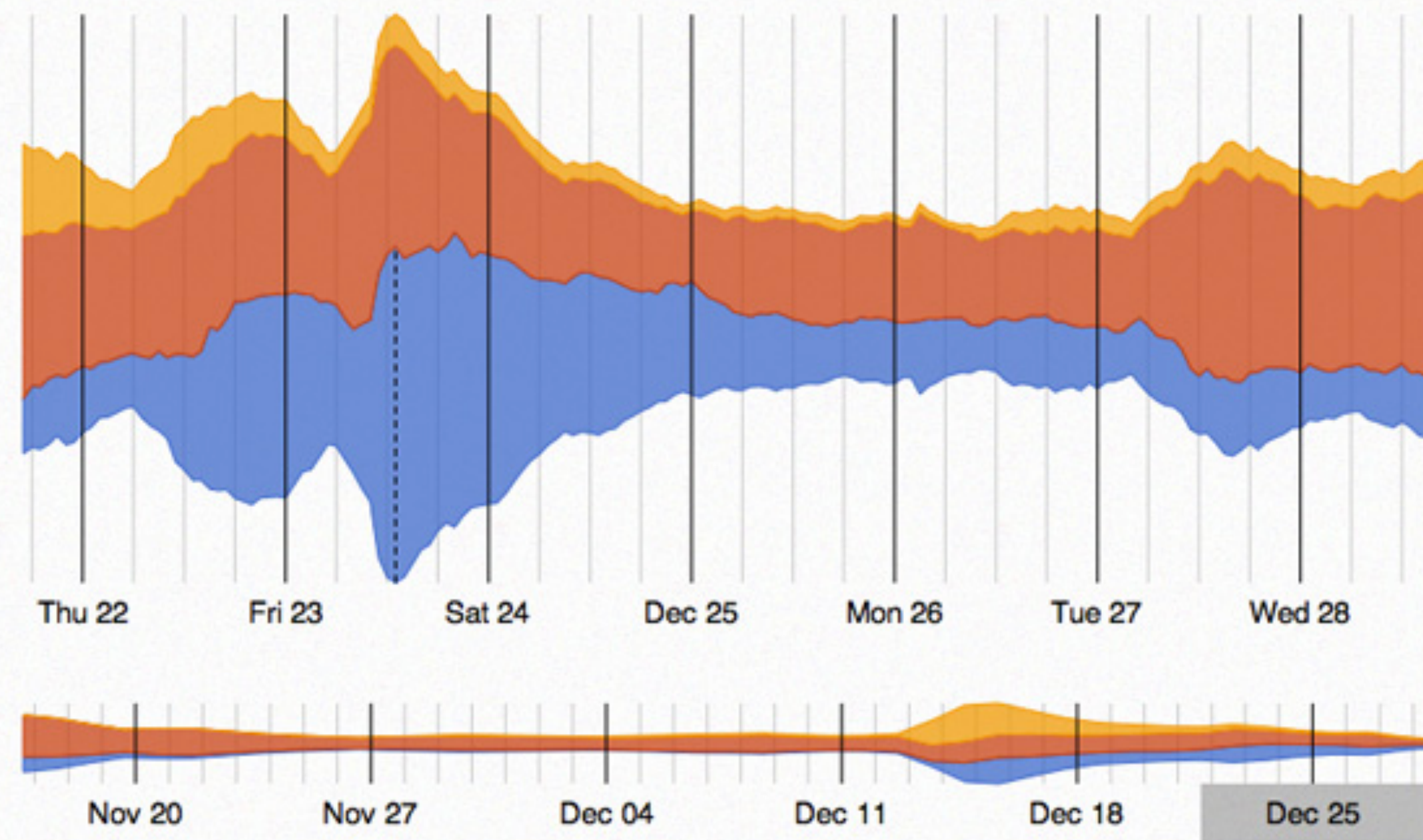
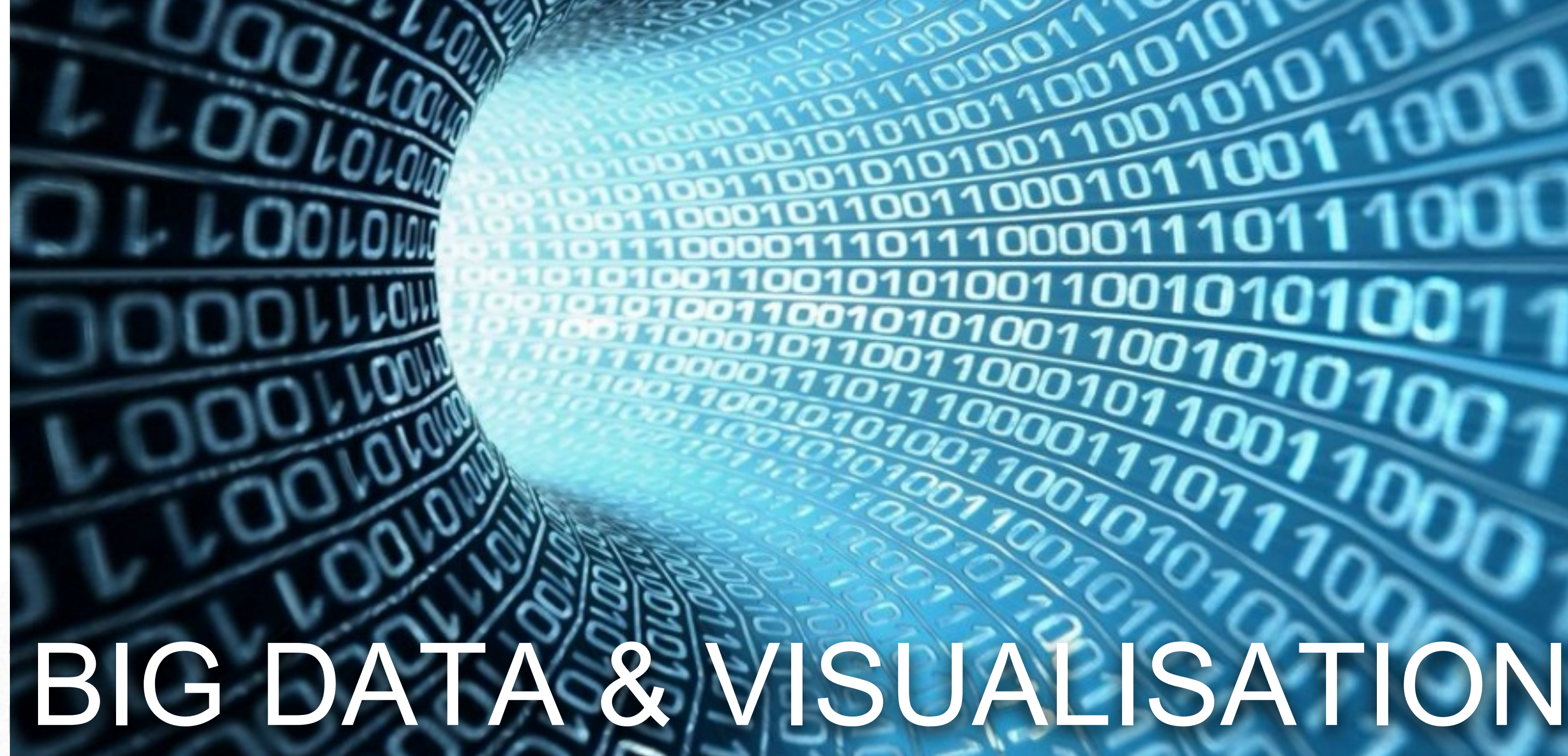


R&D AREAS



SAPO LABS

R&D AREAS



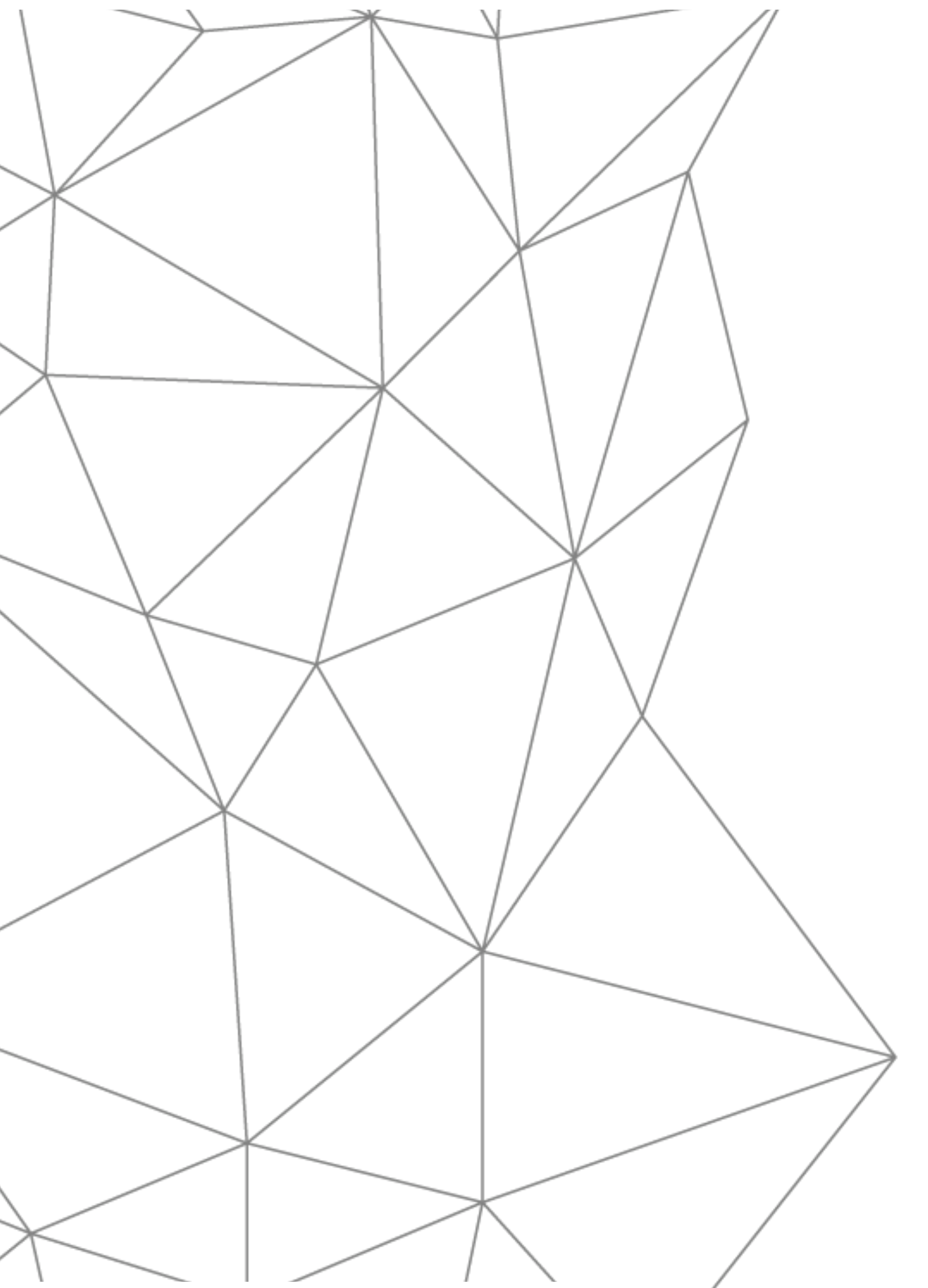
@GaryMWatson it will only work if the twitter community goes after more than one of these companies. don't see the movement spreading #SOPA

tbpHFX 12/23/11 13:00:00
Reported @GoDaddy's Twitter spam, everyone should do the same #SOPA #BoycottGoDaddy

thathurtabit 12/23/11 13:00:00
RT @FredericJacobs: I'm disappointed by Apple, Adobe and Autodesk and others who are supporting #SOPA
<http://t.co/ID5Rfs98>

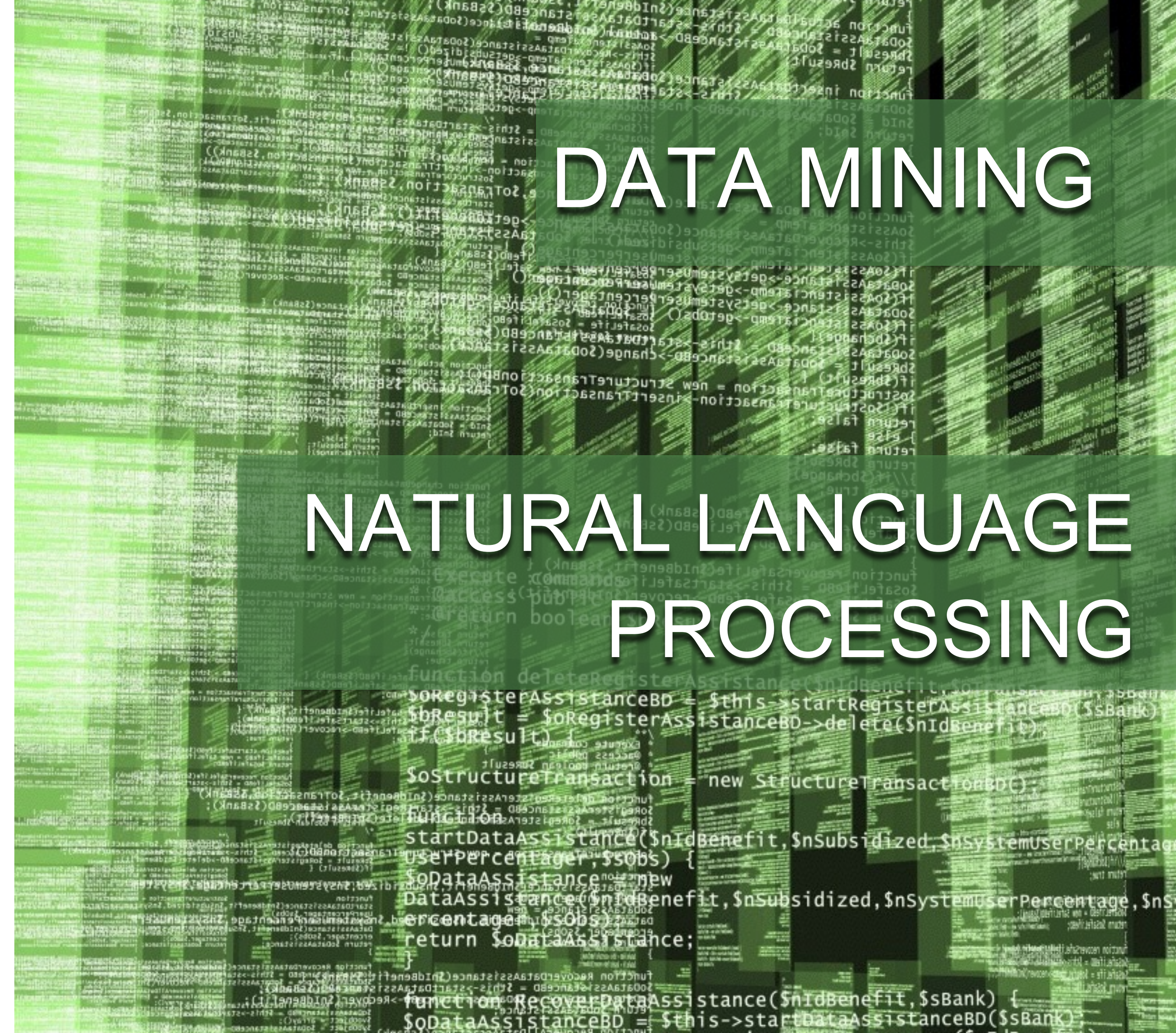
rknLA 12/23/11 13:00:00
RT @hinsterbaker: Excellent work

R&D AREAS



DATA MINING

NATURAL LANGUAGE PROCESSING



R&D AREAS

**MACHINE
LEARNING**

**COMPUTATIONAL
LINGUISTICS**

R&D AREAS

VIDEO SEARCH AND ANALYSIS

R&D AREAS

SOCIAL NETWORK ANALYSIS

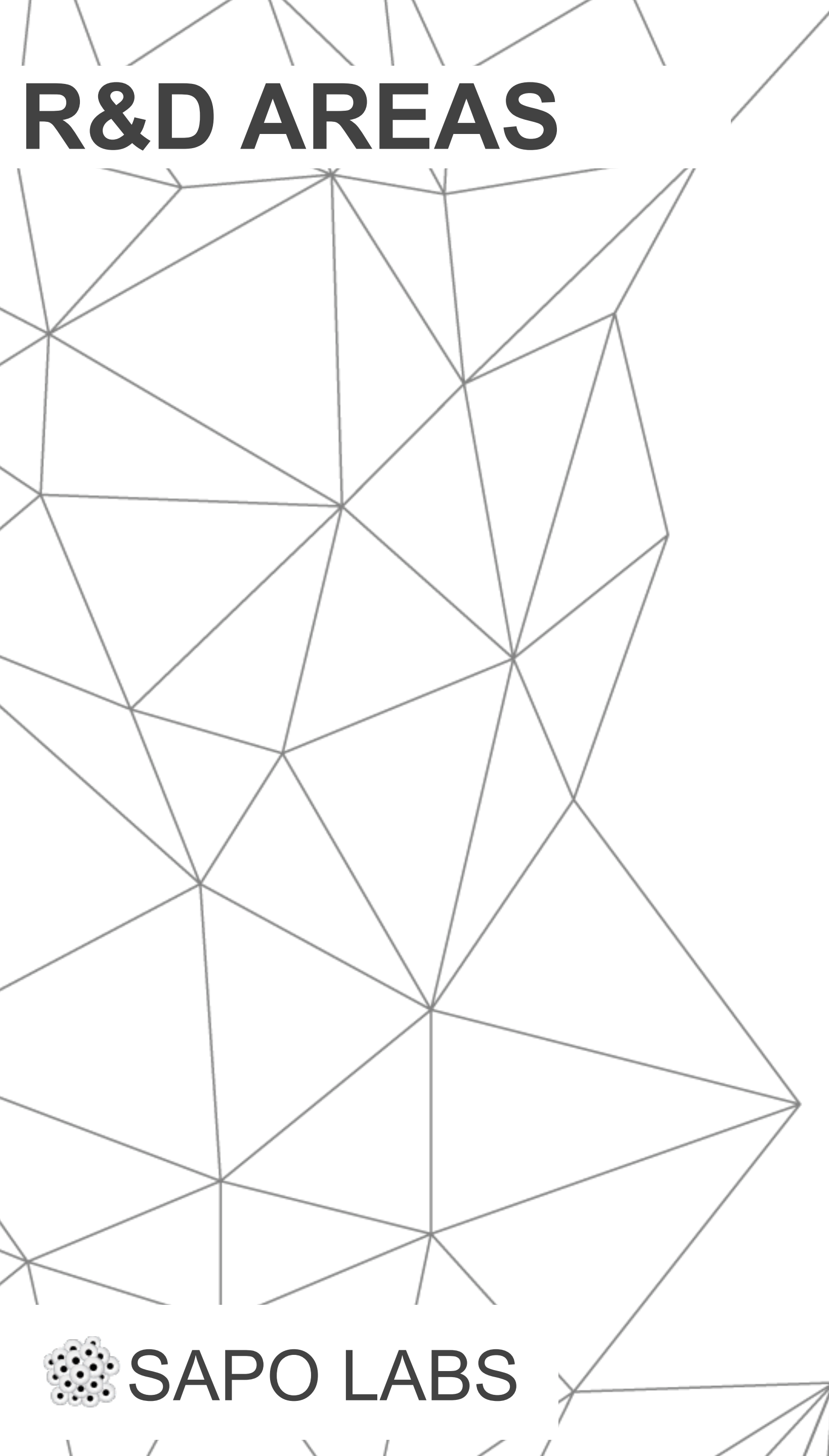


R&D AREAS

 **SAPO LABS**

**CYBER
SECURITY**

R&D AREAS



R&D AREAS



R&D AREAS



R&D PARTNERS



SAPO LABS

R&D PARTNERS



Open Data



SAPO LABS

PROJECTS OPEN DATA

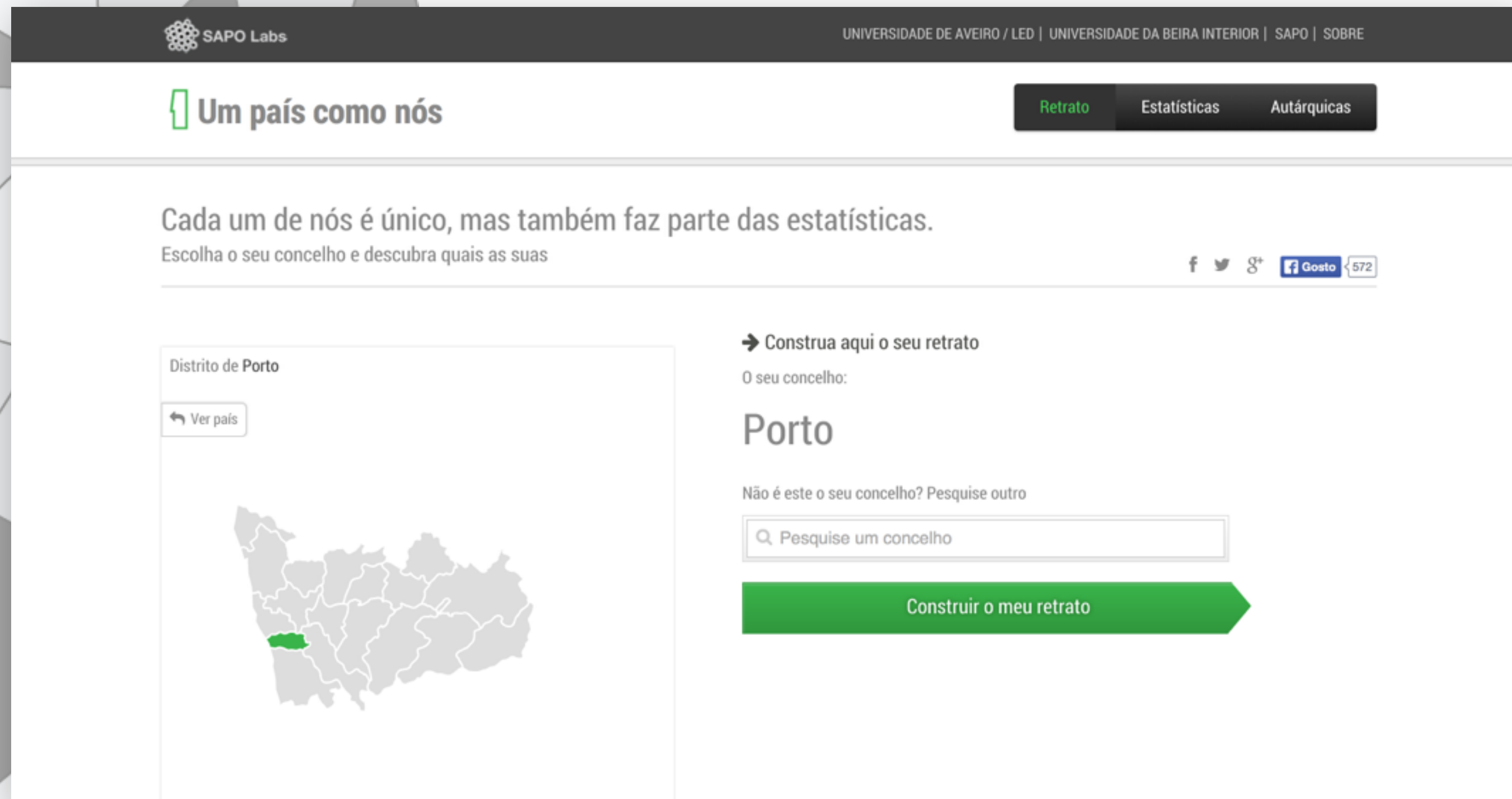
- “Open data refers to the idea that certain data should be freely available to everyone to use and re-use” *(from citek)*
- “(...) predicts that the overall economics gains from opening up public data could amount to €40 billion a year in the European Union” *(from citek)*
- “The main goal is to (...) contribute to the continuous increase of the integration and interoperability of the internet” *(from citek)*

PROJECTS



SAPO LABS

PROJECTS UM PAÍS COMO NÓS



- Data Sources: INE, IEFP, DGAL and PORDATA
- Interactive web tool to connect each of us with the national statistics
- 3 entry points: “portrait”, “statistics” and “elections”

PROJECTS UM PAÍS COMO NÓS



COIMBRA

UM PAÍS COMO NÓS

Reconstruir retrato

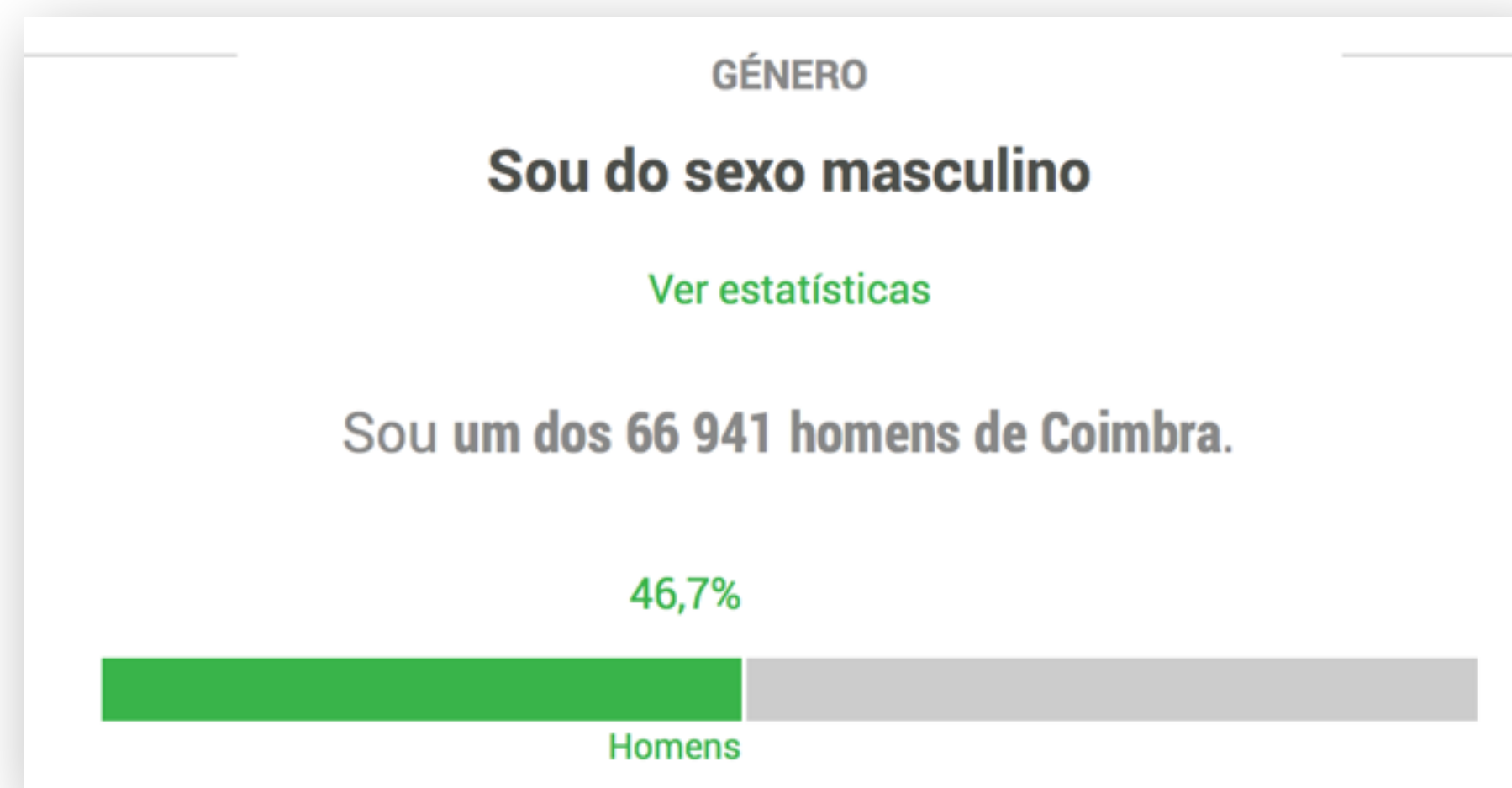
Vivo em Coimbra

Sou do sexo masculino.

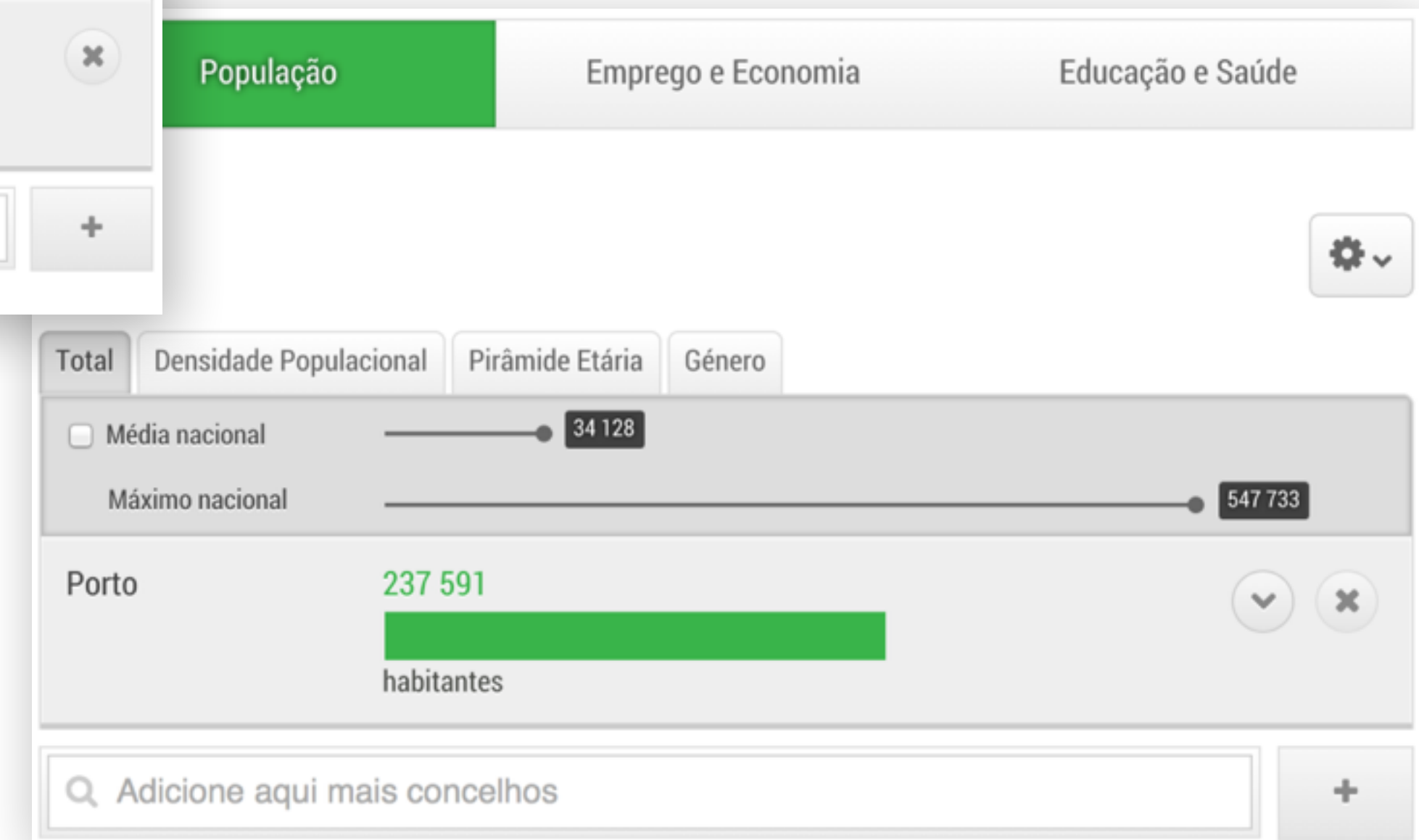
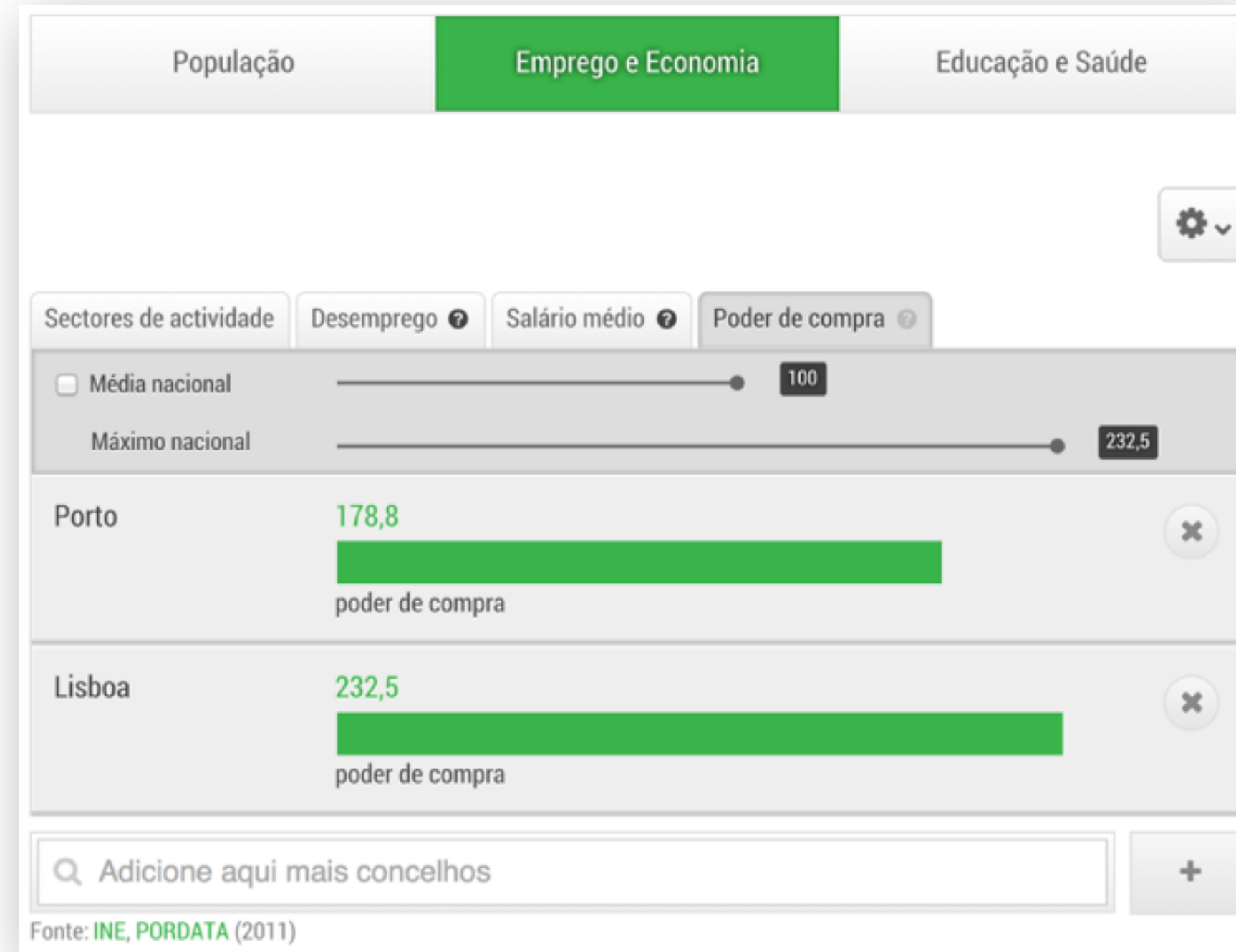
Tenho entre 30 e 34 anos

Estudei até ao ensino Superior

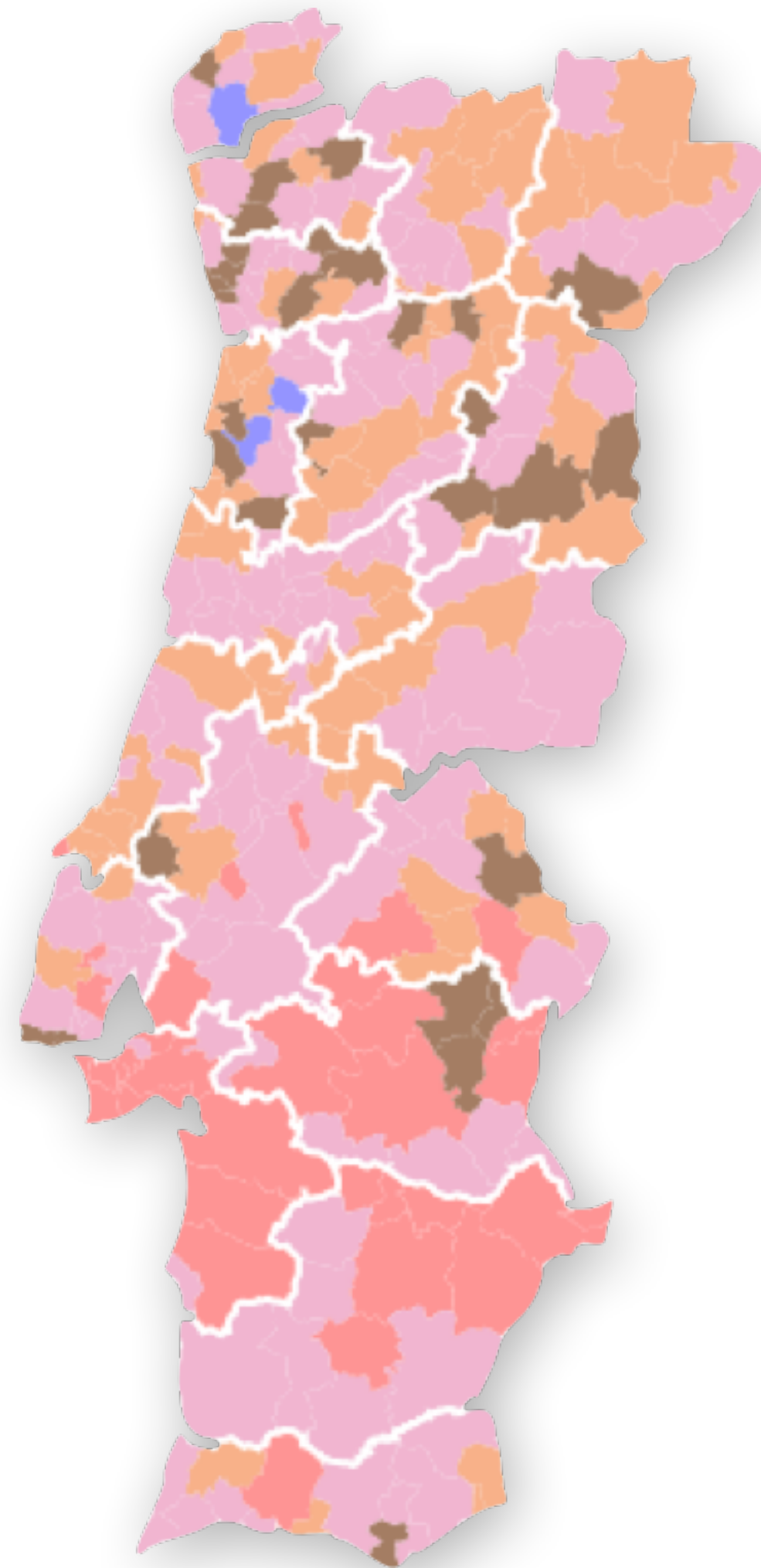
Trabalho no sector Terciário.



PROJECTS UM PAÍS COMO NÓS



PROJECTS UM PAÍS COMO NÓS



Resultados Globais

Votação global dos partidos



Abstenção: **47,4%**

Branco e nulos: **6,82%**

Número de câmaras por partido

■ PS

149 (mais 17 que em 2009)

[Ver câmaras no mapa](#)

■ PSD

86 (menos 31 que em 2009)

[Ver câmaras no mapa](#)

■ CDU

34 (mais 6 que em 2009)

[Ver câmaras no mapa](#)

■ BE

0 (menos 1 que em 2009)

[Ver câmaras no mapa](#)

■ CDS-PP

5 (mais 4 que em 2009)

[Ver câmaras no mapa](#)

■ Independentes

13 (mais 6 que em 2009)

[Ver câmaras no mapa](#)

PROJECTS GRANDE ÁREA

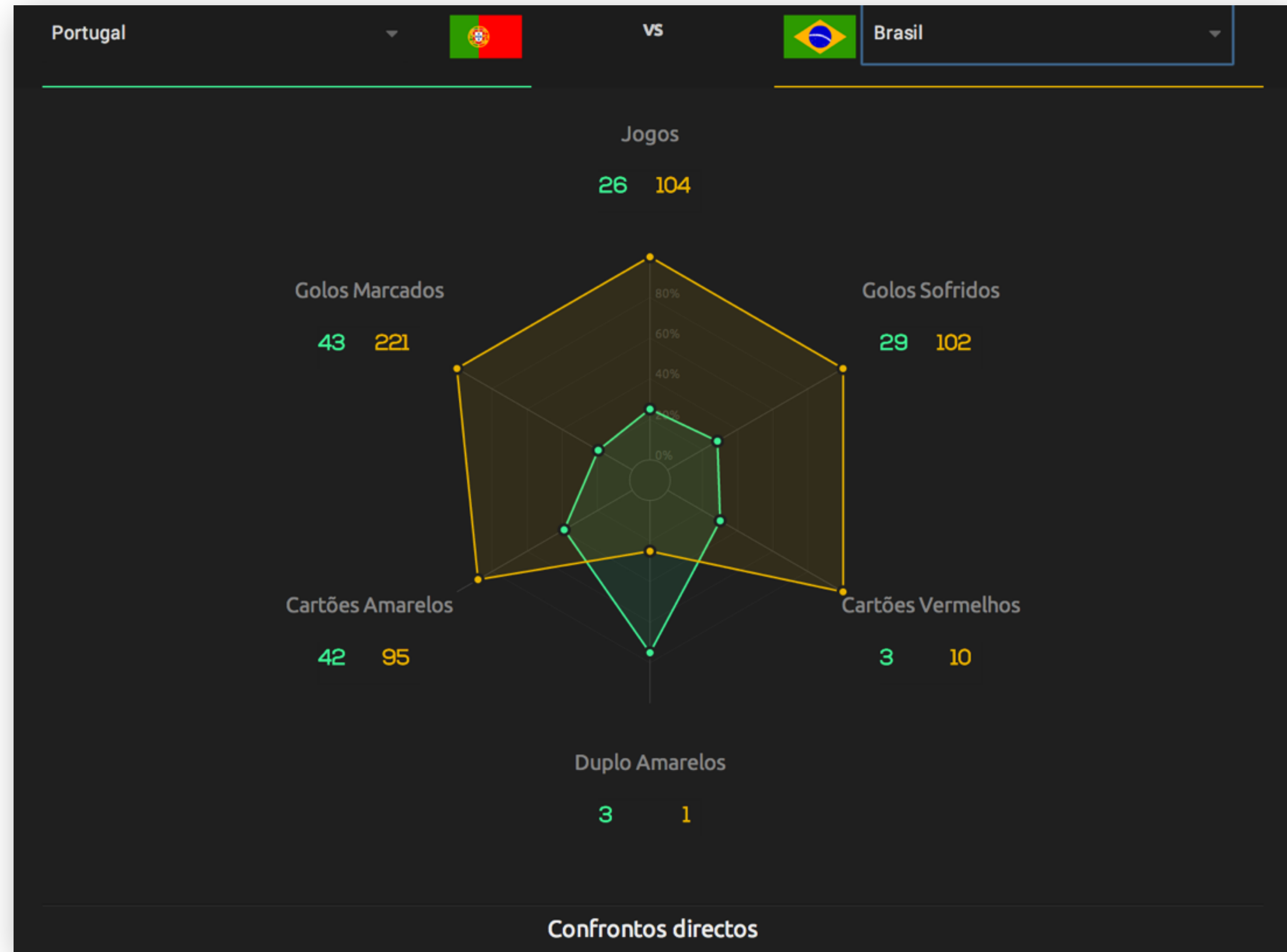


- Data Sources: soccer events
- A different perspective over sports statistics
- Information about teams and players on three perspectives: discipline, efficiency and experience

PROJECTS GRANDE ÁREA



PROJECTS GRANDE ÁREA



PROJECTS GRANDE ÁREA

CRISTIANO RONALDO



Portugal	29 anos	84 kg	1.86 m
Clube: Real Madrid	Posição: Avançado	Nº 7	
281	266	51	
Jogos jogados	Golos marcados	Cartões recebidos	

Histórico das transferências

Espessura da linha = valor da transferência
Tamanho do círculo = N.º de minutos jogados



[< Voltar ao top de jogadores](#)

Grande Área

Os craques e as equipas que batem recordes

SAPO Labs desporto EDIGMA SportInvest

FUTEBOL CLUBE DO PORTO

Porto, Portugal
Estádio: Estádio do Dragão
Nº de troféus por competição

27	40	5
Liga nacional	Taças nacionais	Comp. internacionais

Plantel atual

**Ricardo Quaresma**
Ricardo Andrade Quaresma Bernardo

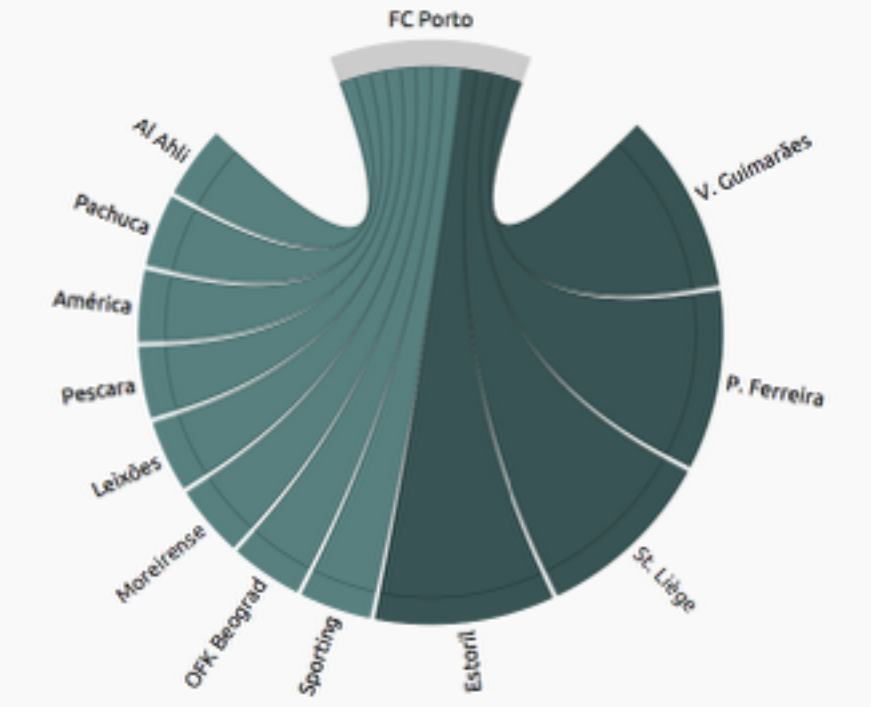
**Helton**
Helton da Silva Arruda

**Fucile**

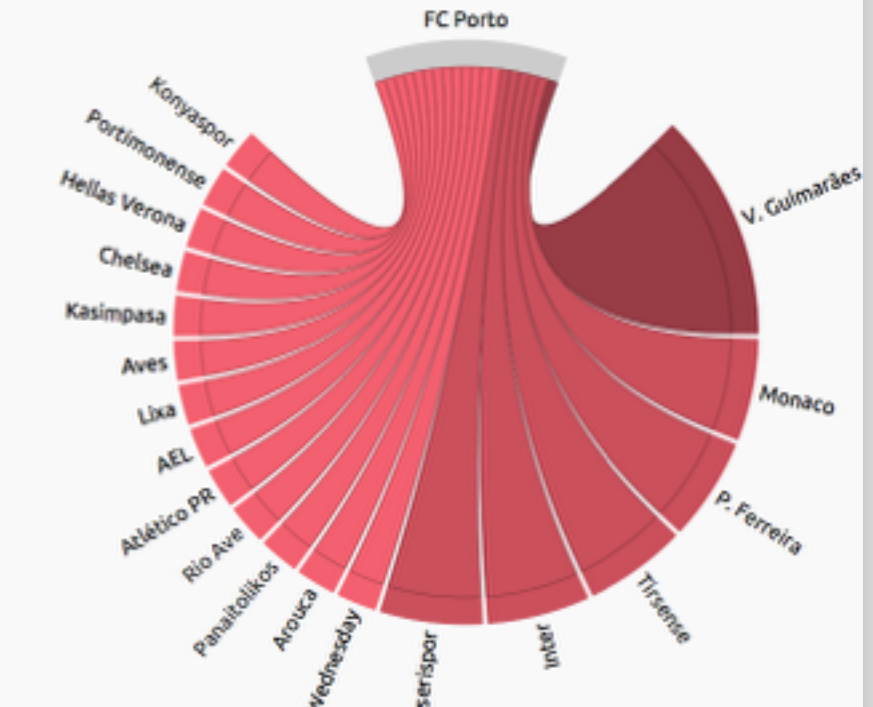
[< Voltar ao top de equipas](#)

Transferências entre o Futebol Clube do Porto e outros clubes

Total de jogadores recebidos: **16**



Total de jogadores transferidos: **27**



Uma temporada [2010-11](#) [2011-12](#) [2012-13](#) **2013-14**

1 vermelhos

11 amarelos

/ 50 jogos jogados

75 Pontos

/ 50 jogos jogados

Uma temporada [2010-11](#) [2011-12](#) [2012-13](#) **2013-14**





- Data Sources: Twitter and Facebook
- Interactive visualisation of the presence of portuguese music artists on the social networks
- Information about artists, albums and music styles

PROJECTS O MUNDO EM PESSOA

The screenshot displays the SAPO Labs website interface. At the top, the SAPO Labs logo is on the left, and navigation links for SAPO.PT, SAPO Labs, FCUL, and Sobre are on the right. The main heading is "O MUNDO EM PESSOA". Below this, a sub-header reads "Citações de Fernando Pessoa nas redes sociais". There are four filter buttons: "Hoje", "Última semana" (which is highlighted), "Último mês", and "Top citações". A large number "322 citações" is prominently displayed. To the right of this number are buttons for "Facebook", "twitter", and "Gosto", along with a count of "1,2 mil". Below the filters, a horizontal list shows the number of citations for various authors: Fernando Pessoa (42), Alberto Caeiro (21), Álvaro de Campos (165), Bernardo Soares (86), and Ricardo Reis (8). The main content area is a grid of quote cards. Each card includes a quote, the author's name, the number of occurrences, and a link to the original text. Some cards also feature a small profile picture and a Twitter handle. The quotes are in various colors and fonts, making them visually distinct.

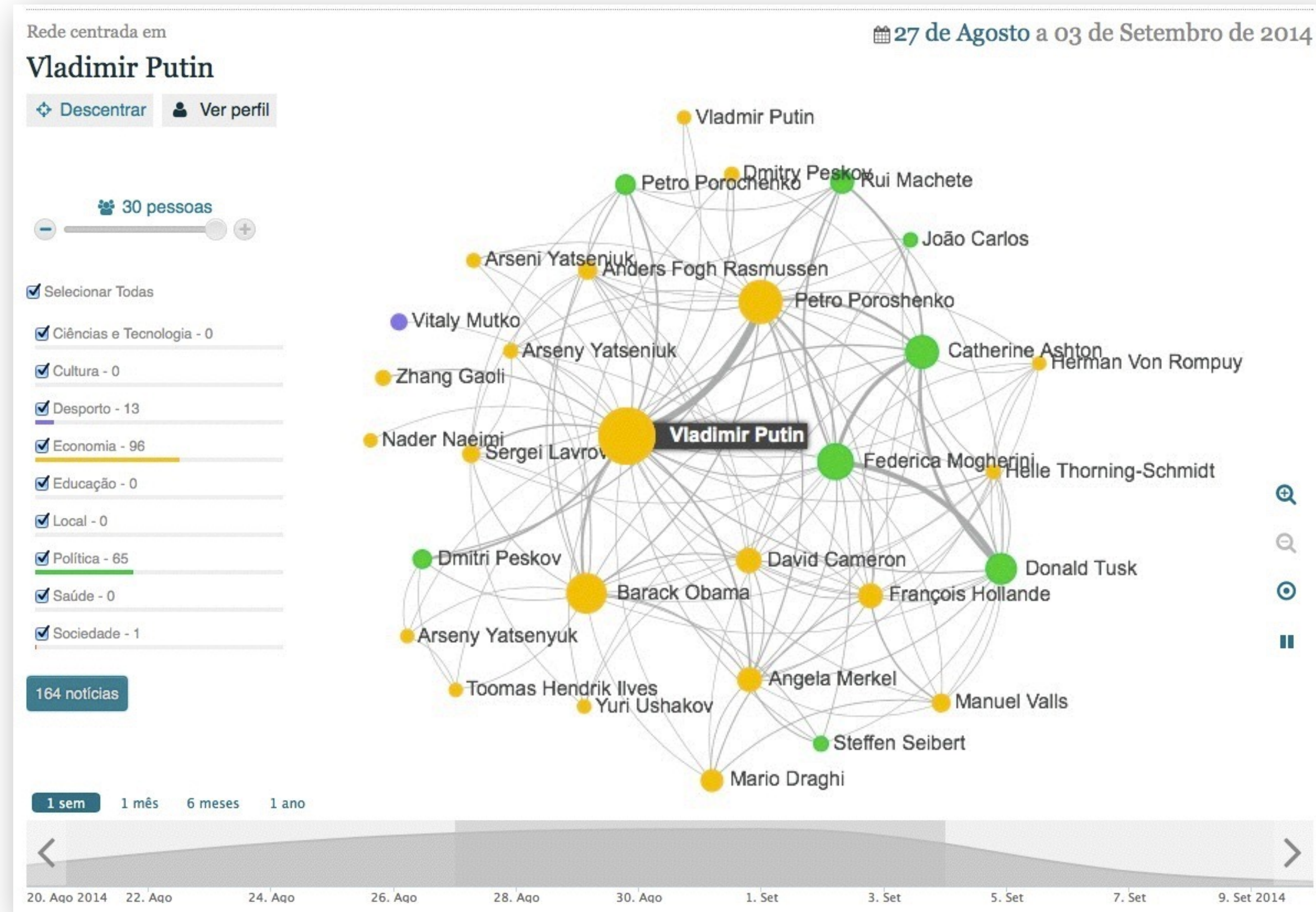
- Data Sources: Twitter and Facebook
- Identify and visualize quotes from Fernando Pessoa poems on the social networks

SAPO MÁQUINA DO TEMPO



- Data Sources: knowledge extracted from portuguese news
- Interactive web tool to explore 25 years of news articles based on networks of co-occurrences of personalities
- All knowledge is extracted automatically, without human intervention
- News can be explored from a particular personality of time frame

PROJECTS MÁQUINA DO TEMPO



TOP ÚLTIMOS 7 DIAS - OS MAIS REFERIDOS NAS NOTÍCIAS



384 notícias ↓ - 5%

 **Fernando Santos**

 treinador



348 notícias ↓ - 48%

 **Cristiano Ronaldo**

 futebolista português



287 notícias ↑ + 19%

 **Pedro Passos Coelho**

 primeiro-ministro e líder social-democrata



167 notícias ↑ + 13%

 **António Costa**

 presidente da Câmara de Lisboa



157 notícias ↓ - 29%

 **Ricardo Carvalho**

 defesa



144 notícias ↓ - 5%

 **Nuno Crato**

 ministro da Educação

159 Notícias

🕒 15 DE OUTUBRO DE 2014, 13:32 | SIC

Diretores pedem que comece "já" o trabalho para colocar professores no próximo ano ➦

🕒 15 DE OUTUBRO DE 2014, 16:05 | RENASCENÇA

Já é conhecido o juiz que vai mediar o conflito entre ministério e professores ➦

🕒 15 DE OUTUBRO DE 2014, 16:05 | RENASCENÇA

Já é conhecido o juiz que vai mediar o conflito entre ministério e professores ➦

🕒 15 DE OUTUBRO DE 2014, 22:05 | EXPRESSO

Ambiente e Economia vencem no aumento da despesa, Agricultura e Educação lideram cortes ➦

🕒 16 DE OUTUBRO DE 2014, 00:05 | DIÁRIO ECONOMICO

Educação leva corte de 704 milhões de euros ➦

8 Citações

🕒 14 DE OUTUBRO DE 2014, 16:16 | DIÁRIO IOL

"O caminho não é recuar na autonomia, é dar às escolas a escolha dos professores" ➦

Referido em mais 3 fontes

🕒 14 DE OUTUBRO DE 2014, 16:16 | DIÁRIO IOL

"Estão marcadas reuniões para esta semana com diretores das escolas TEIP e com autonomia" ➦

Referido em mais 1 fontes

🕒 14 DE OUTUBRO DE 2014, 18:14 | JORNAL DE NEGÓCIOS

"A compensação de professores vai ser feita nos termos da lei, com uma comissão de acompanhamento presidida por um magistrado" ➦

🕒 14 DE OUTUBRO DE 2014, 18:14 | JORNAL DE NEGÓCIOS

"Devo respeitar o trabalho que esta comissão vai fazer e devo de seguida proceder de forma mais célere para que as recomendações dessa comissão sejam postas em prática" ➦

RESOURCES



SAPO LABS

RESOURCES DATASETS

- 16 datasets
- More than 20 million different documents
- Data from public domains, social networks, media, public knowledge, etc.
- Not enough, ask us for more!



RESOURCES DATASETS

Dataset	Description
UmPaisComoNos	Public statistics from Portugal (education, health, population, etc.)
Blogs/PT	4 million blog posts (of national scope) since 2013
CurseWords	A set of curse (bad) words frequently mentioned on social networks.
ObfuscatedMessages	Comprehends a set with 500 sports news comments manually annotated with curse (bad) words.
News/PT/Lusa	4 million news articles from the portuguese news agency LUSA.
News/Comments	20 thousand comments to online news articles.
News/ PersonalitiesOccurrences	Approximately 350 thousand occurrences of personalities on news articles published in 2013.
News/Tags	More than 10 million meta-tags from news published between 2009 e 2014.
News/PT/Online	A set of 6 million news articles published by the most relevant Portuguese media producers since 2009.

RESOURCES DATASETS

Dataset	Description
News/BR/Online	A set of 100 thousand news articles published by the most relevant Brazilian media producers since 2013.
Personalities/Names	A set of 9 thousand names of personalities mentions on news articles.
Personalities/Quotations	Approximately 200 thousand direct and indirect quotations extracted from news articles.
Verbetes/Personalities	130 thousand tuples <name, job/role> extracted from news articles.
Social/Twitter/FernandoPessoa	More than 200 thousand posts from social networks mentioning the poet Fernando Pessoa.
SentimentMessages	Comprehends a set of 1000 twitter messages manually annotated with sentiment towards a target.
Social/Twitter/WorldCup2014	2 million twitter messages regarding WorldCup 2014 event.

RESOURCES APIs ACCESS

- All datasets are available, use them!
- Two methods available: “getDocument” and “getDocuments”
 - <https://services.sapo.pt/Blogs/PT/getDocuments?ESBToken=01.EZs5j5fIZTMf326pOLmC9g>
- Each call returns by default the first 100 documents
- To get the next ones, use the parameter “start”
 - <https://services.sapo.pt/Blogs/PT/getDocuments?start=100&ESBToken=01.EZs5j5fIZTMf326pOLmC9g>

<https://t1hj17.s.cld.pt>



Thank you!

jorge-teixeira@telecom.pt