Research Topics as a Measure

Mapping and modeling the content of science

Carlo Debernardi

ESLS - Project Colloquia

13/10/2022

UNIVERSITÀ DEGLI STUDI DI MILANO

UiA Universitetet i Agder

BEHAVE

Introduction

Research topics (RTs) as a measure, beyond simple mapping

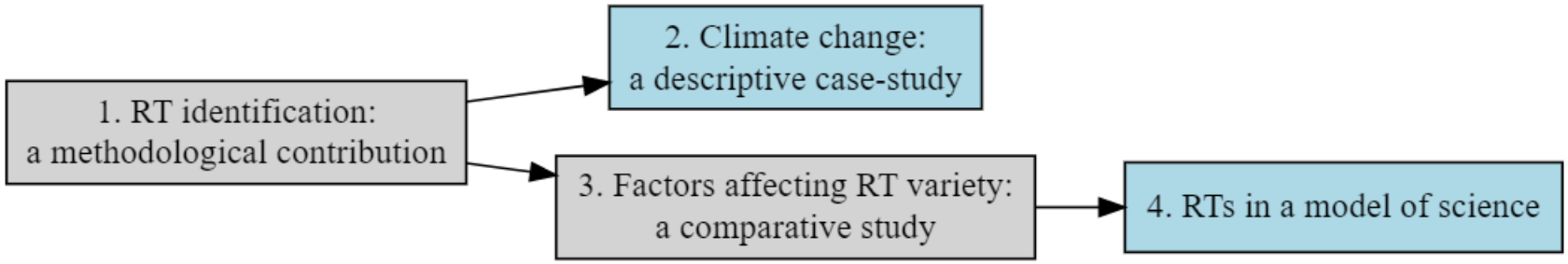
- Interesting *per se* as a phenomenon (what we do, how many different things we do, etc)
- As "control" for science studies (e.g. researchers mobility, etc)

Talking about conceptual dimensions often hinders intuition: *spatial metaphor*

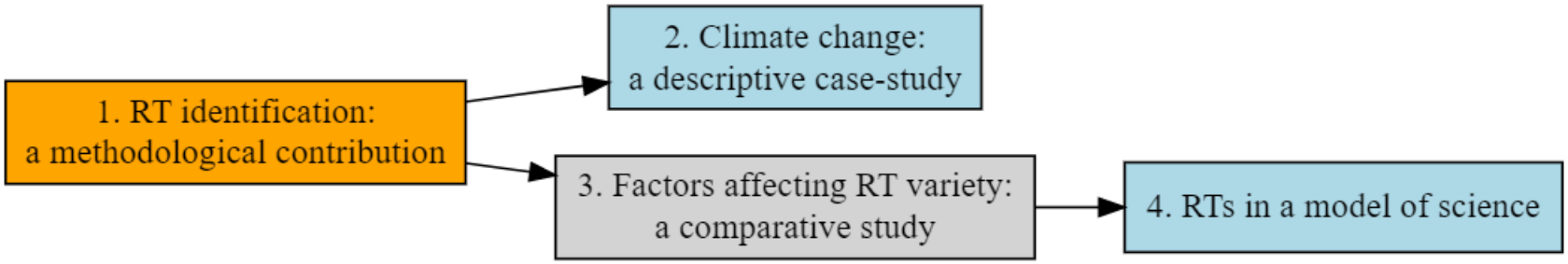
What *exactly* is an RT?

- Terminology and level of abstraction
- "Around 100 core scientists" (Kuhn 1970, Price 1986)
- Informal operationalisation

Roadmap


1. RT identification:
a methodological contribution
2. Climate change:
a descriptive case-study
3. Factors affecting RT variety:
a comparative study
4. RTs in a model of science

Roadmap


1. RT identification:
a methodological contribution
2. Climate change:
a descriptive case-study
3. Factors affecting RT variety:
a comparative study
4. RTs in a model of science

1 RT identification: a methodological contribution

Established approaches:

- Scientometrics: citational analysis (Garfield 1955, Kessler 1963, Small 1973)
- Early textual analysis: keywords co-occurrence (Callon et al. 1983)
- Topic modeling: LDA (Blei et al. 2003)

New developments:

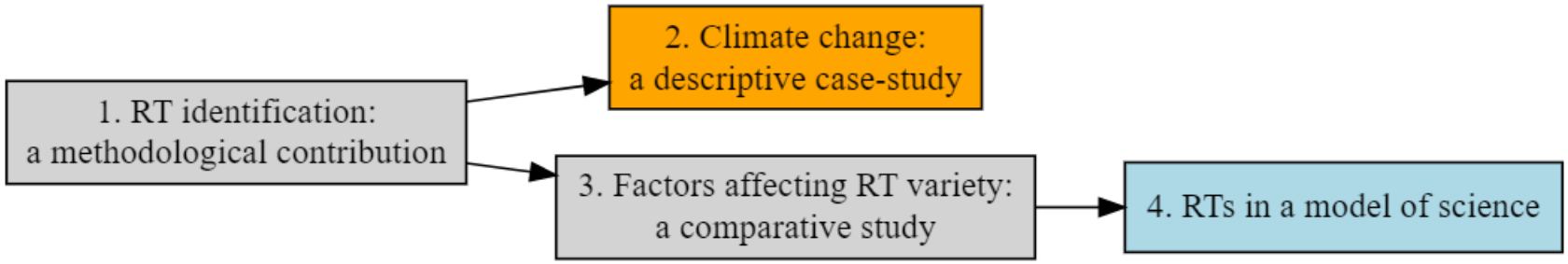
- Vector space embedding (Mikolov et al. 2013)
- Language models (Reimers & Gurevych 2019, etc)

Usage of the new approaches in science studies is still limited...

1 RT identification: a methodological contribution

- COST data (10 calls, around 4k documents)
- Not a "ground truth" but still measure of overlap with human classification
- Benchmarking of textual classification algorithms
- Comparison of short/full texts
- Proposal of a specific combination of techniques

Roadmap


1. RT identification:
a methodological contribution
2. Climate change:
a descriptive case-study
3. Factors affecting RT variety:
a comparative study
4. RTs in a model of science

2 Climate change: a descriptive case-study

Testing in an unsupervised setting (and common use-case): map a research area

Common limitations (e.g. Fu & Waltmann 2022):

- *Absolute* scientific production of only the *top countries*
- Analysis of the *5-10 main* topical trends (coarse grained)

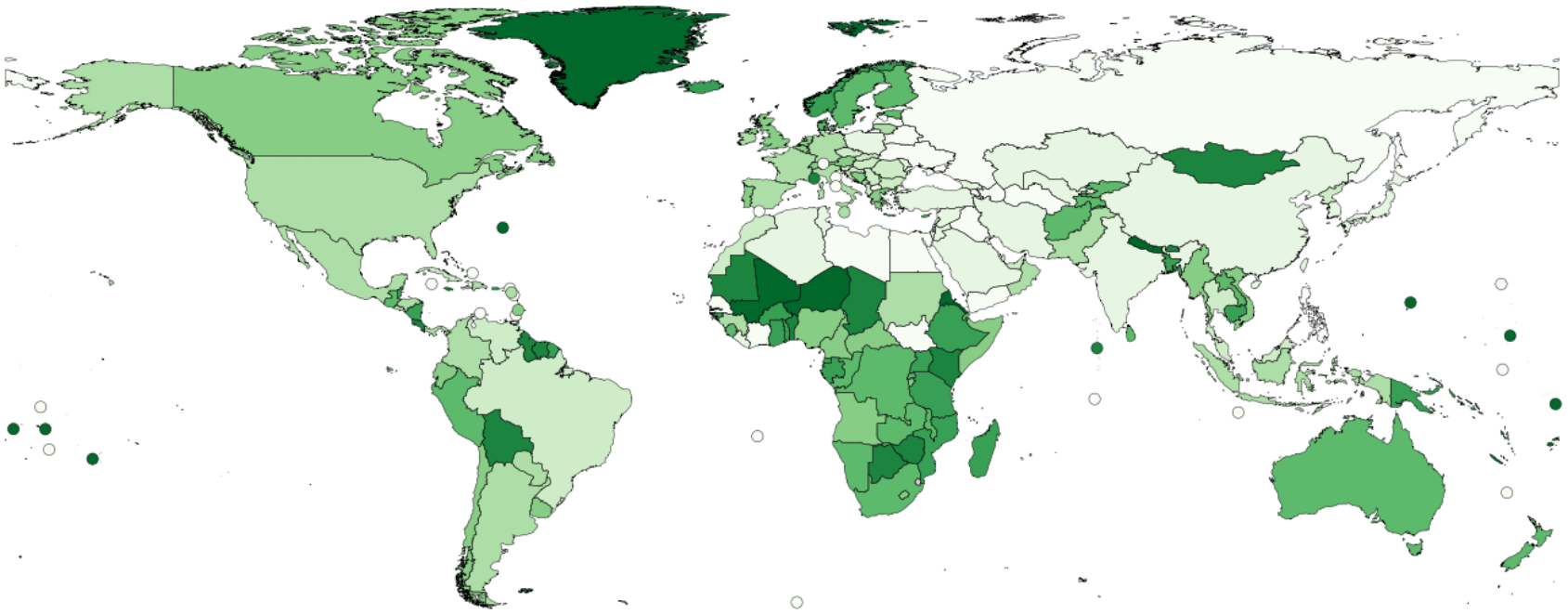
Case study: climate change research

- Public relevance
- Recent growth in size (0-1%)
- Variety of disciplines and RTs

Data:

- Scopus term search
- Time span 1990-2020
- 200k+ documents

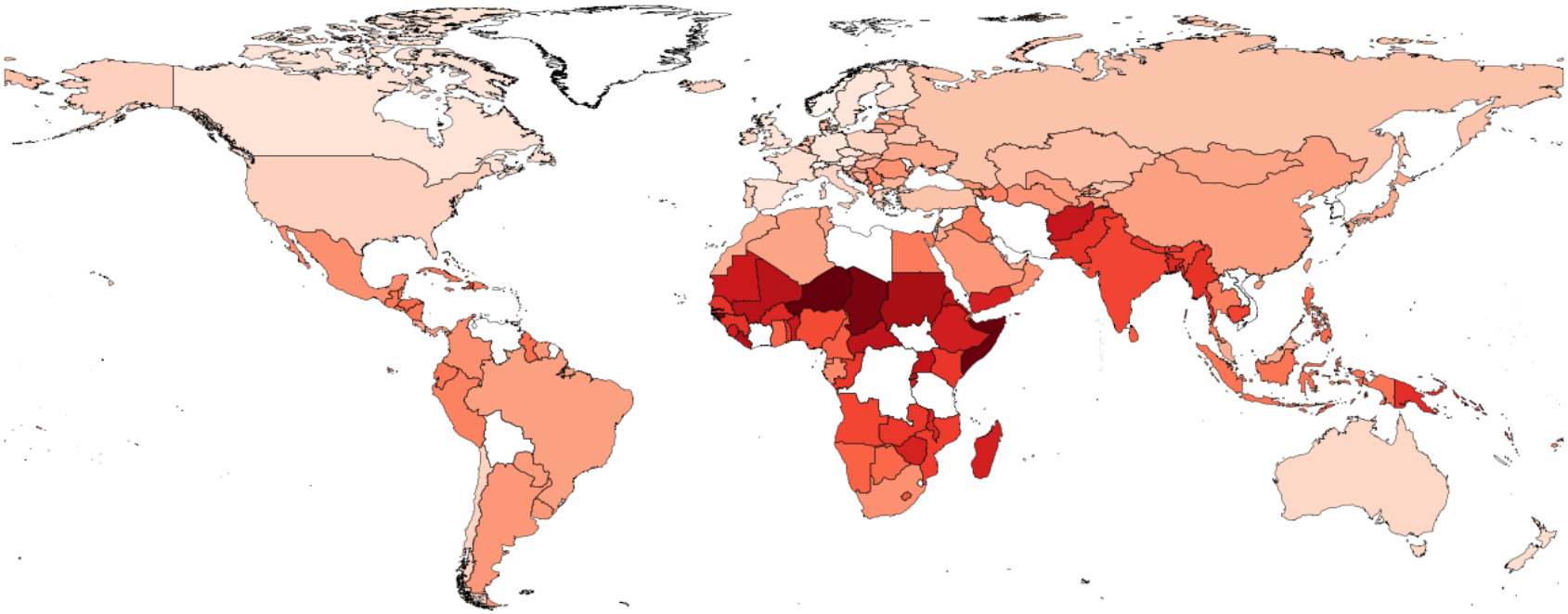
2 Climate change: a descriptive case-study



2016-2020 % of climate change research over country total publications

(darker color = higher value, max around 10%)

2 Climate change: a descriptive case-study



ND-GAIN index of vulnerability to
negative consequences of climate change

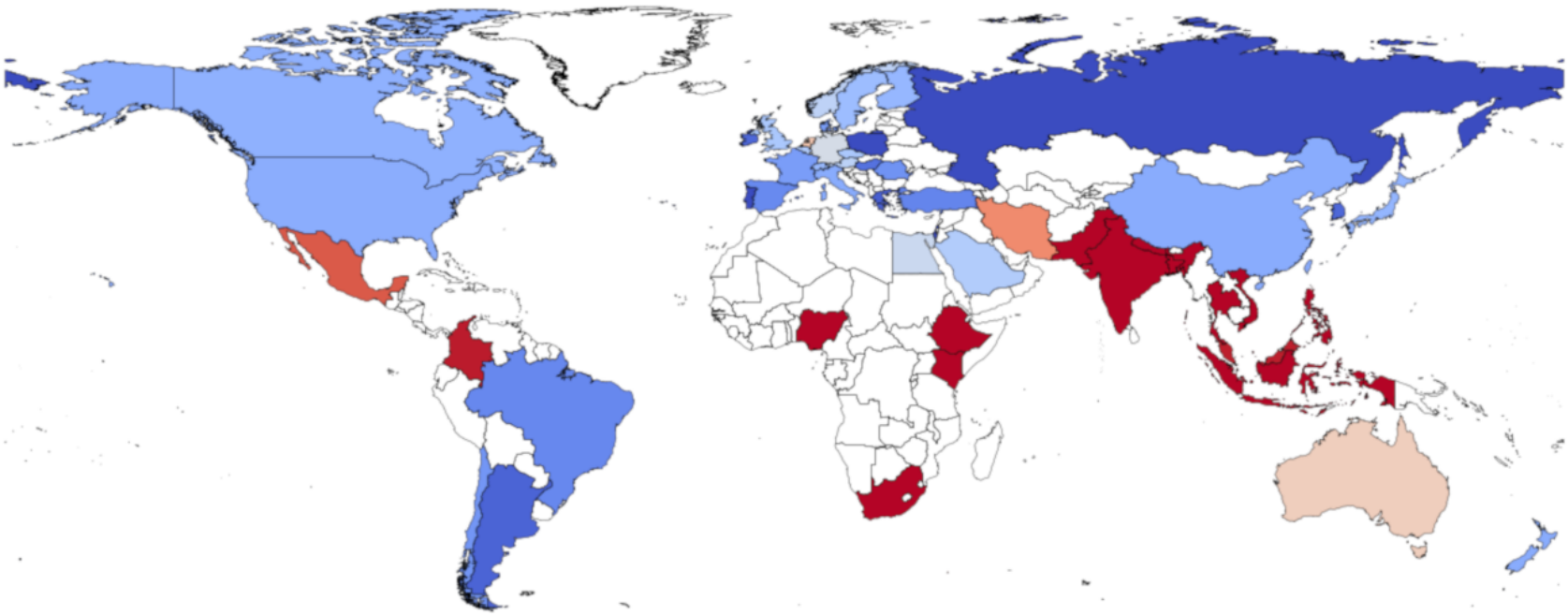
2 Climate change: a descriptive case-study

Analysis with our proposed method 800+ RTs

- Fine grained level of detail (possibly useful for practitioners, ESHS2022)
- Resolution level can be obtained bottom-up afterwards

- From absolute to relative distributions in the space
- Developing countries not only research climate change, they also have a higher representation of topics specific to them

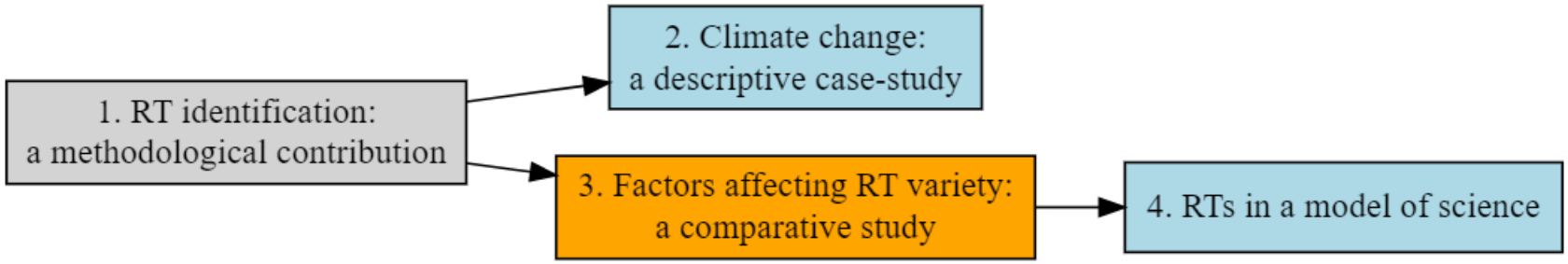
2 Climate change: a descriptive case-study



Topic 787, size 2074

Households in rural areas and small farmers, climate change consequences

Roadmap


1. RT identification:
a methodological contribution
2. Climate change:
a descriptive case-study
3. Factors affecting RT variety:
a comparative study
4. RTs in a model of science

3 Factors affecting RT variety: a comparative study

What factors influence the individual variety of research topics?
i.e. factors shaping *research agendas* (Horta & Santos 2020)

Exploration vs exploitation

March (1991), Kuhn (1977) *"The Essential Tension"*

Shifting topics and methods is *costly*, their choice implies *path dependence* (lock-in mechanism)

Factors vary (also) with: protected space, mutual dependence, task uncertainty (Whitley 2000)

3 Factors affecting RT variety: a comparative study

Scopus data, focus on textual data

Cross country comparison

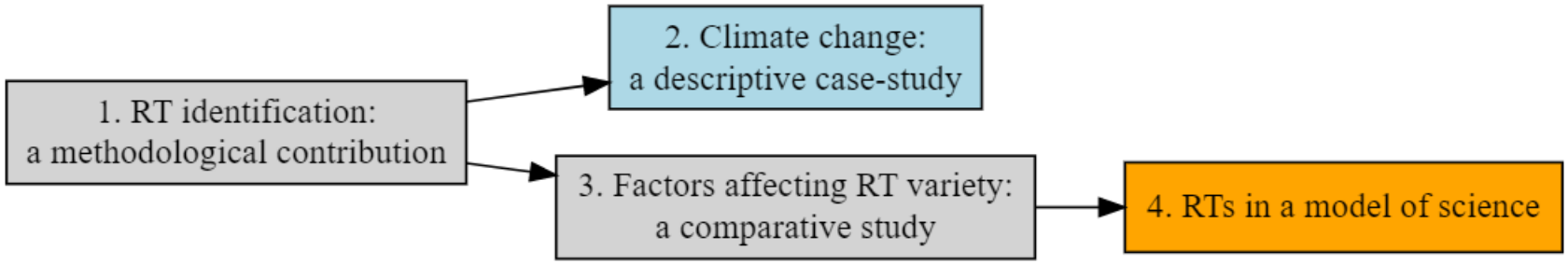
- Italy, Norway, Germany, United Kingdom
- Different career path, funding level, implementation of evaluation policies

Cross discipline comparison

- Sociology, StatMec and Nonlinearity, Molecular Biology
- Different across dimensions of Whitley's typology

Data collection is ongoing

Roadmap


1. RT identification:
a methodological contribution
2. Climate change:
a descriptive case-study
3. Factors affecting RT variety:
a comparative study
4. RTs in a model of science

4 RTs in a model of science

We draw on existing ABM literature on peer-review: Thurner & Hanel (2011), Squazzoni & Gandelli (2012, 2013), Bianchi & Squazzoni (2022)

Limit

Intrinsic quality of manuscripts as proportional (albeit in a noisy way) to the resources the author can access

Our proposal

Resources are an enabling factor, but the final quality of a contribution depends on the specific combination of TDM (theory, data, methods etc)

How to model this complexity?

Spatial representation of an *epistemic landscape*: Weisberg & Muldoon (2009), Alexander et al. (2015), Sobkowicz (2017), Avin (2019)

4 RTs in a model of science

Building blocks: Agents (authors & reviewers), Landscape, Journals, Resources

Algorithm 1 Main loop of the model

while $t \leq T$ **do**
- Compute fitness (scientific significance) of TDM
- Set manuscript quality $q_m = min\{fitness, resources\}$
- Determine journal for submissions
- Match author-reviewers
- Set evaluated quality $q_e \approx f(q_m, distance_J, distance_R, status)$
- Publish top-p papers per journal, according to q_e ranking
- Update (limited) resources according to publications
- Update NK-potential fitness (priority rule)
- Move authors (local search)

end while

4 RTs in a model of science

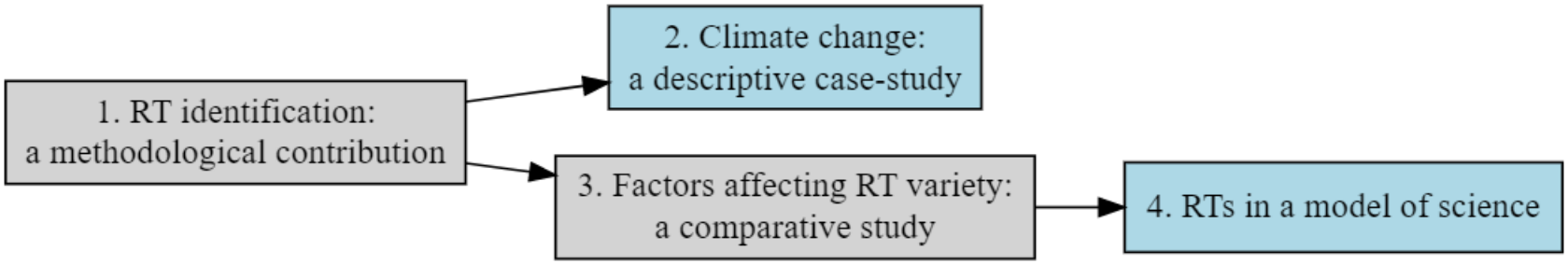
(The model was revisited after SSC2022 conference)

We are now running preliminary simulations

Main insights:

- The evolution of the field matters
- Relative scarcity/abundance of resources seems to have big impact on the role of other mechanisms

Plans for the future: calibration on empirical data from the previous contribution (!)



Thanks for the attention

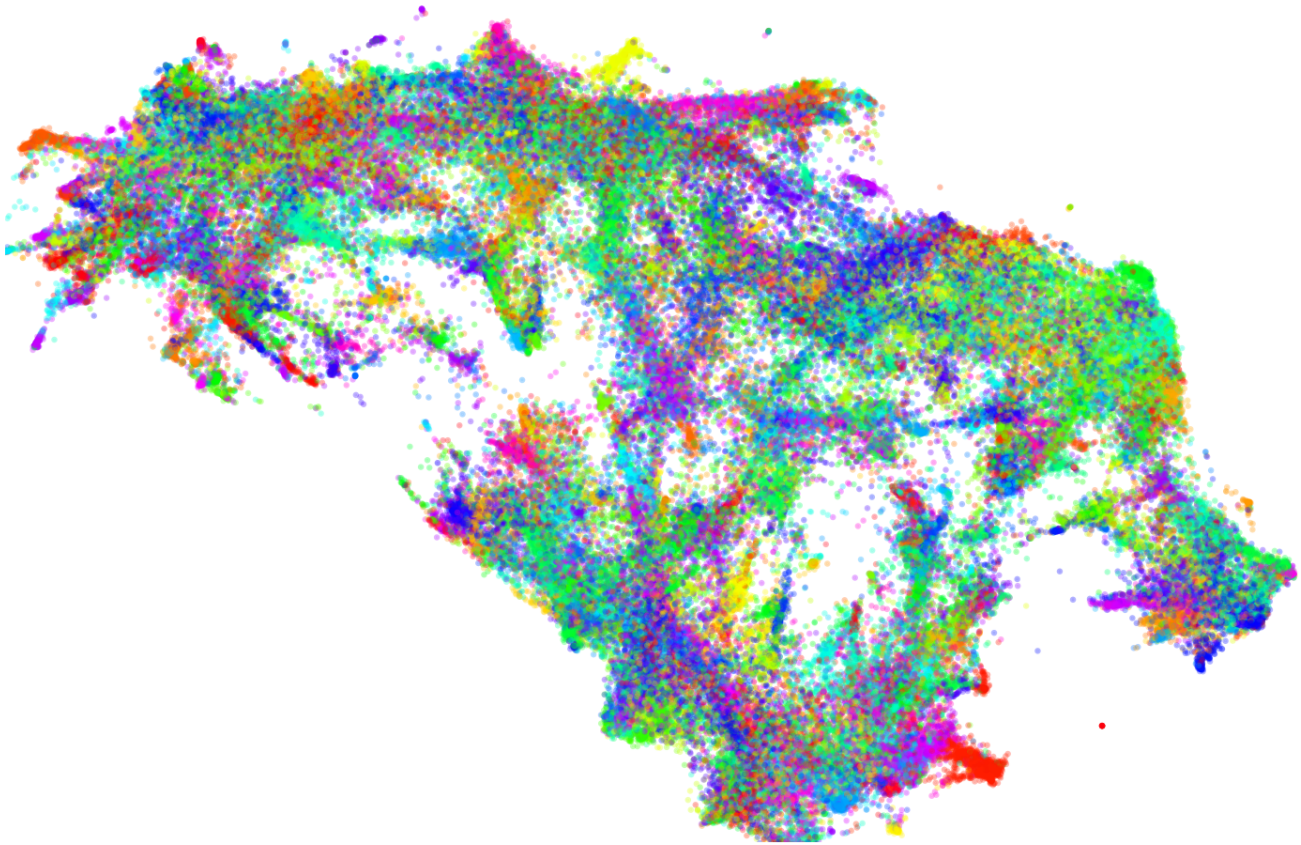
Appendix

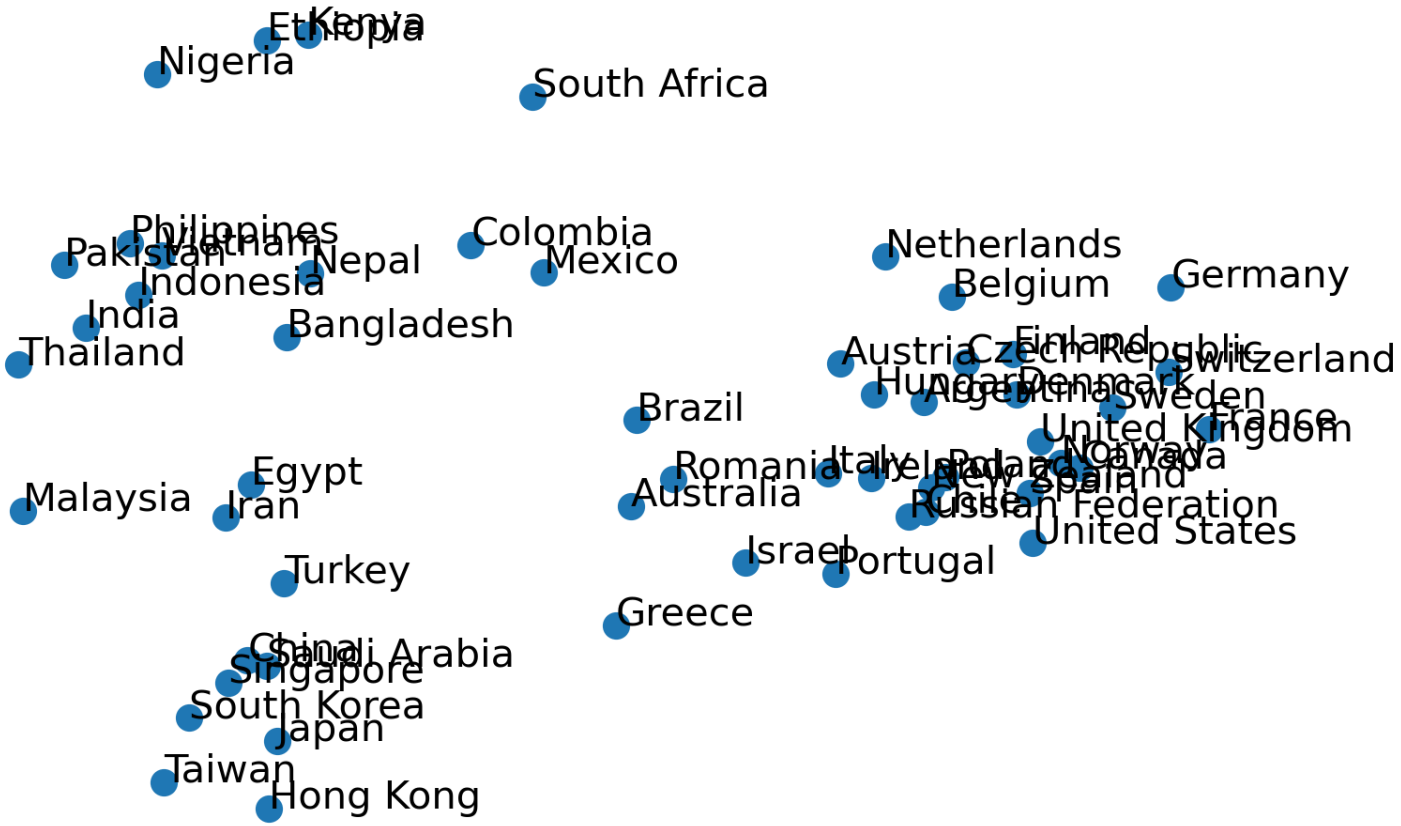
RT identification: our pipeline

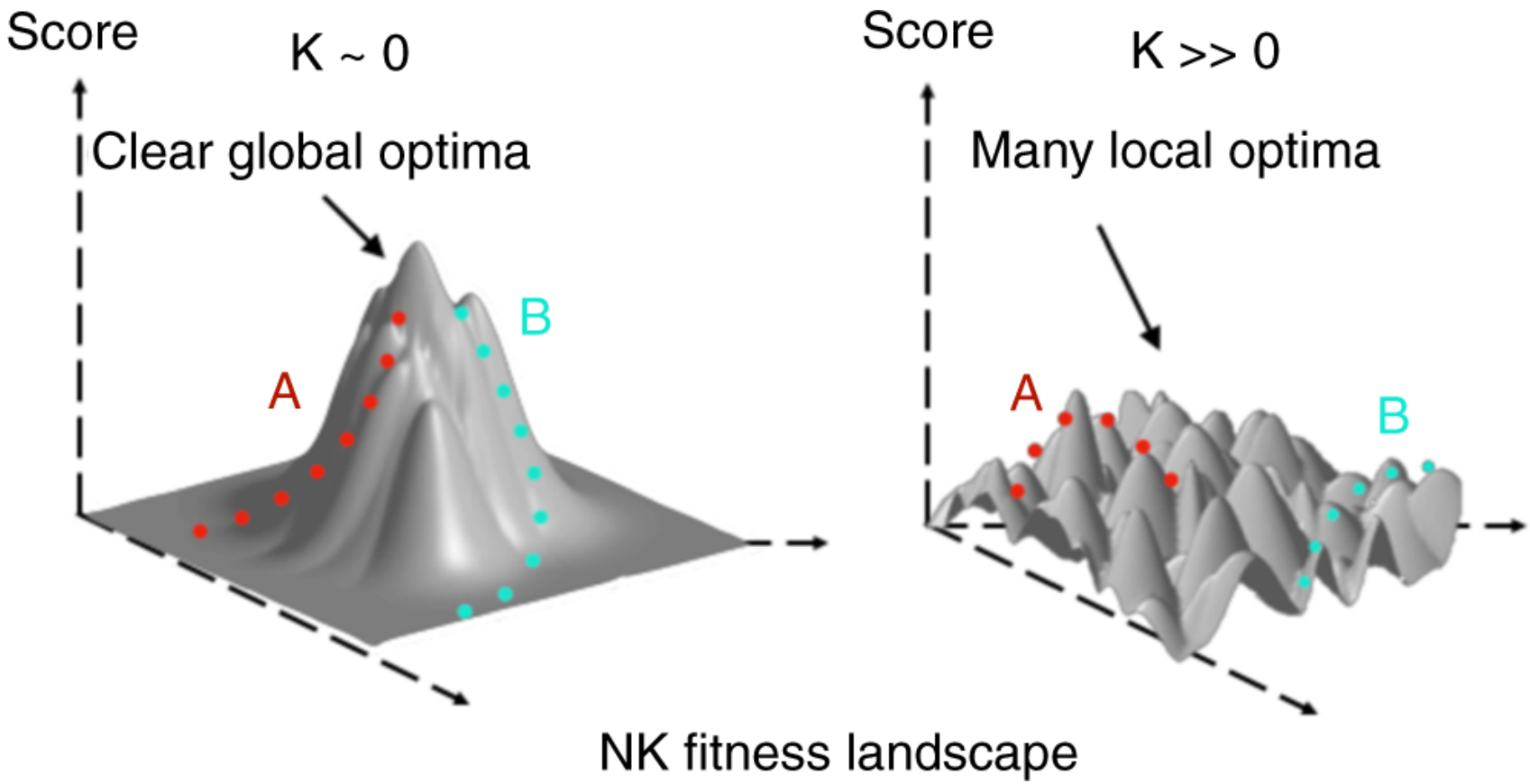
- Document embedding (Reimers & Gurevych 2019)
- Repeated (20 iterations) application of dimensionality reduction (UMAP, McInnes & Healy 2018) and density-based clustering (HDBSCAN, Campello et al. 2013)
- Creation of a co-occurrence matrix of the documents (i.e. how many times doc A and B are labeled in the same cluster)
- Application of a threshold on the matrix in order to link only the documents always appearing together and identification of the connected components. These are the "core" part of the clusters
- The labeling is extended to the unlabeled documents via the following procedure:
  - pick a random unlabeled document with at least a link as strong as 1/2 the maximum (i.e. 10)
  - assign the label via a voting weighted by the links strenght
- The remaining unlabeled are those that where considered noise 50% or more of the times, these are assigned to the cluster with the closest centroid in the embedding space
- All the documents are color-coded in order to provide information on the quality of the clustering

Color-coding of the documents:

- Green: part of the core of a cluster, always labeled
- Yellow: labeled at least 50% of the times
- Orange: labeled less than 50% of the times
- Red: always considered noise




Kenya
Ethiopia
Nigeria
South Africa
Philippines
Vietnam
Pakistan
Nepal
Indonesia
India
Thailand
Bangladesh
Colombia
Mexico
Netherlands
Belgium
Germany
Finland
Austria
Czech Republic
Switzerland
Hungary
Denmark
Sweden
Brazil
United Kingdom
France
Norway
Romania
Italy
Poland
Egypt
Malaysia
Iran
Australia
Russian Federation
United States
Israel
Portugal
Turkey
Greece
China
Saudi Arabia
Singapore
South Korea
Japan
Taiwan
Hong Kong


Score
K ~ 0
Clear global optima
A
B
Score
K >> 0
Many local optima
A
B
NK fitness landscape