



# what is ddc?

- ddc (1.0) is an experimental dissertation-centric database
  - currently a 5 (6) country pilot dataset.
- the ddc approach combines multiple RISIS resources in order:
  - to enrich PhD Production data
  - to estimate broad labor market (LM) outcomes post dissertation
- ddc consists of 2 complementary components:
  - Component 1 enriches PhD Production data (DD)
  - Component 2 estimates a novel indicator of post-dissertation LM outcome (DC)

# why ddc?



- mounting evidence (in the literature) and concern (in public policy circles) about
  - striking increases in PhD production in recent years
  - the uncertain labor-market outcomes of trained PhDs
  - the effects on the research system in general and the research career in particular
- we need a reliable & generalizable approach
  - To understand changing PhD production patterns across countries and fields of science
  - To study the rate at which PhD students pursue careers inside (outside) academia over time
  - To appreciate factors that may shape labor market outcomes.

# what challenges?

Designing such an ideal approach faces particular challenges:

- **law: how to deal with gdpr issues**
  - Working with individual data & publishing derived information
- **data: how to organize & recruit RISIS tools/services**
  - Automating ingestion/cleaning/integrating raw data
  - Linking to ETER and CWTS publication
- **statistics: how to design a predictive setting to estimate LM outcomes**
  - Observed dissertations in term of total population
  - Observed link with publication with actual publication
  - Estimation of LM outcome in relation to 'real' outcome

➔ Host of problems that have haunted other attempts (UOE CDH)

# why RISIS?



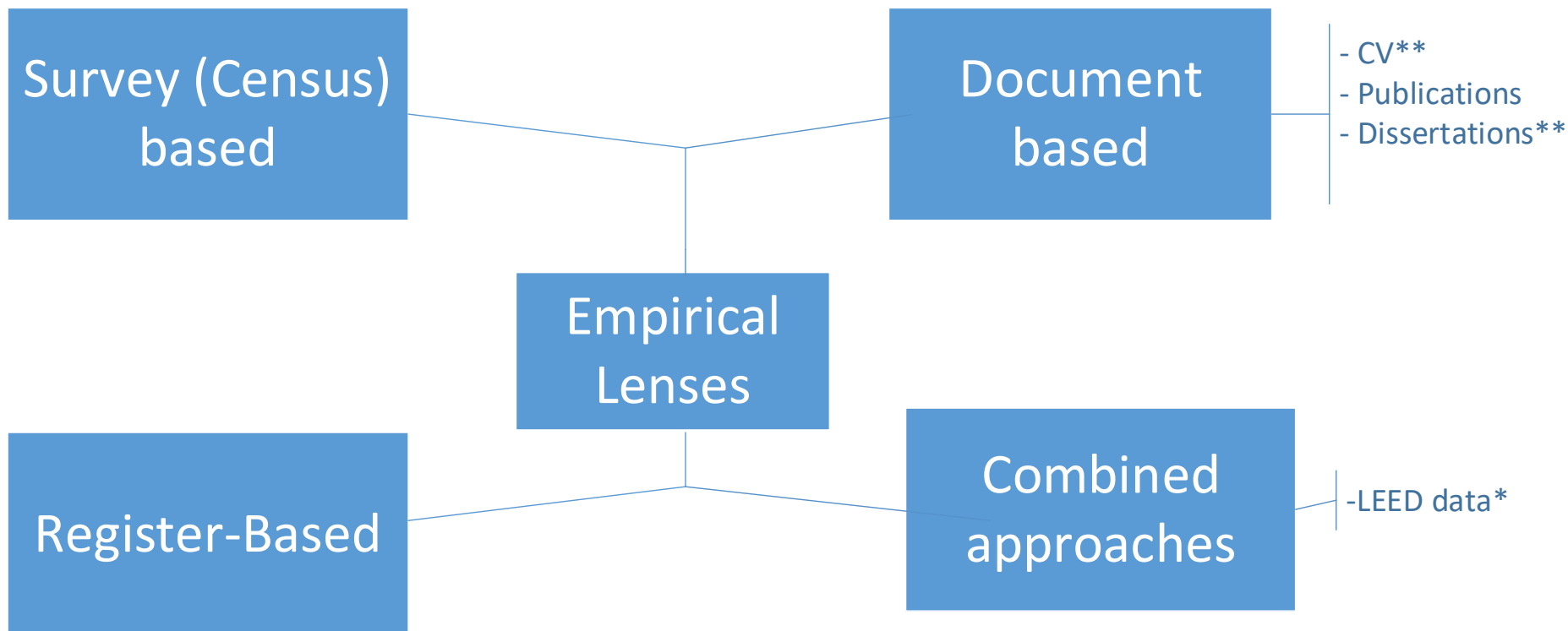
- consistency of design for international data collection and analysis
- secure framework for accessing and working with confidential data
- potential to verify estimated LM outcomes (eg) by cross-validation
- data framework to collect ancillary data
- RCF/GATE tools substantially improve analytical avenues
- scientific and data community in one
- the ability to experiment and learn

# Empirical lenses

Building Blocks

# Basic empirical approaches to study changing research careers

# RISIS







# why dissertation-based?

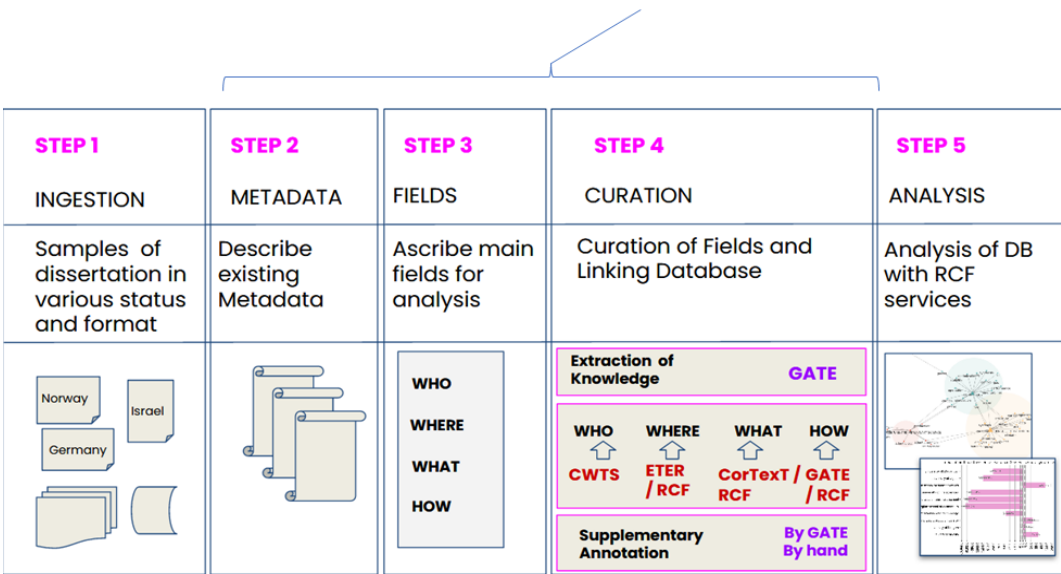


- Dissertation data are rich data sources which are available at a known rate
- Defining characteristics of the dissertation in a given context (country  $x$ , field  $y$ , year  $z$ )
  - First-order elements include:
    - its form (monograph, by article, as art-work, model, etc) and
    - its extent (nr of pages, sections, articles, references, figures/tables, etc).
  - Second order elements
    - **The Who:** primarily the candidate, but also the supervisor, and any co-authors
    - **The Where** (of involved entities): university affiliation (dissertation locus), collation with other research entities (labs, companies), funding agencies
    - **The What:** the topic of research (Jel codes, faculty, references, research questions)
    - **The How:** the models/methods used in the dissertation, including equipment (eg. tele/microscopes, software, etc).



# experimental design

component 1: Enriched PhD production database



component 2: estimation procedure to identify LM outcome

- Dissertation (metadata,text) is key
- Conceptual step
  - Conceptualize (non)academic careers of trained doctorates based on the literature
  - Design a workflow to construct DDC based on repositories of dissertations
- Baseline step
  - Establish the population-frame based on existing data
  - Anchor the DDC in existing approaches
- Implementation step
  - Pilot the DDC workflow on 5(6) countries
  - Collect/compile data based on dissertations and metadata
  - Enrich basic data with the help of RCF, GATE, CWTS, ETER
  - Estimate an indicator to proxy for career outcome
- Exploration step

# 1. enrich PhD production db

# RISIS



## Layer 1: Population frame (from register)

- Using RCF/GATE for data collection and text annotation

## Layer 2: Dissertation frame

- The dissertations provide basic counts of who studied what at which universities in which year.
- Useful and portable information especially for comparisons across contexts (country x, field y, year z)
- Value is enhanced by enriched data from RISIS data services

## Layer 3: Enrichment of DDC information from RCF/GATE

- DDC, ETER, Cortext, Geocoding GATE
- DDC and Web of Science/CWTS Link with publication (for component 2)

# 2.1 LM outcome indicator

# RISIS



**Step 1: Assume post-dissertation publication discriminates between (non-) academic careers**

- Population 1. Individuals who continue to publish with academic affiliation= “Academic”
- Population 2. Individuals who continue to publish under other affiliation = “Other Research”
- Population 3. Individuals who don’t continue to publish= ‘non-academic career’ or ‘out of LM’

**Step 2: Use RISIS to match PhDs with publication records (CWTS)**

- Automated linking of cohorts (2010, 2014) for 6 countries (country, FoS, year...)
- Curated step on subsample: supervised verification of 450 PhDs
  - Calculate precision and recall

**Step 3: Verify the true labor-market outcomes for subpopulations for cohorts**

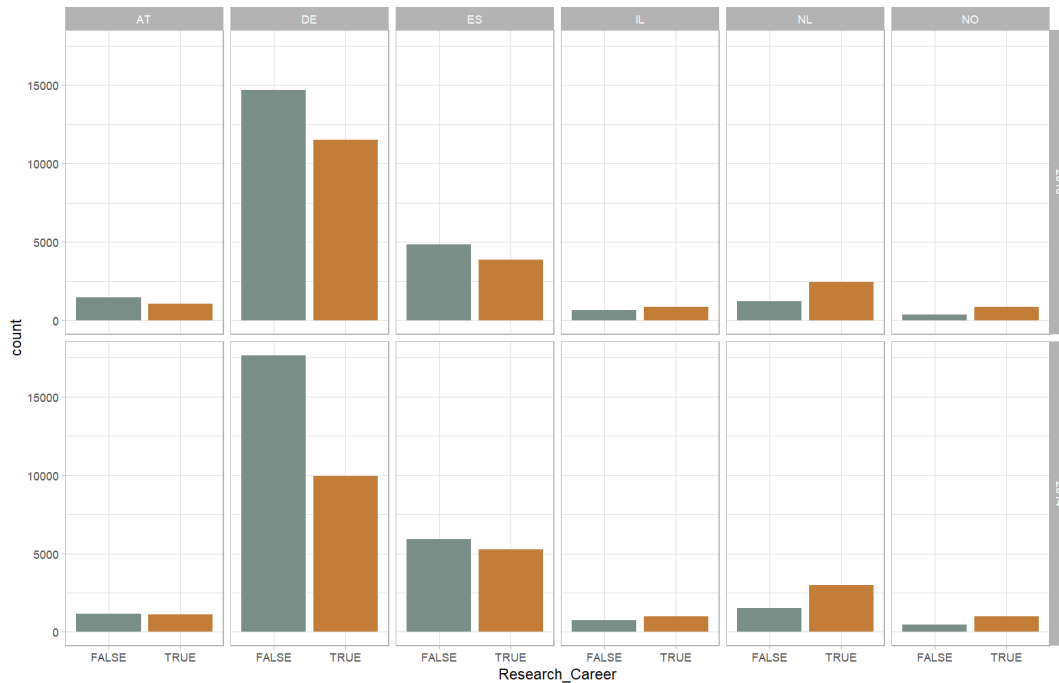
- Spain evaluates the LM outcome for the sample cohorts for 2010 & 2014
- Israel and Norway evaluate the full cohorts for 2010 & 2014.

**Step 4: Estimate a “career indicator”**

- model the proxy outcome of publication based on step2.
- model the ‘true’ LM outcome (academic sector) given predictors from above
- evaluate the accuracy of the outcomes from the models

## 2.2. CWTS Link

Use RISIS to match PhDs with publication records



	LOGISTIC	SVC	RANDOMFOREST	MLP
<b>Germany</b>				
Precision	0.91	0.80	0.78	0.87
Recall	0.70	0.65	0.68	0.71
F1	0.79	0.75	0.72	0.78
<b>Israel</b>				
Precision	0.89	0.89	0.85	0.87
Recall	0.80	0.74	0.81	0.82
F1	0.84	0.80	0.83	0.84
<b>Netherlands</b>				
Precision	0.90	0.90	0.82	0.90
Recall	0.73	0.75	0.76	0.79
F1	0.80	0.81	0.78	0.83
<b>Norway</b>				
Precision	0.92	0.89	0.84	0.89
Recall	0.85	0.89	0.80	0.87
F1	0.88	0.89	0.81	0.88

positive identification of research career  
(publication 0-5 years after degree) for each  
cohort

scores broadly support the strategy  
(note differences in recall)

## 2.3. Verify true LM outcome

# RISIS

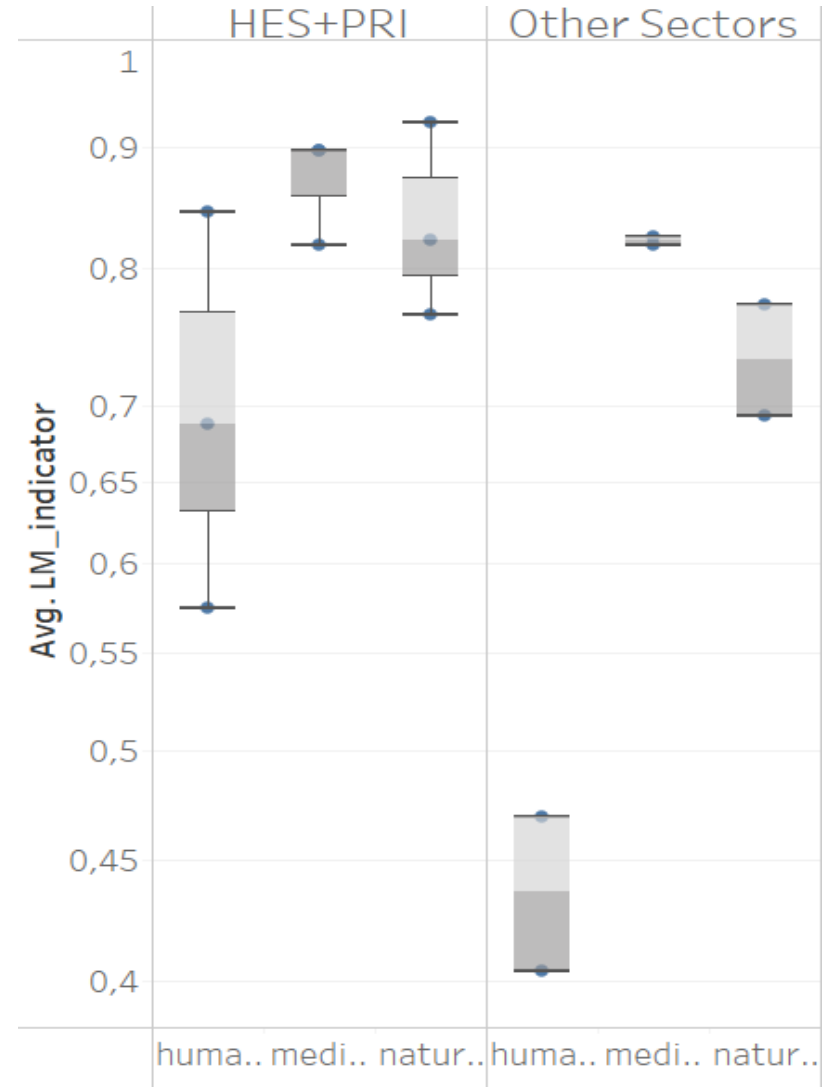


### Norwegian example

- data-source: Recruitment Monitor (Statistics Norway, 2023)
- PhD production: 2010 & 2014
- LM outcome in 2021

Publication intensity significantly higher for Academic Careers

- Differs by FoS
- Other important factors include gender, age, citizenship



# 3: inference & evaluation framework\*



## Model-based predictions

- Model Selection (Category Logistic, Mixed Model Extensions, 'Black Box' Alternatives - BART and RF...)

## Predictive Inference Framework

- Sample partition methods
- Use of full-sample 'gold standard' data

## Forecast Evaluation Metrics

- Performance of model as compared to actual observation
- 'forecast' ~ conditional probabilities given covariates of outcome category membership
- Assess Accuracy, Calibration and Refinement.

Variable Summary

Name	Description	Class	Type
AP	automated proxy	response	binary (0,1)
CP*	curated proxy	response	binary (0, 1)
LMP*†	labor market placement	response	categorical (0, 1, 2)
FoS	field of science	covariate	categorical
Yr	year of dissertation	covariate	numeric discrete {2010, 2014}
Country	country of dissertation publication	covariate	categorical
X	background characteristics of individuals	covariate	multiple
Y	characteristics of degree university	covariate	multiple
Z	characteristics of dissertation	covariate	multiple

\* CWTS curated verification for cohort samples  
 † verification of LM outcome for cohort samples



# Examples & Applications

what can the ddc do

- enriched PhD Production dataset
- career examples

# Datafication of DDC datasets

# RISIS



- Example presented by (INRAE, Paris)

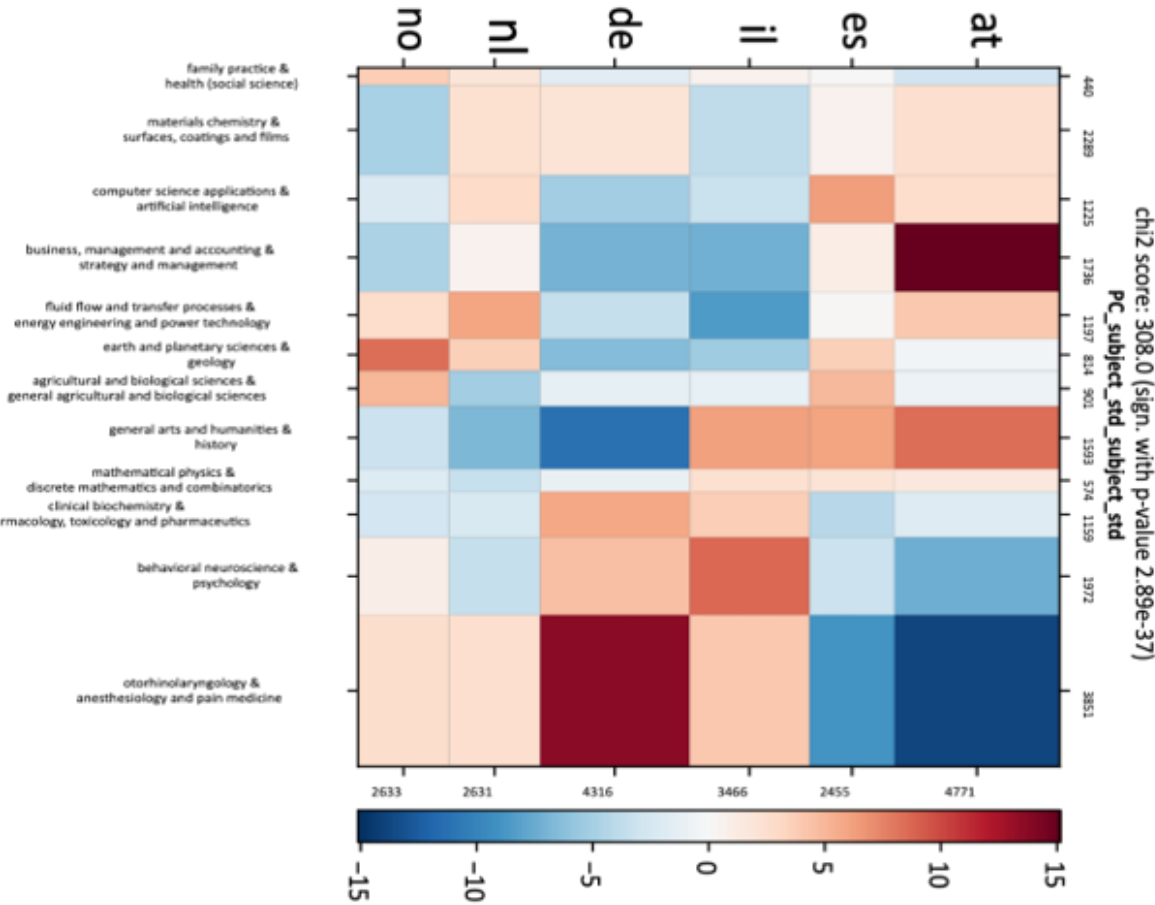
**Initial Task Gate** has deployed a service enabling NER and other extraction of information (accordingly) to be aligned in rows

## Development for the Use Case

- Integrating DDC files of various countries within a description data framework (RCF-MetaSchema)
- Added value of cleaning, normalizing (e.g.translating titles), classifying (Domain and Subject)
- Importation of EU DDC Dataset in RCF
- Analysis by RCF Scenarios to demonstrate the capacity to answer research questions

# PhD production example

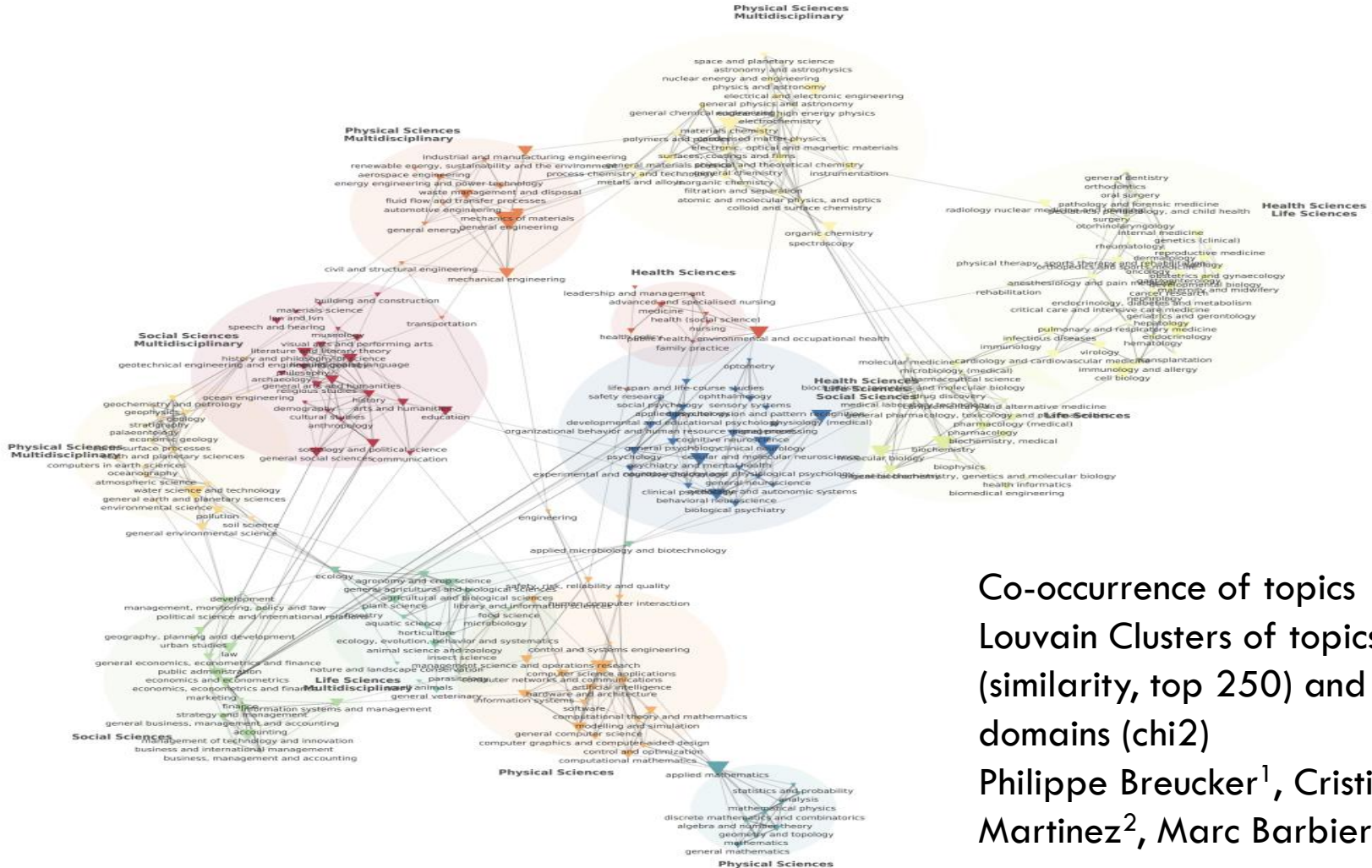
# RISIS



National specialization  
Contingency Matrix (Chi2 score) Cluster of Subjects Against Countries  
Philippe Breucker,  
Cristian Martinez,  
Marc Barbier

# PhD production example2

# RISIS



Co-occurrence of topics  
Louvain Clusters of topics  
(similarity, top 250) and  
domains (chi2)  
Philippe Breucker<sup>1</sup>, Cristian  
Martinez<sup>2</sup>, Marc Barbier<sup>1</sup>

# Postdocs abroad & career RISIS (IL)



The role of a postdoc in STEM: (Israel)

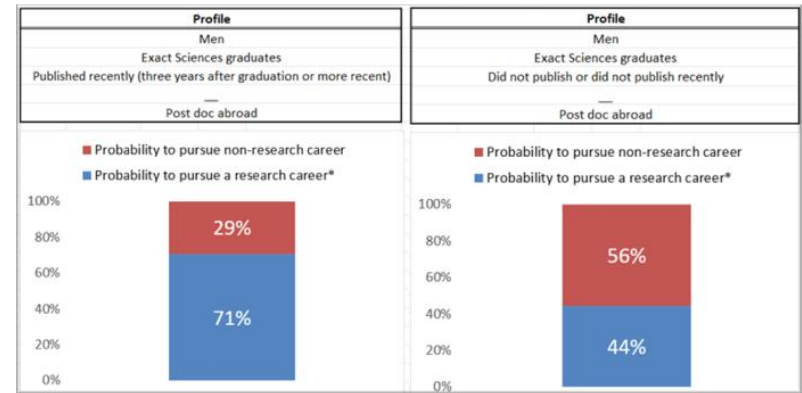
Time: 2010, 2014

Full-count approach (3450):

Eran Leck, Daphne Getz and Ella Barzani (SNI)

Career data was matched and augmented for ~93% of the population.

- Post-doc fellowship abroad and having recent publications in the WOS are strong predictors of a research career.
  - Interaction with postdoc affects predictive force of WOS
- Doctoral degree graduates originating from STEM fields more likely to pursue career outside academia



Independent variables	B	S.E.	Wald	df	Sig.	Exp(B)
WOS Field (Exact Sciences dummy)	-1.436	0.097	220.708	1	0.000	0.238
Recent publication in WOS	1.108	0.093	141.953	1	0.000	3.029
Post doc abroad dummy	1.633	0.118	190.540	1	0.000	5.121
Post doc in Israel dummy	0.770	0.157	24.081	1	0.000	2.159
Gender (Female dummy)	-0.067	0.088	0.571	1	0.450	0.935
Constant	-0.419	0.094	20.042	1	0.000	0.658

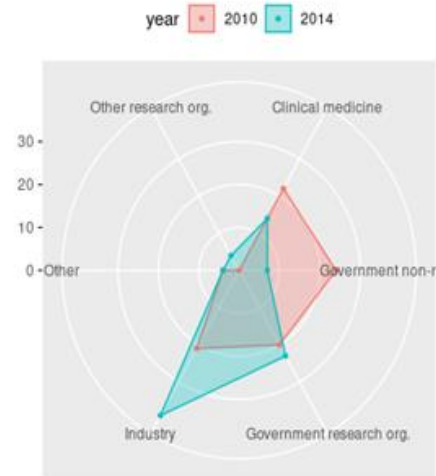
# Career Hybridization (ES)

# RISIS



- Career example: (Spain)
- Sampled approach (450):
  - hybridization (~6 percent, increasing)
  - Changes across cohorts
    - 2014 cohort more apt to move into non-academic sector

A



B

