25 July 2018

Identifying and visualising tonality/emotion and geographical information in the text of ECA Special Reports using IBM Watson and Tableau
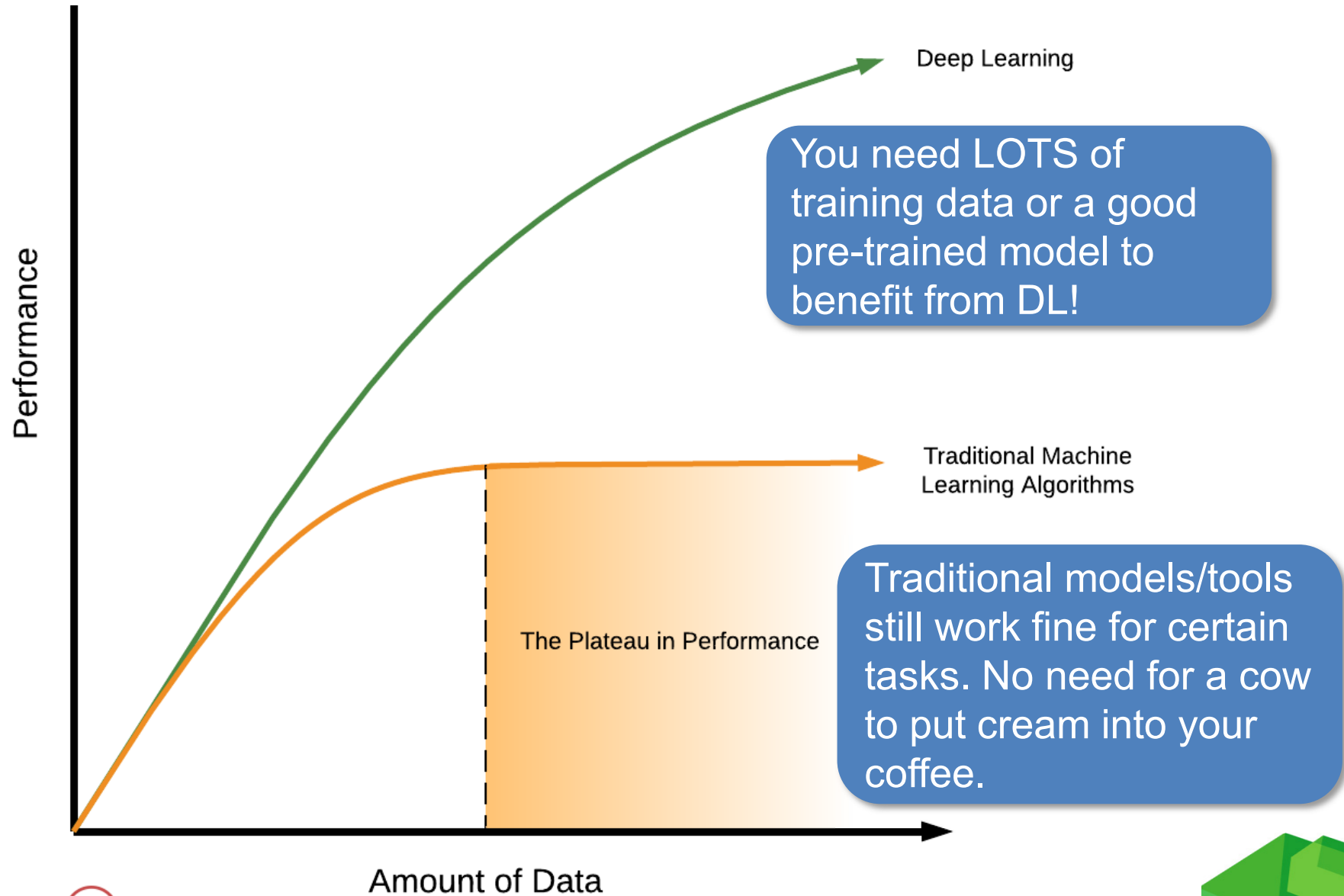
Zsolt VARGA, ECA.SEC-GEN.SG3.HU
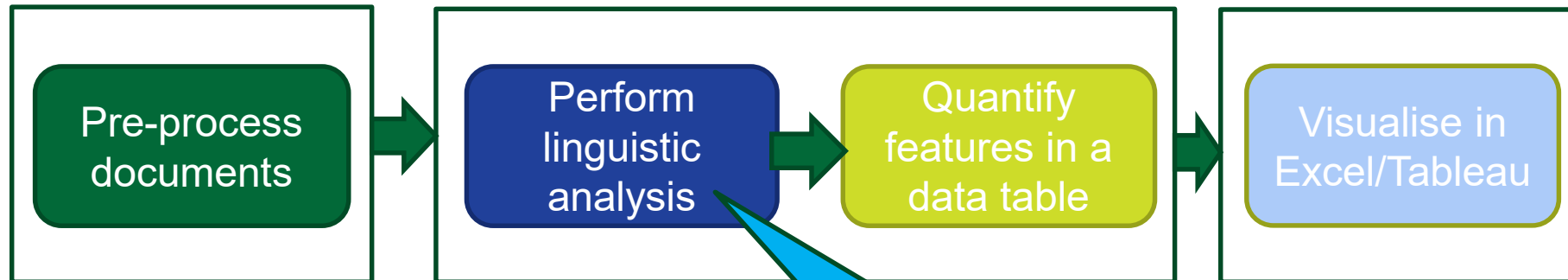
zsolt.varga@eca.europa.eu

Neural vs. Traditional

Text mining pipeline

Pre-process documents → Perform linguistic analysis → Quantify features in a data table → Visualise in Excel/Tableau

- Named Entity Recognition
- Vocabulary analysis
- Sentiment analysis

EUROPEAN COURT OF AUDITORS

Geo-coverage of ECA special reports

Geo-coverage based on report content

Geographical coverage of SRs by report, 2010-2016
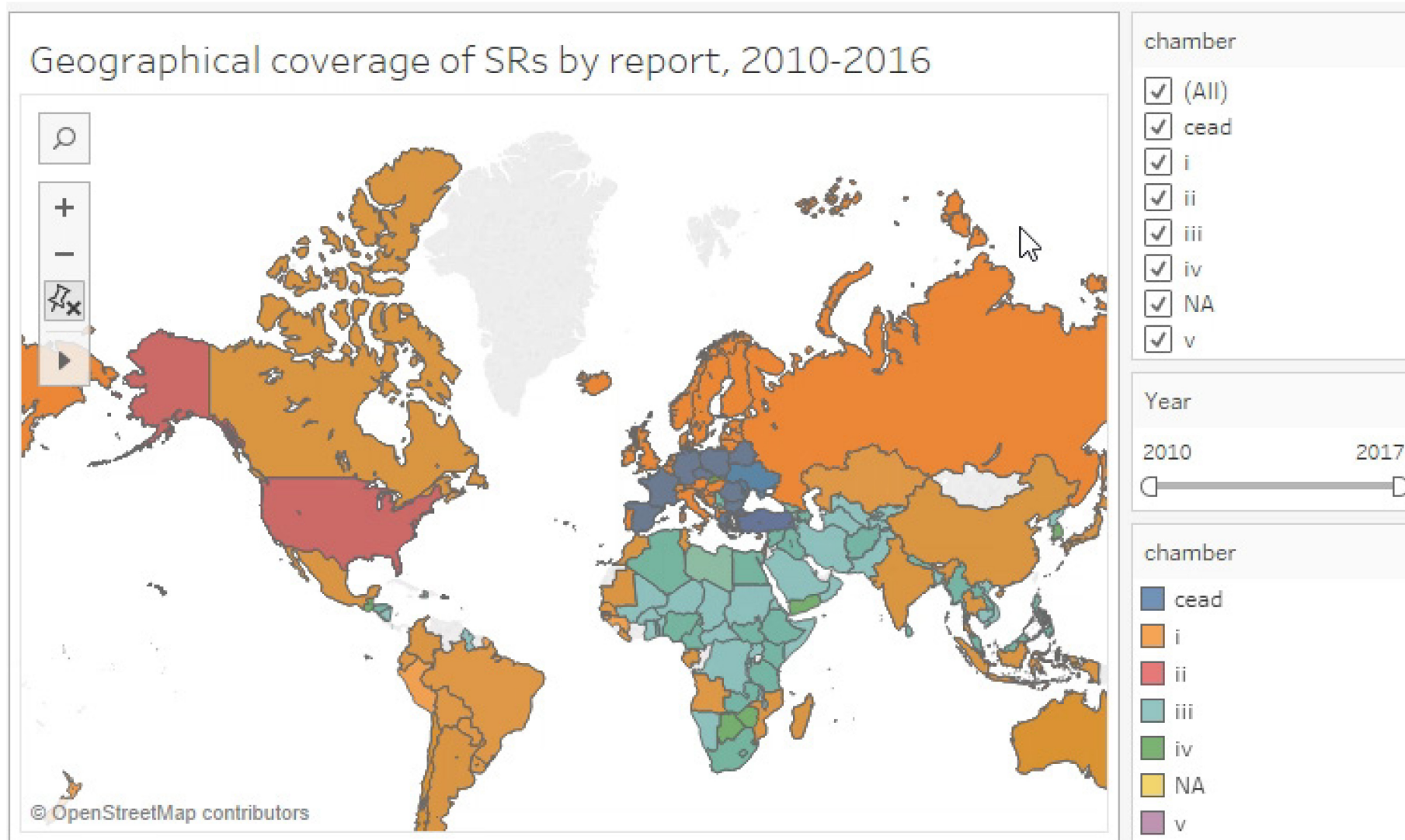
Geo-coverage based on report content (2010-2016)

Chamber III seems to cover the whole world!

Chamber II is more focused on Europe and our partners, the USA, Turkey and Russia

CURIA RATIONUM

EUROPEAN
COURT
OF AUDITORS
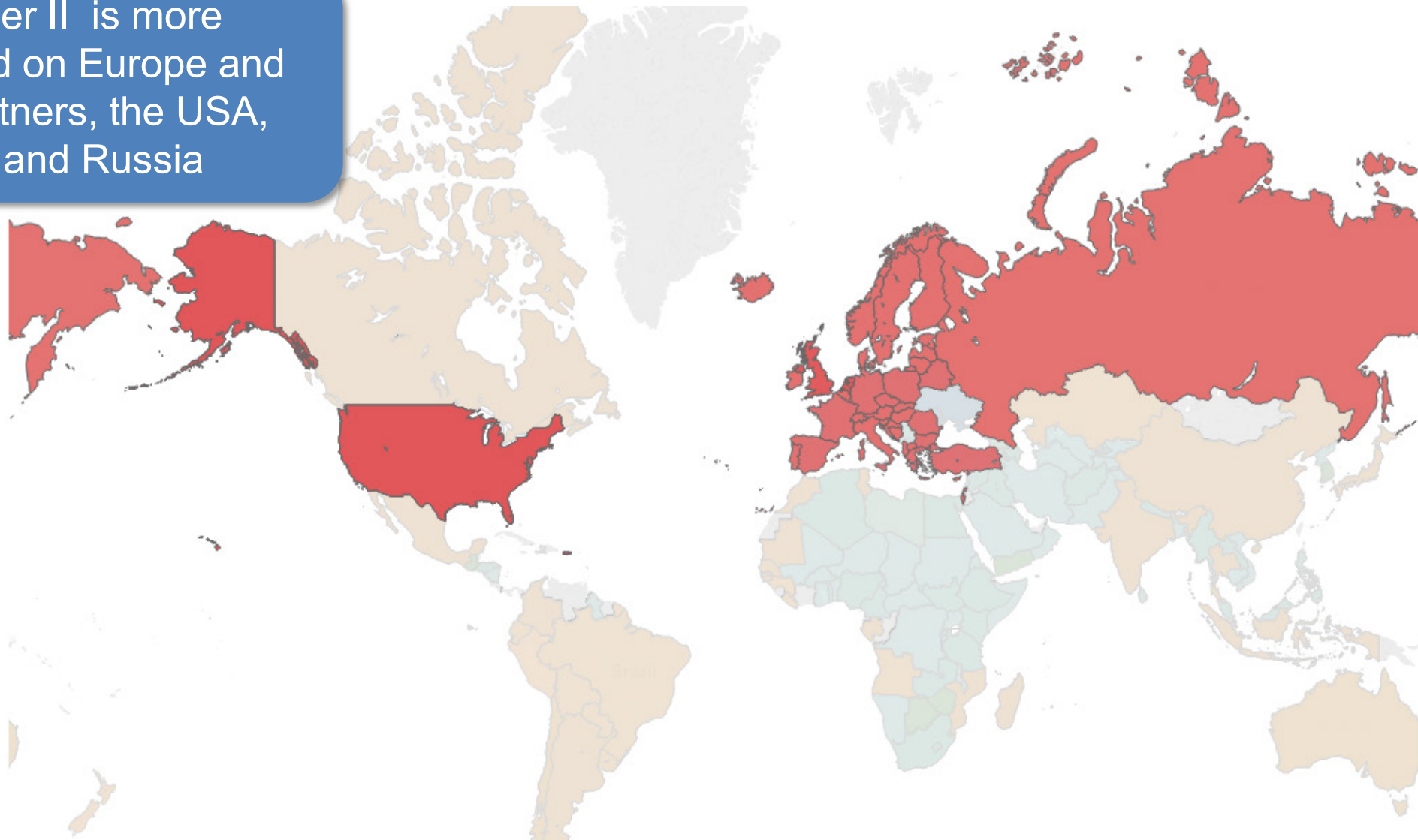
Geo-coverage based on report content (2013-2014, min. 25 refs)

Countries covered in 2013-2014

Countries mentioned at least 25 times
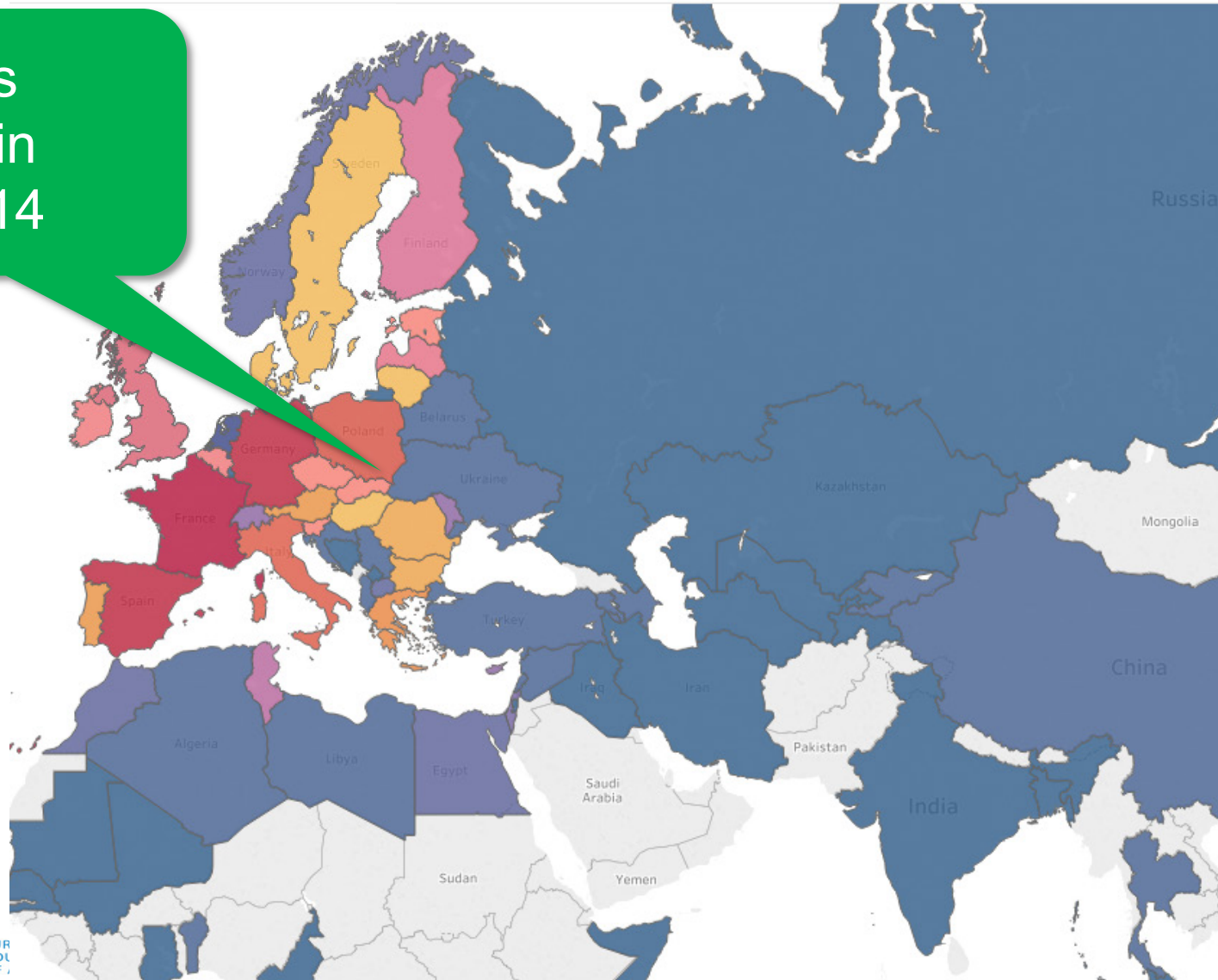
Year
2013   2014

Minimum no. of references
25   98

SRs by country

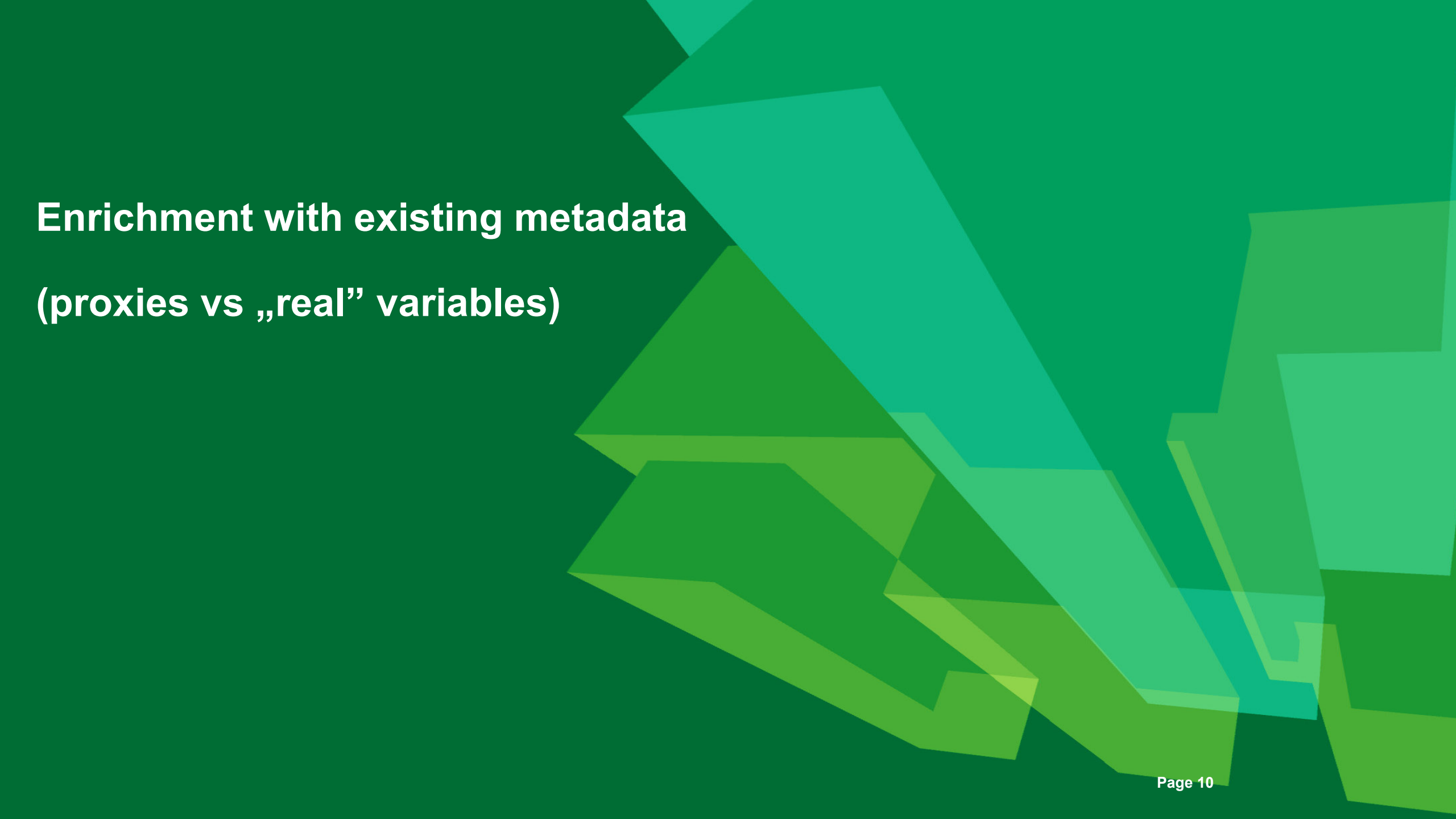| Location	SR name	
austria	RS-Renewable-Energy	6/2014
egypt	RS-Egypt	4/2013
haiti	RS-Support_Haiti	13/2014
israel	RS-Palestine	14/2013
italy	RS-EBF	15/2014
kyrgyzstan	RS-Central_Asia_FINAL	13/2013
poland	RS-Axis3	6/2013
	RS-Renewable-Energy	6/2014
serbia	RS-Serbia	19/2014
sweden	RS-Axis3	6/2013
tajikistan	RS-Central_Asia_FINAL	13/2013

Geo-coverage based on report content (2013-2014)

Countries covered by
SR Agri-Environment
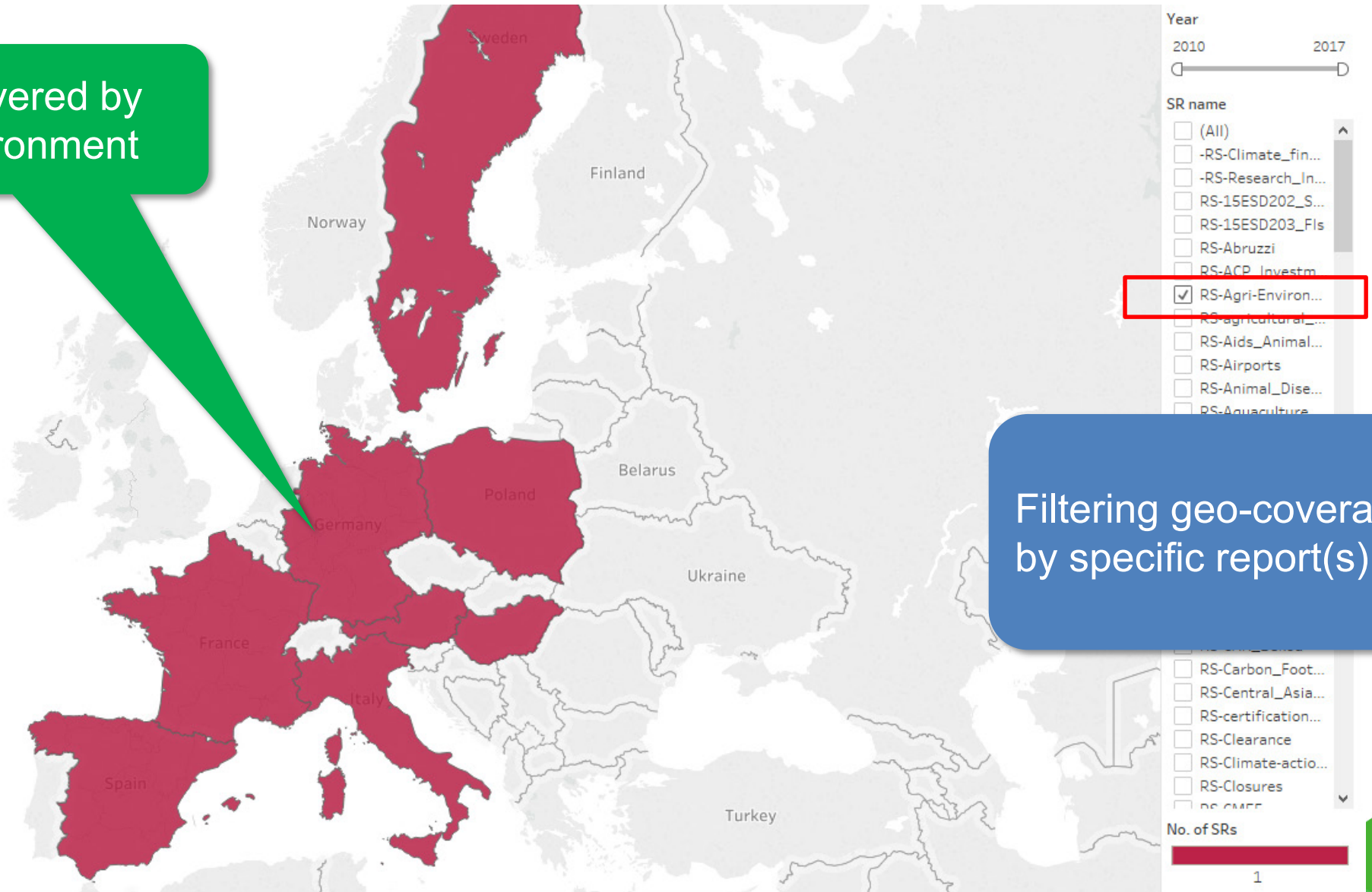
Filtering geo-coverage
by specific report(s)

Year
2010   2017

SR name
- (All)
- -RS-Climate_fin...
- -RS-Research_In...
- RS-15ESD202_S...
- RS-15ESD203_Fis
- RS-Abruzzi
- RS-ACP_Investm
- ☑ RS-Agri-Environ...
- RS-agricultural_...
- RS-Aids_Animal...
- RS-Airports
- RS-Animal_Dise...
- RS-Aquaculture

- RS-Carbon_Foot...
- RS-Central_Asia...
- RS-certification...
- RS-Clearance
- RS-Climate-actio...
- RS-Closures
- RS-CMEF

No. of SRs

1

Enrichment with existing metadata

(proxies vs „real" variables)

Geo-coverage + persons associated with reports

Auditors who worked on Ireland-related topics in the 2015-2017 period
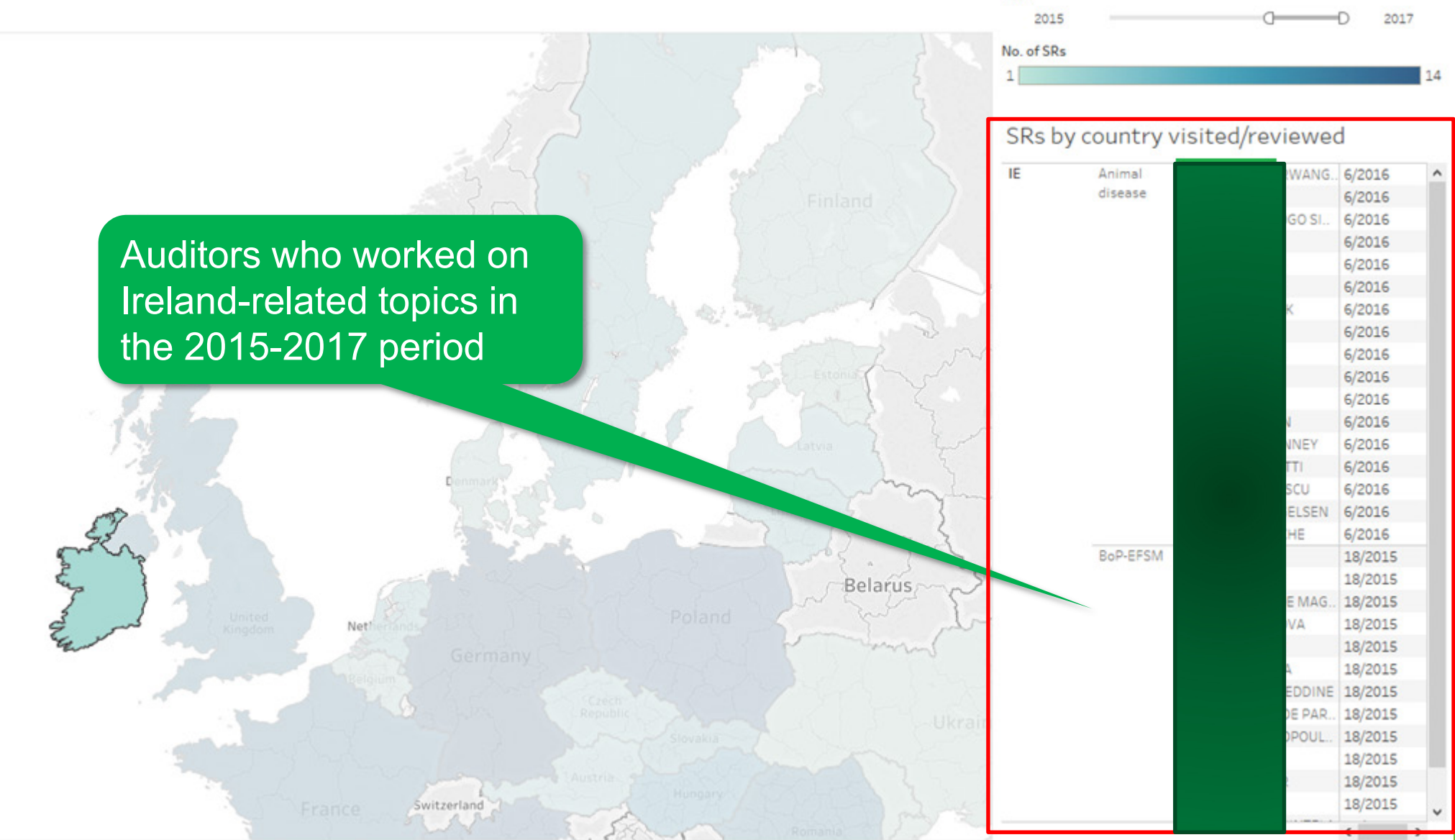
Tonality and emotion in ECA special reports

Sentiment analysis based on SR summaries

- At the current technological level sentiment analysis is unreliable on long texts.
- The Executive Summary provides a general overview of the reports and sets the tonality
- Subjectivity analysis was performed by a word-list based weighting method (SpaCy)
- Sentiment and tonality analysis was done by IBM Watson, using a pre-trained ensemble model (SVM, multi-label classifier)
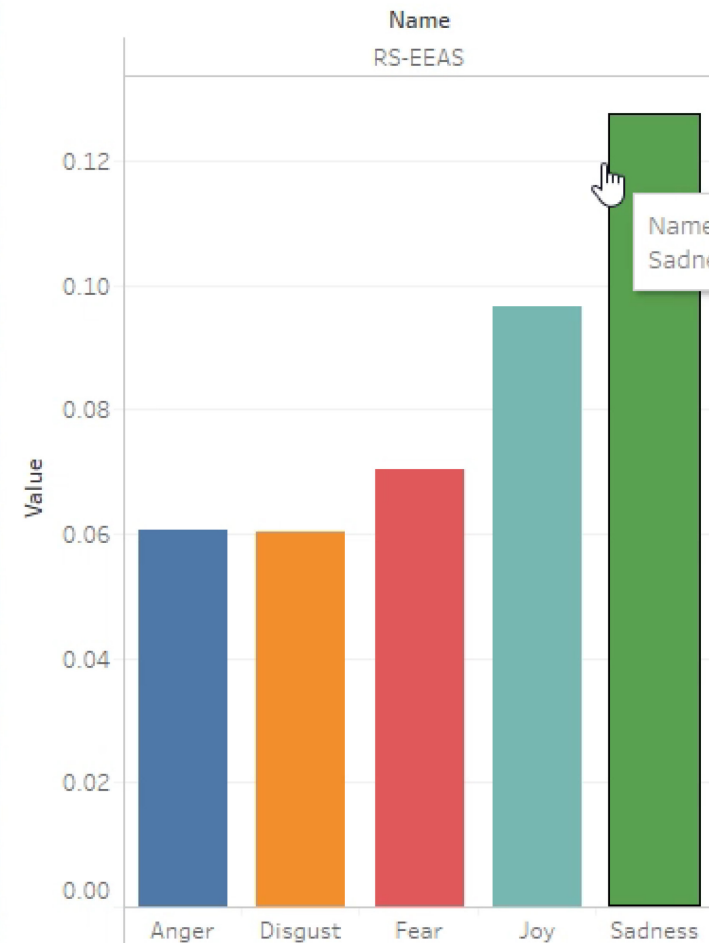
Sentiment analysis

SR Danube II – Water Quality

Watson identified „Sadness" because we are missing things (there are deficiencies and problems)

Watson identified „Disgust" as a relevant emotion, because of wastewater-related words
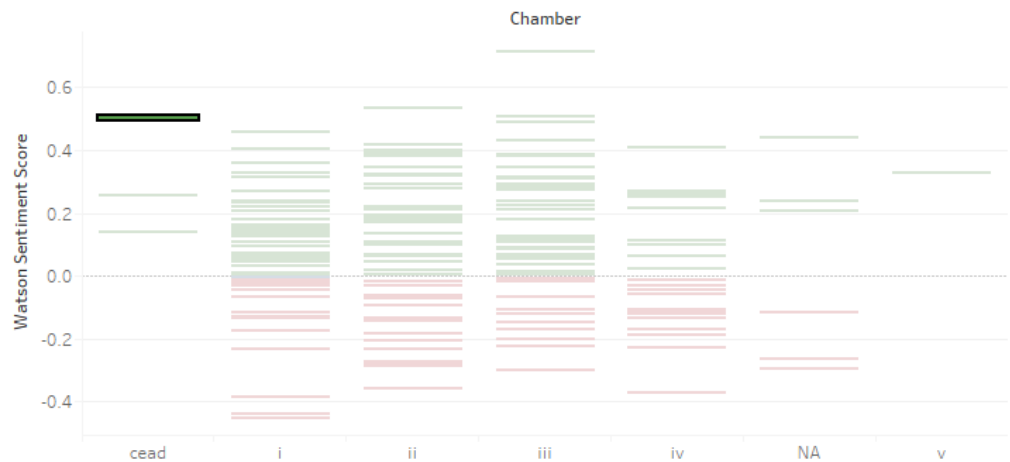
Follow-up report 2012

Connecting all this together

Geo-coverage with colours showing tonality

Geo-coverage with Watson Sentiment Score

chamber (regex SR summ..
- ☑ (All)
- ☑ cead
- ☑ i
- ☑ ii
- ☑ iii
- ☑ iv
- ☑ NA
- ☑ v

Year
2010   2017

Avg. Watson Sentiment S..

-0.2965   0.4637

© OpenStreetMap contributors
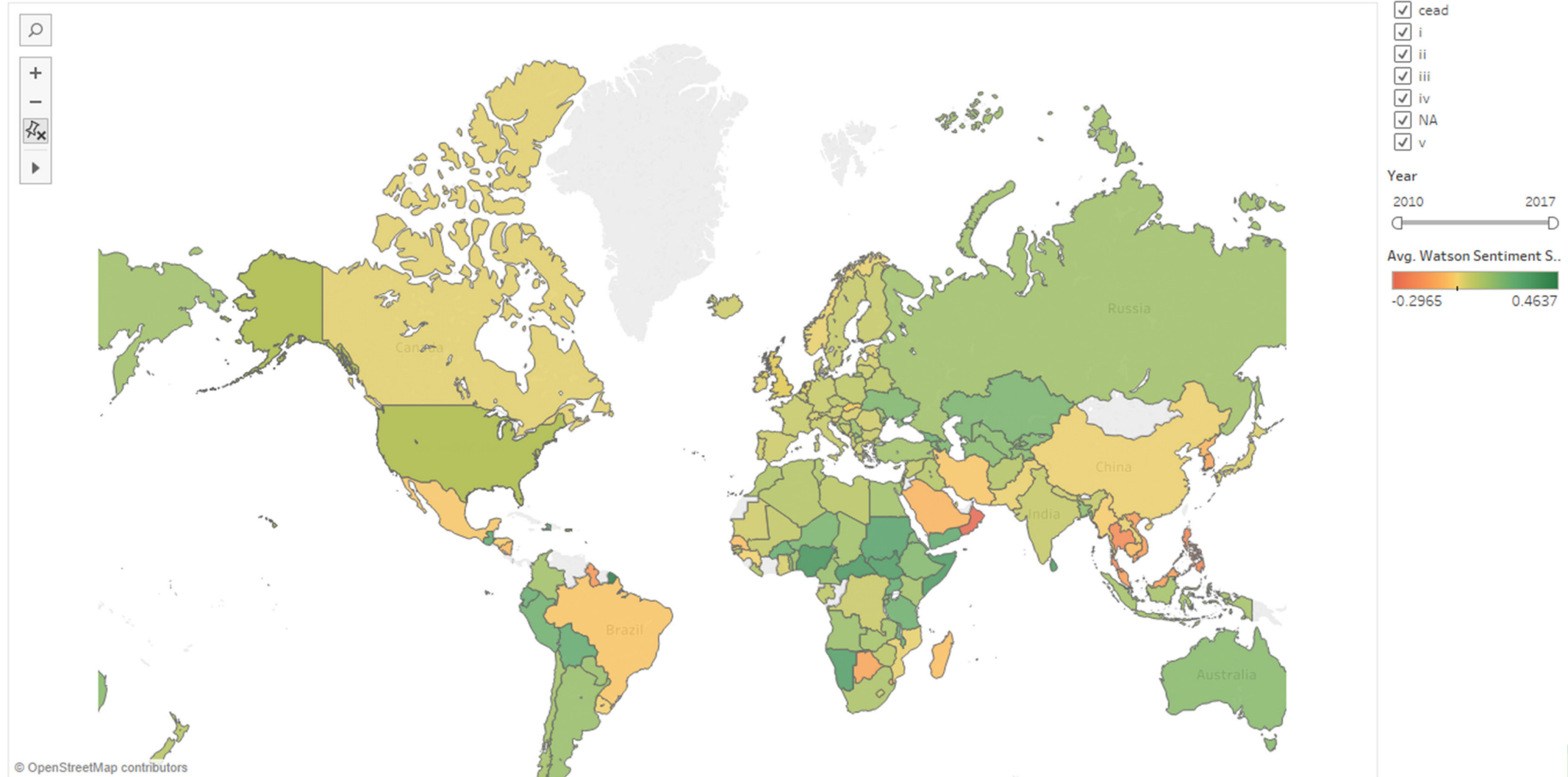
Closing remarks

Typical dilemmas:

- Off-the-shelf vs Custom developed
- Free vs Commercial
- Public/open vs Private/internal data
- Artificial vs Natural intelligence ☺

Start small, think BIG!

Find out if there's a „data lab" or somebody already experimenting with text mining within your organisation and team up with them. If not, be a champion yourself!

What you need:

- **Python** (Anaconda distribution) and the **SpaCy** library
- **IBM Watson** Natural Language Understanding (free tier cloud access)
- **Tableau** (trial version)
- *Documents you want to analyse* (convert them to plain text format beforehand and make sure they are not classified)

QUESTIONS?

Email: zsolt.varga@eca.europa.eu

LinkedIn: www.linkedin.com/in/zsvarga/

GitHub: github.com/zseebrz

Tableau Public: public.tableau.com/profile/zsolt.varga#!/

(the scripts and visualisations will be published
on GitHub and Tableau Public soon)