Challenges and Solutions in the Construction of Chinese Patent Database

Deyun YIN
deyun.yin@wipo.int
Research Fellow
Innovation Economics Section,
Economic and Statistic Division,
World Intellectual Property Organization

2019/12/07

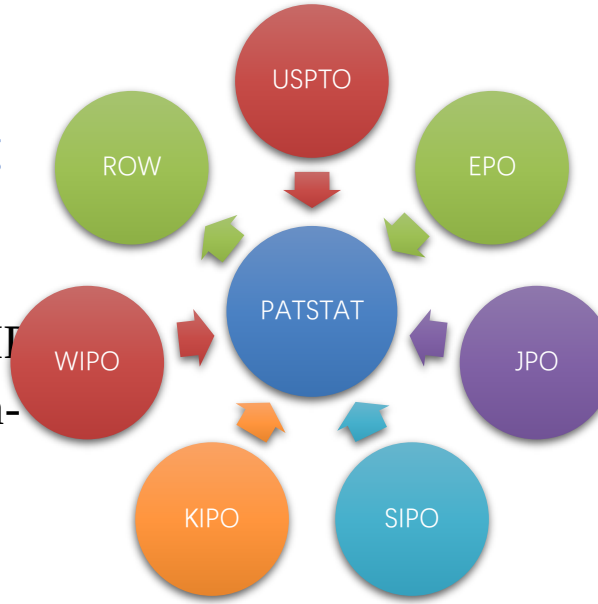
Contents

- **Topic1: Construction of CNIPA database: challenges, solutions and recent progress**
  - **Inventor name Disambiguation**

  make individual-level study of Chinese innovation possible
  - **Standardizing applicants' names**
  - **Construction of citation database**

- **Topic 2: Improving PATSTAT's coverage, geocoding and map global innovation network of innovation hotspots**

  1. Supplementing missing information from CNIPA, JPO, KI
  2. International patents (PCT or patent family >=2 or foreign-oriented patents)
  3. identification of technological clusters with DBSCAN
  4. Mapping global innovation network



Contents

- Topic 1: Construction of CNIPA database

- Inventor name disambiguation
- Applicant name harmonization
- Citation

- Topic 2: improving PATSTAT database geocoding and mapping global network of innovation hotspots

Introduction

Why CNIPA data?

1. **World 2nd largest database.**

2. **Important for studying China's innovation**

   Coverage of Chinese patent in major databases in USPTO, EPO, JPO

Introduction

But CNIPA data suffers from the following limitations:

	Challenges	Solutions
Inventors	Lack Unique identifiers Lack inventor addresses	● USPTO: Patentsview ● EPO: Pezzoni, et al., 2014 ● JPO: Ikeuchi et al., 2018 ● SIPO: Yin et al., 2019 ● KIPO
Applicants	No harmonized names	● USPTO ● EPO ● JPO: IIP company name list ● SIPO: CPDP (He et al., 2018)
Citations	Lack of reliable citation data	● Patent search report ● Extraction of within-text citation by IncoPAT

Contents

- **Topic 1: Construction of CNIPA database**
  - **Inventor name disambiguation**
  - Applicant name harmonization
  - Citation

- **Topic 2: improving PATSTAT database geocoding and mapping global network of innovation hotspots**

Inventor Name Disambiguation
Introduction

Scientometrics
https://doi.org/10.1007/s11192-019-03310-w

Check for updates

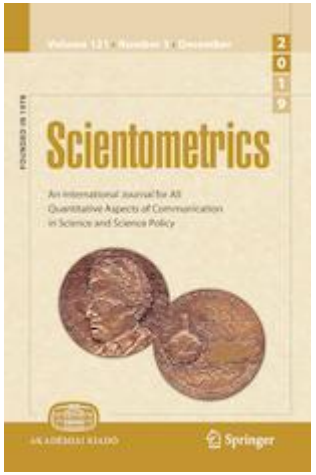
Large-scale name disambiguation of Chinese patent inventors (1985–2016)

Deyun Yin[1] · Kazuyuki Motohashi[1] · Jianwei Dang[2]

Received: 16 April 2018

Deyun YIN, Kazuyuki MOTOHASHI, Jianwei DANG (2019). **Large-scale Name Disambiguation of Chinese Patent Inventors (1985-2016)**, Scientometrics, p1-26.

Abstract
This study presents the first systematic disambiguation result of Chinese patent inventors in State Intellectual Property Office of China patent database from 1985 to 2016. With a list of 66,248 inventors owning rare names and a hand-labeled data of 1465 inventors, our supervised learning algorithm identified 3.99 million unique inventors from 1.84 million Chinese names referring to 14.68 million patent-inventor records. We developed a method for constructing high-quality training data from a third-party rare name list and provided evidence for its reliability when large-scale and representative hand-labeled data is crucial but expensive to obtain. To optimize clustering results on large-scale dataset with highly unbalanced distribution, we also modified robust single linkage by adding constraints to the maximum distance within clusters generated. Varying across different training and testing data, as well as clustering parameters, our algorithm could yield F1 scores to 93.36% before clustering and 99.10% after clustering, with final splitting errors of 1.05–1.34% and lumping errors of 0.21–0.83%. Besides, we also applied this framework in standardizing applicants' names according to their text similarity and geographical information based on the high-resolution geocoding data of all addresses within mainland China.



Keywords Disambiguation · Patent · Inventor · Machine learning · Gradient boosting decision tree · Single linkage

Inventor Name Disambiguation
Introduction

Background

1. **Productivity:** identifying and targeting productive ("star") engineers (prolific or with higher influence) and examine factors that influence organizational or inventor's productivity
2. **Mobility:** Tracking inventor's patenting career and mobility among institutions and regions, examine determinants of inventor mobility and its impact on knowledge spillover or firm performance.
3. **Networks:** inventor teams, collaborative and citation networks .

Why should we disambiguate Chinese inventors?

- CNIPA: World largest patent database and rising share of papers & patents published by Chinese
- East Asian names, especially Chinese names: most challenging problems

What is disambiguation? Identifying "who owns which"

Goal: Disambiguate Chinese inventors and give an unique identifiers for each different inventors in SIPO .

Inventor Name Disambiguation
Introduction

Common Challenges & roots of confusion

Western Names

- **Synonym:** different names referring to one person:
  - Typographical errors/misspelling, name variant, and abbreviations
  - e.g.
  - Lee Fleming – 5,029,133 (1991)
  - Lee O. Fleming -- 5,136,185 (1992)
- **Major issue for disambiguating western names, negligible for Chinese names:** 2~2.5‰ synonym problem in rare names, fewer than 2‰ for common names as rare names are prone to misspelling or miswritten.

Eastern Asian Names

- **Homonym:** same name referring to different person:
  - Repetition of common names (同姓同名)
  - e.g. name "张伟"
    - 9,684 patent records (would generate 46,885,086 comparison pairs).
    - 61,037 papers in Wanfang database
  - **most severe problems in Chinese names** as Chinese names are less variate.
  - → highly unbalanced clustering task and large lumping errors generated by single linkage clustering

Unique challenges:

1. **Lack inventor address:** only address of the first applicant.
2. **No training and testing set:** need to choose appropriate training set & models

Inventor Name Disambiguation
Net Contribution

- **Methodological Contribution:**

① An algorithm customized according to characteristics of Chinese names;

② An evidence for substituting hand-labeled records with rare name data when large-scale, representative and unbiased hand-labeled records is impossible to collected.

③ Robust Single Linkage with dynamic constraints: adding constraints to the maximum distance on clusters generated by RSL→ a robust and flexible method to avoid creating excessively large clusters with high lumping errors.

- **Data contribution:**

① A cleaned Chinese inventor dataset with classification of ethnicity (Chinese names, Japanese, and other foreign names);

② 21,359 hand-labeled records participated by 1,465 academic and industrial inventors from a broad range of technological fields for evaluating and comparing performances of disambiguation classifiers.

③ a list of 66,248 inventors owning real rare names (0.4 million records)

④ 1.84 million Chinese names → 3.99 million unique inventors (14.68 million)

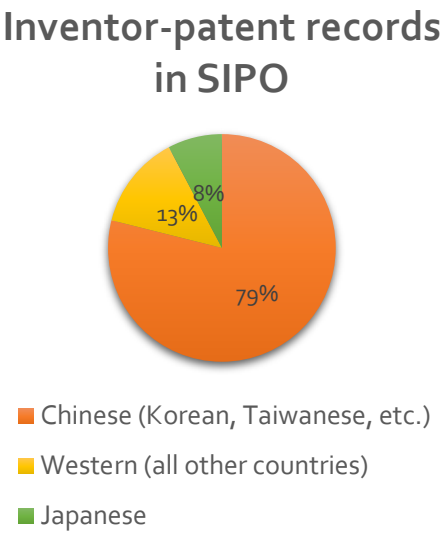
Inventor Name Disambiguation
Literature Review

Methodology	Author	Application	Training Set	Algorithms	Evaluation	Claimed Result
Rule-based (Heuristic)	Lai et al., 2009	USPTO 1975~2008	NA	Similarity profile + *ad hoc* threshold	Extension of Torvik et al. (2005) and the first dataset public available	Splitting: 9.18%, Lumping: 0.76%
	Zhang et al., 2014	SIPO 2000-2009	NA	String matching of applicant, province, IPC class	First and unique attempt that disambiguated Chinese inventors	NA
	Pezzoni et al., 2014	PATSTAT 2011	NA	Similarity profile + *ad hoc* threshold and weight	First disambiguation of PATSTAT data	Precision:74%, Recall: 70%
	Morris, 2017	PATSTAT 2014	NA	using high-resolution geocoding data	Simple, straightforward rules that disambiguate assignees and inventors at the same time	Splitting: 10.5%, Lumping: 9.5%
Semi-supervised	Li & Lee Fleming, 2014	Full USPTO	Statistically generated labels with rare names as a part of input data	Naïve Bayes + *ad hoc* rules automatically generated training sets	Pioneering work based on machine learning method	Splitting 3.62%, Lumping: 2.34%, F1 score from 3rd party: 92.7314%
	Ikeuchi et al., 2017	Japanese inventors in JPO	Rare name data as a part of input data	Naïve Bayes + *ad hoc* rules	First systematic disambiguation of Japanese patent data	Splitting: 2.41%, Lumping: 0.29%
Supervised	Ventura, et.al., 2015	A small subset of USPTO	Optoelectronics + Academic life scientists	Random Forest + Hierarchical clustering	First supervised method while the training set have large biases when applying to whole USPTO dataset	Splitting: 2.09%, Lumping: 1.26%
	Kim, Khabsa, & Giles, 2016	Full USPTO	Random mixture of IS and E&S dataset	Random Forest + DBSCAN	Similar to method of Ventura, et.al., (2015) with refinement such as detailed feature selection	Precision: >99%, Recall: 97%,
Unsupervised	Balsmeier et al., 2016	Weekly updated Full USPTO	NA	K-Means clustering	Completely automated process	NA

Inventor Name Disambiguation
Data

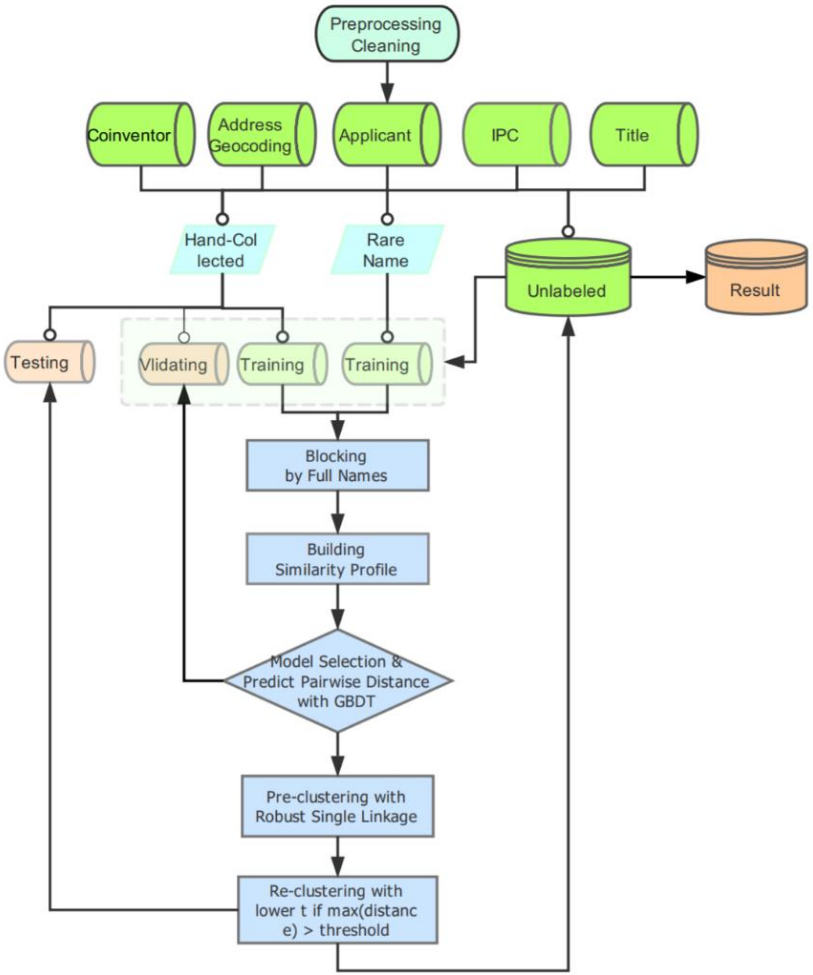
Summary of nationality identified by names in full SIPO patent inventor data

Names of country	Unique patents	Unique names	Inventor-patent pairs	Ida_seq /names	Max	Criteria
Chinese (Korean, Taiwanese, etc.)	4,9 m	1,84 m	14,685,617	7.96	9680	No points within name, do not have a Japanese Family name and the length of name equal to or lower than 4 characters
Western (all other countries)	0.899 m	1,17 m	2,492,036	2.14	509	With a point in names, e,g. “P·T·贾特”, “D·罗布”
Japanese	0.569 m	0.35 m	1,435,067	4.05	1402	With typical Japanese Family name e.g. “伊藤彰浩” , “毒岛真”
In Total	6,25 m	3,36 m	18,6 m			



Inventor Name Disambiguation
Data

Unit of Analysis:	**inventor-patent pairs.** • **Ida_seq = Patent id + inventor sequence**
Input	Inventor characteristics • Inventor names
	Patent characteristics • Applicant names: job • Applicant address • Geocoding of applicant address (Baidu Map's API) • All inventors: co-inventors • IPC code: technological area • Title: technological area
Process: pairs for comparing ida_seq	• Non-match: 0 = two ida_seq pairs belonging to 2 different person • Match: 1= two ida_seq pairs belonging to the same person
Output	• Inv_id = Unique inventor identifier



Inventor Name Disambiguation Algorithms

ntor	id	ida_seq	name	inventor_all	ipc_class	ipc_group	title	la
丁友昉	2068	03100363-1	['丁友昉']	['丁友昉', '王德培', '刘廷志']	['A23K', 'A...	['1-14', '1-...	多种微生物秸秆发酵饲料与...	39.07842290307
丁友昉	2068	03105080-1	['天津科技大学']	['丁友昉', '杜连祥', '王德培'...	['A23K']	['1-14']	利用多菌种联合厌氧发酵生...	39.07842290307
丁友昉	2068	200410037613-2	['包头天赋食品']	['方向东', '丁友昉', '秦刚']	['C12N', 'A...	['1-20', '1-...	多菌种微生物发酵剂的制作...	40.60827433188299
丁友昉	2068	200410072845-1	['丁友昉']	['丁友昉', '秦刚', '王德培']	['C12G']	['3-02']	发酵虫草、枸杞保健酒及制...	39.08885978382113
丁友昉	2068	200410072995-2	['天津科技大学']	['王德培', '丁友昉', '秦刚', ...	['A23K', 'A...	['1-14', '1-...	多种微生物发酵精饲料和发...	39.08885978382113
丁友昉	2068	200610014479-2	['天津科建科技发展']	['王德培', '丁友昉']	['C12N', 'C...	['9-24', '1-...	复合微生物β一葡聚糖酶和β...	39.1135305284678
丁友昉	2068	200610015434-1	['天津达美科技']	['丁友昉', '王德培', '杨扬']	['C12N', 'C...	['9-24', '1-...	一种固态机械发酵木聚糖酶...	39.11334710771231
丁友昉	2068	200710058626-2	['天津科技大学']	['王德培', '丁友昉', '高年发']	['C12P', 'C...	['7-52', '1-...	固定化微生物发酵丙酸的生...	39.07842290307
丁友昉	2068	200710058627-2	['天津科技大学', '唐山赛纳...	['王德培', '丁友昉', '周念闽']	['A61K', 'A...	['35-74', '9...	一种用于牲畜的益生活菌微...	39.07842290307
丁友昉	2068	200710059578-3	['天津科技大学']	['王德培', '揣玉多', '丁友昉'...	['A23K', 'A...	['1-00', '1-...	微生物发酵的紫甘薯饲料及...	39.07842290307
丁友昉	2068	200810152951-2	['北京阔利达生物技术开发', ...	['王德培', '丁友昉', '李桂祥'...	['A23K']	['1-14']	发酵废弃蔬菜饲料和制备方法	40.4164229624551
丁友昉	2068	200910067649-2	['北京阔利达生物技术开发', ...	['王德培', '丁友昉', '李桂祥'...	['A23K']	['1-16']	微生物活菌饲用添加剂及制...	40.4164229624551
丁友昉	2068	201410206221-1	['丁友昉']	['丁友昉', '王德培', '刘廷志'...	['A23K', 'A...	['1-00', '1-...	多种益生菌秸秆发酵饲料的...	39.08066616659715

Input data

Sample of Similarity profile

ida_seq1	ida_seq2	match	app_i	inventor_i	ipc_c	ipc_g	applicant	geo	address	title	keyword1	keyword2
200610014479-2	201510139861-2	0	0	0	0	0	0	1	.2631579	0	0	0
200610014479-2	200610015434-1	1	0	2	8	3	0	5	.9473684	.4487628	0	0
200610014479-2	200710058626-2	1	0	2	2	0	0	3	.5714286	.55	0	0
200610014479-2	200710058627-2	1	0	2	0	0	0	3	.5714286	.4118687	0	0
200610014479-2	200710059578-3	1	0	2	3	2	0	3	.5714286	.5343137	0	0
200610014479-2	200810152951-2	1	0	2	0	1	0	1	.2916667	.491453	0	0
200610014479-2	200910067649-2	1	0	2	0	0	0	1	.2916667	.5861111	0	0
200610014479-2	201410206221-1	1	0	2	0	1	0	3	.3333333	.3507615	0	0
200610014479-2	201510731936-6	0	0	0	0	0	0	1	.2083333	.3605556	0	0
200610014479-2	201510736690-4	0	0	0	0	1	0	1	.2083333	.3611111	0	0
200610014479-2	200610032408-4	0	0	0	0	0	0	1	.0869565	.3728632	0	0

Inventor Name Disambiguation
Algorithms

Step 4: Clustering

Algorithm: Pseudocode for our modified version of RSL with constraints (RSLcd)

Input data: precomputed distance matrix predicted by GBDT

Input: parameters in RSL: τ; k, etc.

Input: Additional user-defined parameters in RSLcd:

1) upper limit: u; $0.5<u<=1$
2) shrinking rate: r $0<r<1$
3) block size: Z Z>=2

1. Initialize the mutual reachability distance threshold τ
2. **For** each pairwise distance within block **do**:
3. If block size >Z: Pre-clustering with RSL
4. **While** there is any cluster whose max distance >= u**: do**
5. $\tau = \tau * r$ # shrink the value of τ with shrinking rate r
6. Re-clustering with new τ while keeping all other parameters unchanged
7. **End**
8. Combine new results with previous ones and update labels
9. **Return** labels

Inventor Name Disambiguation
Result

Figure 2: Feature selection: F1 score with 4 kinds of features' combinations (before clustering)

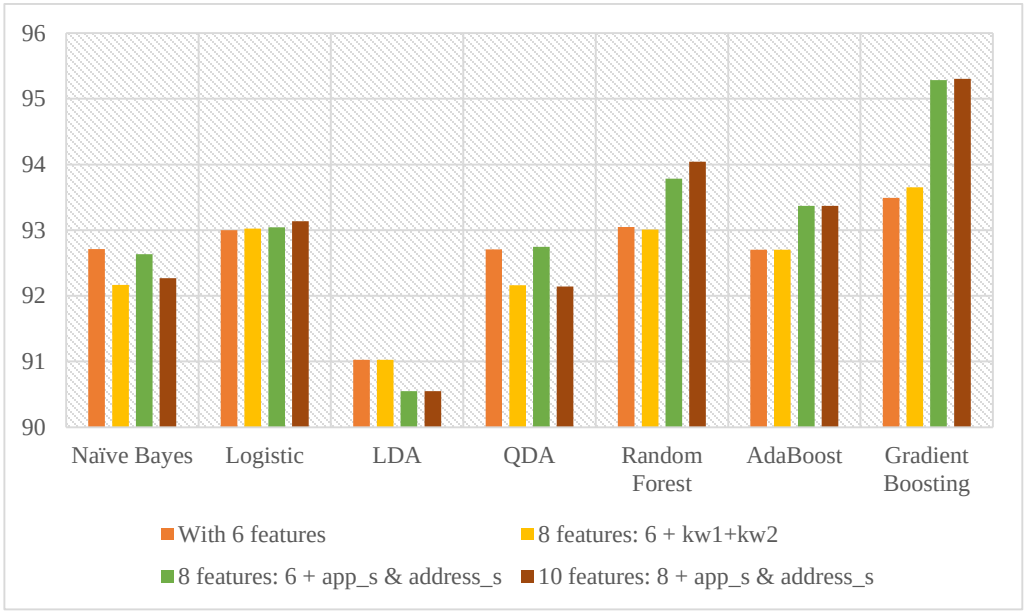
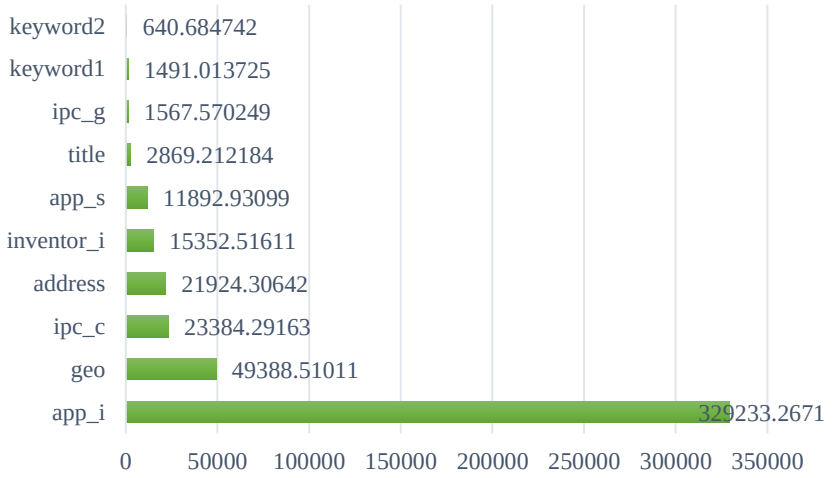


Figure 2-3 Gini importance of features in Gradient Boosting Classifier



Inventor Name Disambiguation
Result

Figure 4. Comparison of clustering algorithms (after clustering)

Inventor Name Disambiguation
Result

Table 8. Comparison of clustering methods with three kinds of training sets (Evaluation after clustering)

Training set	Clustering_τ	F1 Score (%, std)	AUC (%)	Score	Splitting Error (%)	Lumping Error (%)
Rare	**RSL_0.15**	**97.70 (0.97)**	**96.95**		**3.69**	**0.80**
	RSLcd_0.15	97.88 (1.11)	97.08		3.71	0.16
	RSL_0.5	82.16 (19.03)	84.88		1.22	59.18
	RSLcd_0.5	**99.10 (0.53)**	**97.47**		**1.34**	**0.21**
	Mean Diff	1.40***	0.52***		-2.35***	-0.60***
$Hand_{train}$	**RSL_0.15**	**93.45 (5.17)**	**93.76**		**2.79**	**18.08**
	RSLcd_0.15	98.21 (29.63)	97.43		2.86	0.53
	RSL_0.5	73.02 (0.74)	77.50		0.90	83.32
	RSLcd_0.5	**99.03 (0.48)**	**95.09**		**1.05**	**0.82**
	Mean Diff	5.58***	1.33**		-1.74***	-17.26***
Rare+ $Hand_{train}$	**RSL_0.15**	**96.84 (2.19)**	**96.21**		**3.71**	**4.09**
	RSLcd_0.15	97.82 (1.07)	96.94		3.80	0.54
	RSL_0.5	78.03 (22.77)	81.51		1.18	70.55
	RSLcd_0.5	**98.94 (0.50)**	**96.57**		**1.34**	**0.83**
	Mean Diff	2.10***	0.36***		-2.38***	-3.26***

Inventor Name Disambiguation
Conclusions

- **Algorithm:** 2-step supervised learning:
  - Step 1: Gradient Boosting Decision Tree (GBDT)
  - Step 2: Modified algorithm of Robust Single Linkage (RSLcd)

Result:

- F1 scores to 93.36% before clustering and 99.10% after clustering, with final splitting errors of 1.05%–1.34% and lumping errors of 0.21%–0.83%.
- from 1.8 million Chinese names → 3.99 million unique inventors (14.68 million patent-inventor records)

Conclusion and net contributions:

- First systematic disambiguation of Chinese inventors with machine learning techniques and a reliable dataset for inventor-level studies.
- Rare name data as qualified training set
- Dynamic constrained RSL: Inhibit the chaining effect of (robust) single linkage

Inventor Name Disambiguation
Conclusions

Statistical description of disambiguation result on all Chinese inventors in SIPO

Obs.	14.68 million	**Percentiles**	
Min	1	**1%**	1
Max	8516	**10%**	1
Mean	50.35275	**25%**	3
Std. Dev.	298.2924	**50%**	8
Variance	88978.38	**75%**	27
Skewness	18.62185	**90%**	76
Kurtosis	443.5064	**99%**	642

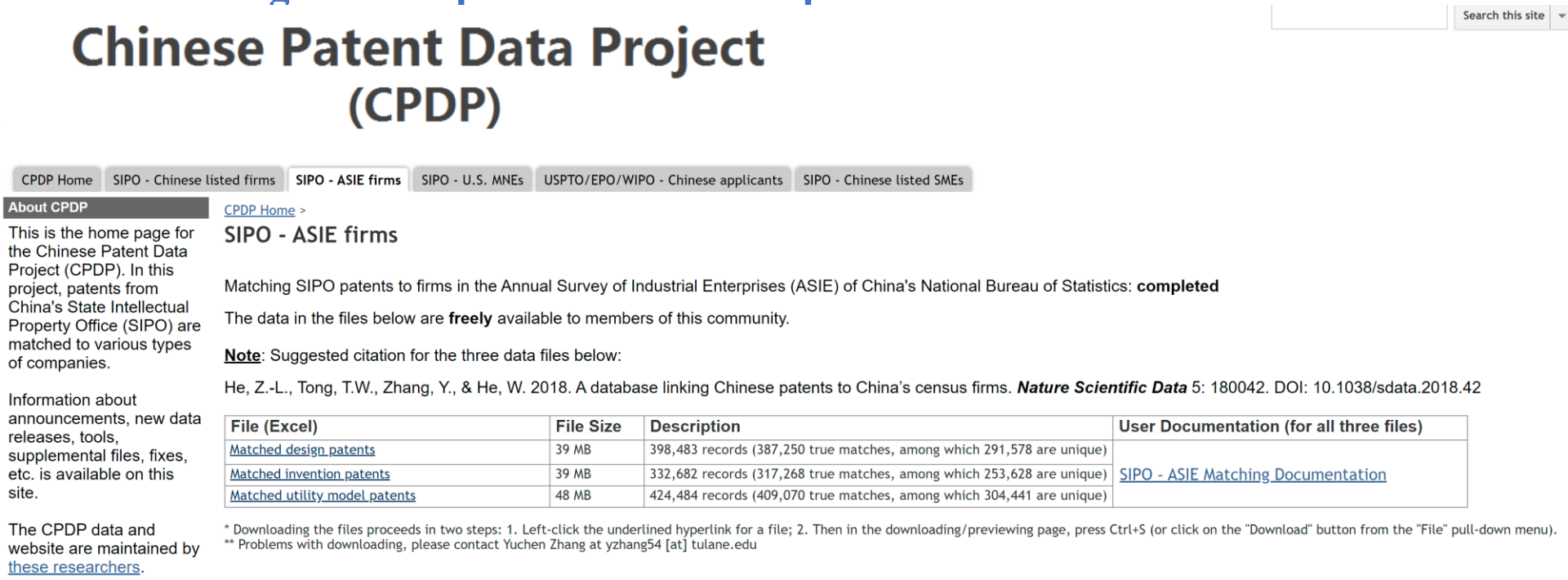
Contents

- **Topic 1: Construction of CNIPA database**
  - Inventor name disambiguation
  - **Applicant name harmonization**
  - Citation

- Topic 2: improving PATSTAT database, geocoding and mapping global network of innovation hotspots

Applicant Name Harmonization

- **Linking China's patent data to companies' data**



Chinese Patent Data Project
(CPDP)

Search this site

CPDP Home | SIPO - Chinese listed firms | SIPO - ASIE firms | SIPO - U.S. MNEs | USPTO/EPO/WIPO - Chinese applicants | SIPO - Chinese listed SMEs

About CPDP

This is the home page for the Chinese Patent Data Project (CPDP). In this project, patents from China's State Intellectual Property Office (SIPO) are matched to various types of companies.

Information about announcements, new data releases, tools, supplemental files, fixes, etc. is available on this site.

The CPDP data and website are maintained by these researchers.

CPDP Home >

SIPO - ASIE firms

Matching SIPO patents to firms in the Annual Survey of Industrial Enterprises (ASIE) of China's National Bureau of Statistics: **completed**

The data in the files below are **freely** available to members of this community.

Note: Suggested citation for the three data files below:

He, Z.-L., Tong, T.W., Zhang, Y., & He, W. 2018. A database linking Chinese patents to China's census firms. ***Nature Scientific Data*** 5: 180042. DOI: 10.1038/sdata.2018.42

File (Excel)	File Size	Description	User Documentation (for all three files)
Matched design patents	39 MB	398,483 records (387,250 true matches, among which 291,578 are unique)	
Matched invention patents	39 MB	332,682 records (317,268 true matches, among which 253,628 are unique)	SIPO - ASIE Matching Documentation
Matched utility model patents	48 MB	424,484 records (409,070 true matches, among which 304,441 are unique)	

* Downloading the files proceeds in two steps: 1. Left-click the underlined hyperlink for a file; 2. Then in the downloading/previewing page, press Ctrl+S (or click on the "Download" button from the "File" pull-down menu).
** Problems with downloading, please contact Yuchen Zhang at yzhang54 [at] tulane.edu

Notice: Should distinguish different type of applicants when pre-processing and stemming the names:

- Removing spaces, brackets, and name suffix such as "股份有限公司", "有限公司", "公司", "所", "院", "中心", as well as firms' prefix like "省", "市", "北京", "(北京)", "深圳".
- Keep the prefix representing addresses of **universities and research**, such as "北京大学", "浙江大学", etc., the address prefix is their sole identifier.
  - Heilongjiang University➔ Heilongjiang Group
  - Zhongshan Group ➔ Zhongshan University

Applicant Name Harmonization

- **Problem setting:** the synonym problem

  Goal: cover the name variant, changing names, typos, abbreviations
  subsidiaries, associate ventures, and joint ventures?

- **Rule-based (Dictionary based, Corporate trees) vs Machine learning approach**

  Assumption: high similarity in two or more dimensions indicates a higher probability of matching.

- **Our approach**

  e.g., 深圳/TCL/数字技术/有限公司
  Province/city + name stem+ industry+ type

  ① Preprocessing
  ② Stemming
  ③ Blocking by Name stems,
  ④ String similarity of full names, addresses (geocoded), inventors, technology profiles, etc.

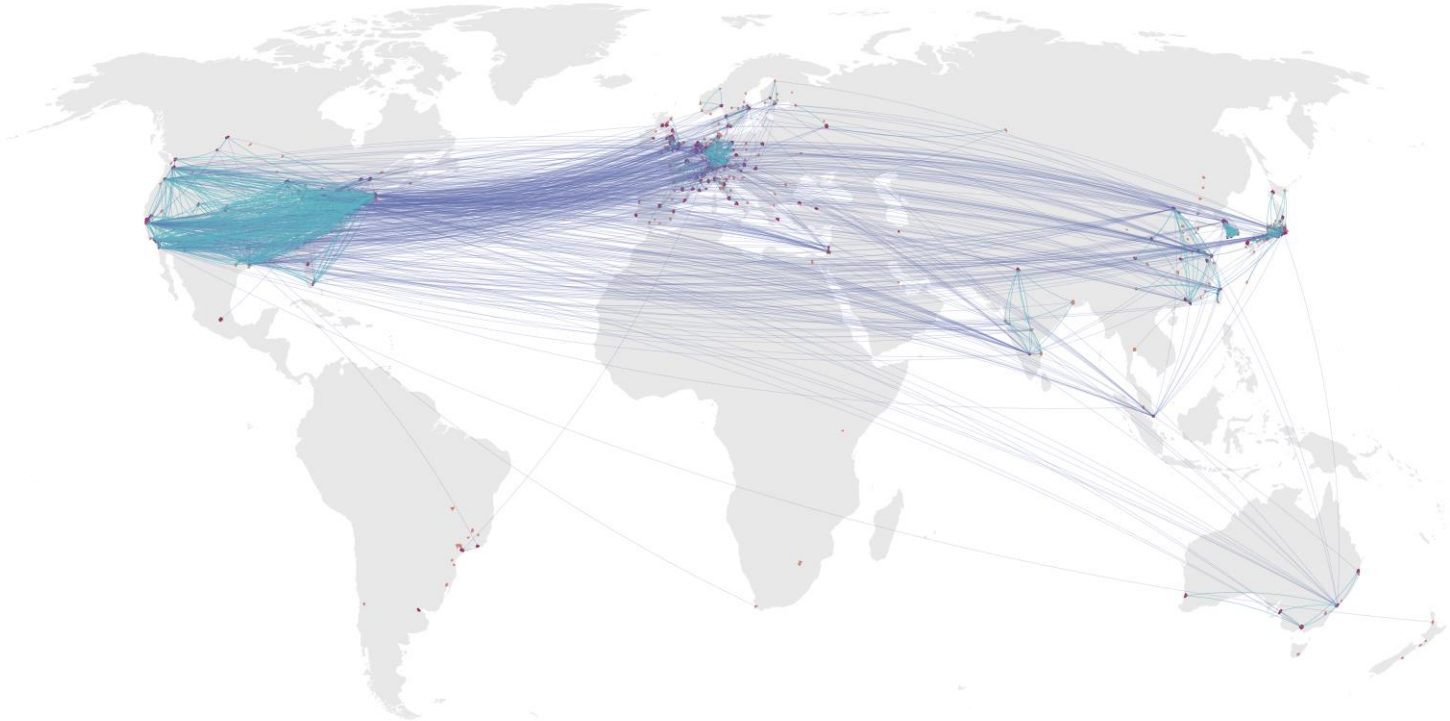
Applicant Name Harmonization

- **Steps:**
  - **Preprocessing:**
  - **Stemming:** extract the most special words (e.g., “万达”, “中兴通讯”, “TCL”, “ABB”) in a name with TF-IDF algorithm.
  - **Blocking** by the stem, applicant type and generating comparing pairs: from 0.38 million unique applicant names, we generated 5.54 m record pairs to compare.
  - **Comparing the Levenshtein similarity** of preprocessed names and record pairs.
  - **Clustering** with thresholds directly, or
  - **Predicting distances** based on trained models and clustering based on distance matrices. Two datasets were collected for training and testing our algorithm: one is our manually standardized applicants’ names of firms filed patents and listed on the National Equities Exchanges and Quotations (NEEQ), the other is SIPO’s linked data with China’s Annual Survey of Industrial Enterprises (ASIE), which is provided by the Chinese Patent Data Project (He et al., 2018).

Citation

Chinese Citation Database

- CNIPA Official Website (search report + examiner-added+ some extracted from full-text)
- Google Patent
- PATSTAT: (international search report)
- IncoPAT: (international search report + extracted from full-text)



Contents

Topic 2: improving PATSTAT database, geocoding and mapping global network of innovation hotspots

Invitation

The World Intellectual Property Organization (WIPO) invites you to the launch of the

World Intellectual Property Report 2019

The Geography of Innovation: Local Hotspots, Global Networks

followed by a panel discussion on
Where will future technology break through?

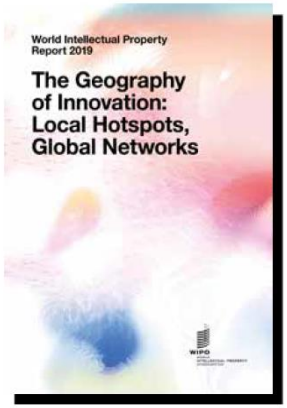
Tuesday, November 12, 2019
3.00 p.m.

WIPO New Conference Hall
34, Chemin des Colombettes, Geneva

To register for the event, visit **wipo.int/wipr**

#WIPR19

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Where exactly is innovation taking place? Relying on millions of patent and scientific publication records, the *World Intellectual Property Report 2019* documents how the geography of innovation has evolved over the past few decades. It finds that innovation is increasingly global and intertwined.

At the same time, a limited set of innovation hotspots lead the way and are at the center of global innovation networks. In addition to macro analysis of global trends, the report explores the geography of innovation through two case studies of technology fields undergoing rapid change – autonomous vehicles and agricultural biotechnology. On the basis of its findings, the report makes the case for economies to stay open in the pursuit of innovation.

Published every two years, the *World Intellectual Property Report* is WIPO's flagship analytical report.

Program

3.00 p.m. to 3.10 p.m.
Opening remarks

Mr. Francis Gurry, Director General, WIPO

3.10 p.m. to 3.40 p.m.
Presentation of the *World Intellectual Property Report 2019*

Mr. Carsten Fink, Chief Economist, WIPO

3.40 p.m. to to 4.40 p.m.
Panel discussion – Where will future technology break through?

Ms. Silke Reinhold, Head of Electronics and Mobility Patents and Design Rights, Volkswagen AG, Germany

Prof. AnnaLee Saxenian, Dean, School of Information, University of California at Berkeley, United States of America

Prof. Jie Tang, Harbin Institute of Technology and Former Vice-Mayor, Shenzhen, China

4.40 p.m. to 5.00 p.m.
Open discussion

5.00 p.m.
Reception

Tied in: the global network of local innovation

Mapping the global innovation networks of hotspots

Ernest MIGUELEZ[2], Julio RAFFO[1],
Massimiliano CODA-ZABETTA[2], Christian CHACUA[2], Deyun YIN[1]
Francesco LISSONI[2,3], Gianluca TARASCONI[3]
yindeyunut@gmail.com
1. IES, ESD, World Intellectual Property Organization
2. GREThA UMR CNRS 5113—Université de Bordeaux
3. ICRIOS - Bocconi University

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

2019/12/10

Introduction

Abstracts:
To enable the study of global geography of innovation, this paper constructed a comprehensive globally geocoded dataset of patent applications in PATSTAT (1970-2017) and scientific publications in Web of Science database (1998-2017). Based on Density-based spatial clustering of applications with noise (DBSCAN), it identified 174 GIHs and 313 SNCs worldwide, which together concentrate 85% of all patents and 81% of all scientific publications produced worldwide. This manuscript also explored several questions concerned with two current phenomena on the way knowledge is produced and shared worldwide: its geographical spread at the international level and its spatial concentration in few worldwide geographical hotspots.

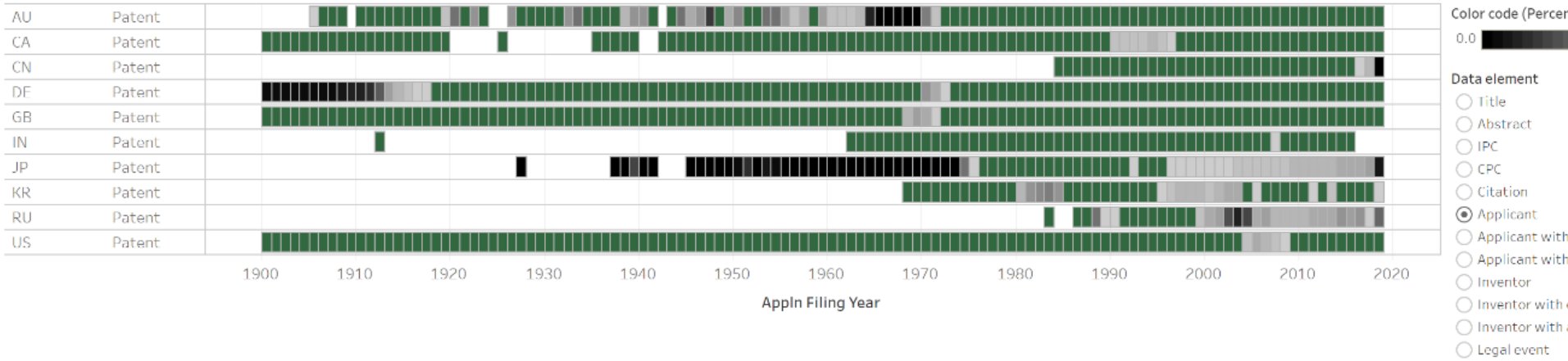
Keywords:
Geocoding, Patent, DBSCAN, Innovation, Global Innovation Network

Data & Methodologies

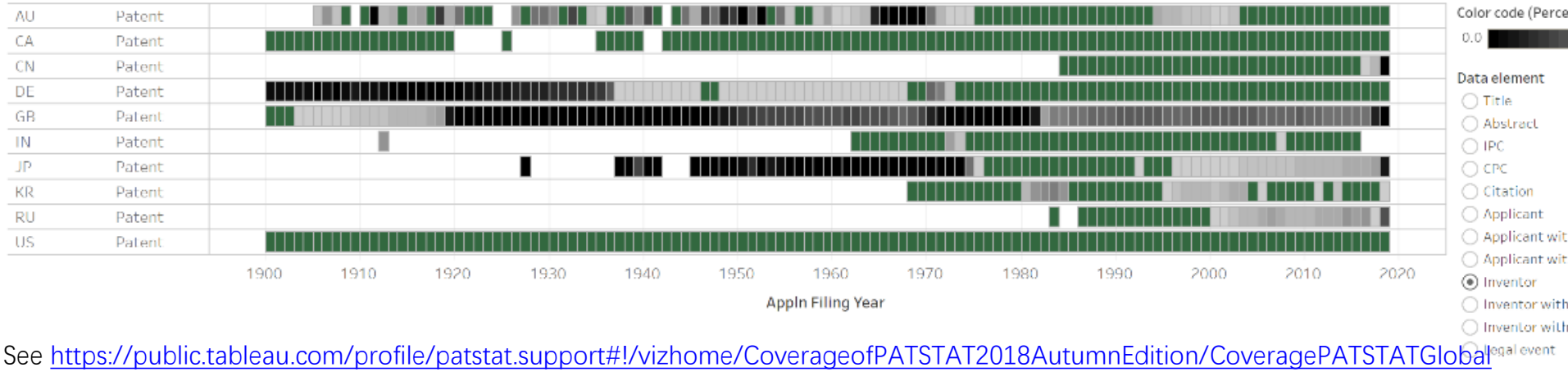
Data

Coverage of PATSTAT 2018 Autumn

Applicant



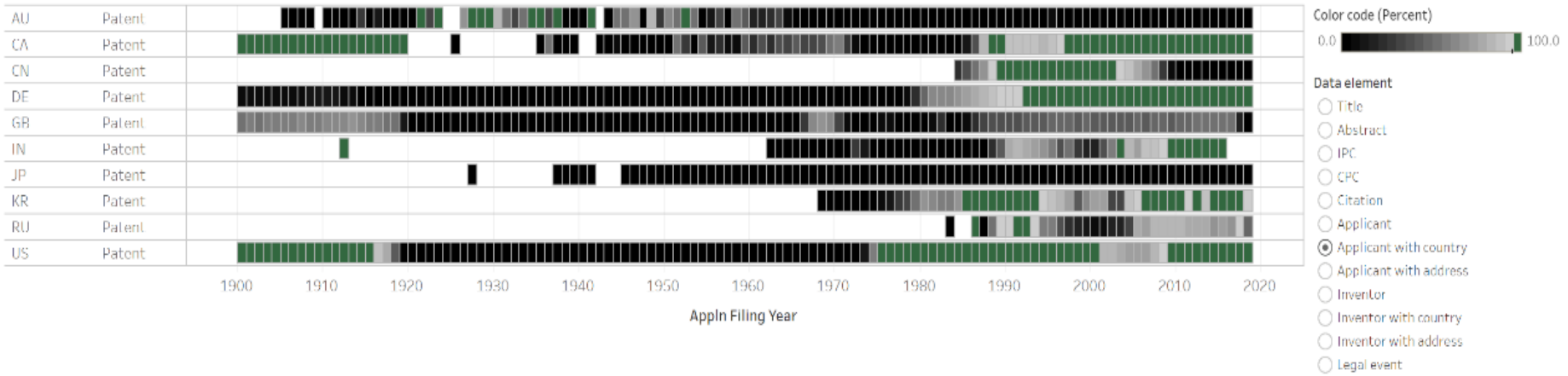
Inventors



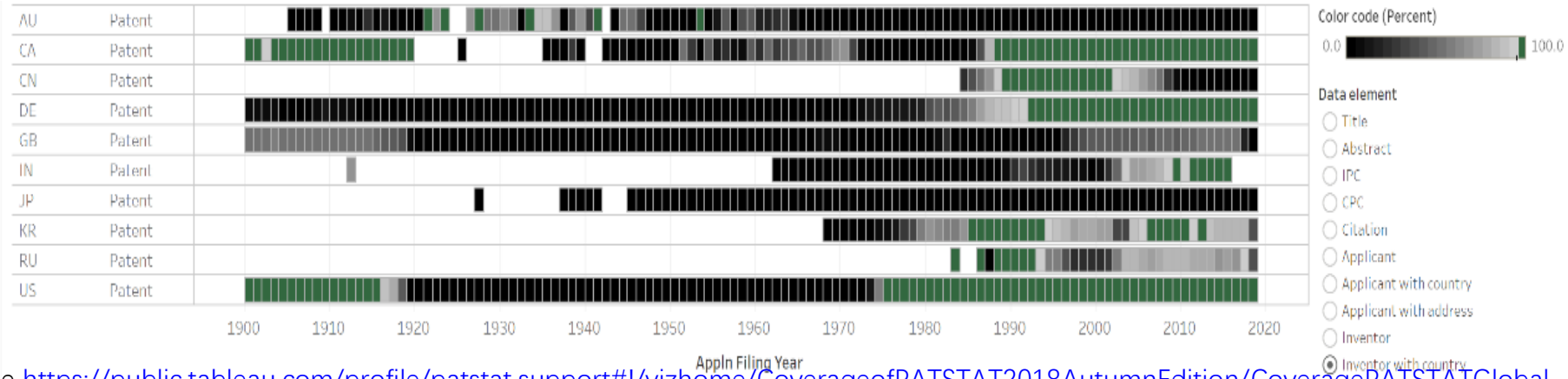
See https://public.tableau.com/profile/patstat.support#!/vizhome/CoverageofPATSTAT2018AutumnEdition/CoveragePATSTATGlobal

Data

Coverage of PATSTAT 2018 Autumn
Applicant with country



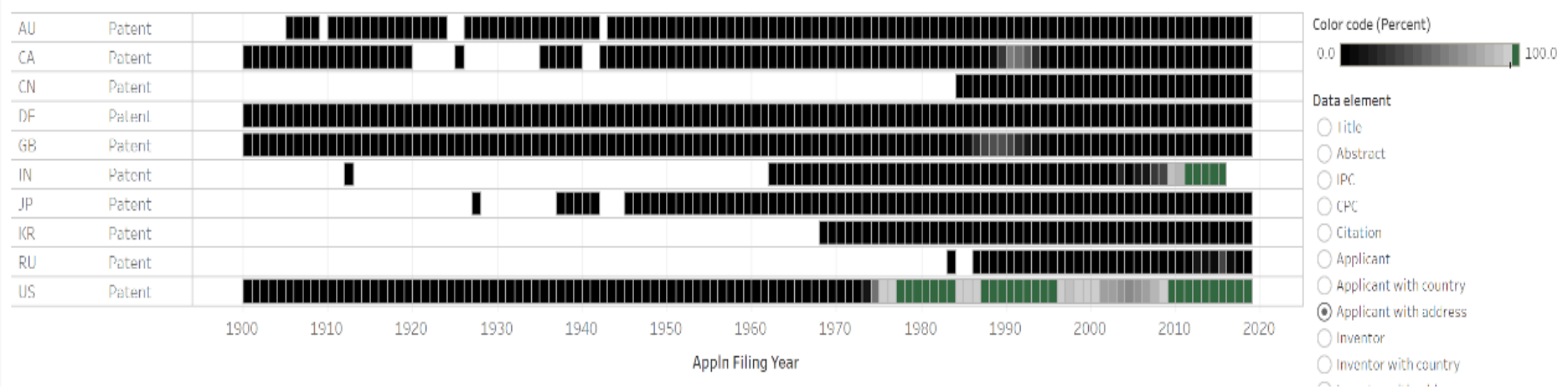
Inventors with country



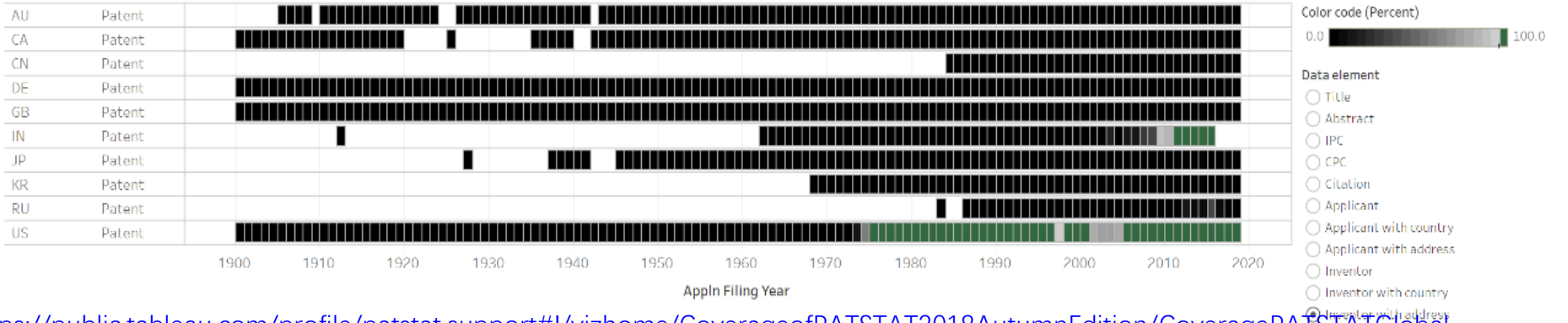
See https://public.tableau.com/profile/patstat.support#!/vizhome/CoverageofPATSTAT2018AutumnEdition/CoveragePATSTATGlobal

Data

Coverage of PATSTAT 2018 Autumn
Applicant with address



Inventors with address



https://public.tableau.com/profile/patstat.support#!/vizhome/CoverageofPATSTAT2018AutumnEdition/CoveragePATSTATGlobal

Data

Patent data

Sources

- **PATSTAT: Rassenfosse et al (2019)** +
- USPTO: PatentsView +
- PCT: WIPO Statistical Database +
- SIPO: Yin & Motohashi (2018) +
- JPO: Ikeuchi et al (2017) +
- EPO, JPO, USPTO: Morrison et al. (2017)
- **Geocoded at up to rooftop level**

Coverage

- 48 years (1970-2017)
- 168 patent offices
- 24.8 M (of 34M) patent families (73%)
  - 7.8 M (of 9M) international patent families (87%)
  - 17.5 M (of 25.5M) domestic patents (79%)
- 22 M inventors

Scientific Publication data

Source

- Clarivate's Web of Science, SCIE
- **Geocoded at postal code or sub-city level**

Coverage

- 20 years (1998-2017)
- 25.9 million records (93%)
- 20 million papers (98%)
- 66 M authors

Methodologies

Step 1: Density-based spatial clustering of applications with noise (DBSCAN)

To make the result internationally comparable

✓ **International patent families + All Wos**

① Patent family >=2

② PCT patents

③ Foreign-oriented patents: applicants' origin of country ≠ filing offices

✓ **Domestic patents (National patents, Singletons)**

Methodologies

GINs Vs. SNCs

- **DBSCAN on all records ➔ Global Innovation Hotspots (GIHs)**

  or, more plainly, hotspots.

- **DBSCAN on specialized industries: ➔Specialized Niche Clusters (SNCs):**

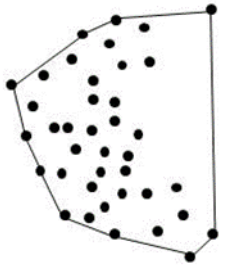
  To allow for scientific and technological specialization, the above method is repeated for 25 sub-samples of the same publication and patent data, which refer to 12 scientific fields and 13 technological ones, respectively. Only the resulting polygons of these 25 iterations not contained within a hotspot are kept. From these, the overlapping polygons are merged in the same way as for hotspots.

  The final outer areas are referred to as (SNCs) or, more plainly, niche clusters.

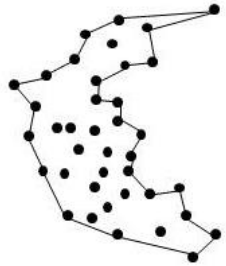
Methodologies

- **Step 2: Delimiting clusters' borders with concave hulls**

- ☹ **The convex hull** in a set X of points in the Euclidean space is the smallest convex set that contains all the X points. That is to say, is a polygon encompassing all the X points, with straight, short lines connecting the outer points in space among X. Any point inside the outer shape belongs to the polygon.



- ☺ **Concave hull:** A k-nearest neighbors approach for the computation of the region occupied by a set of points. The concave hull starts from the convex hull, but removes the concave areas that do not have any point inside. (Adriano et al., 2007)



- **Step 3: Final boundaries of clusters: merging patent layer with publication layer)**

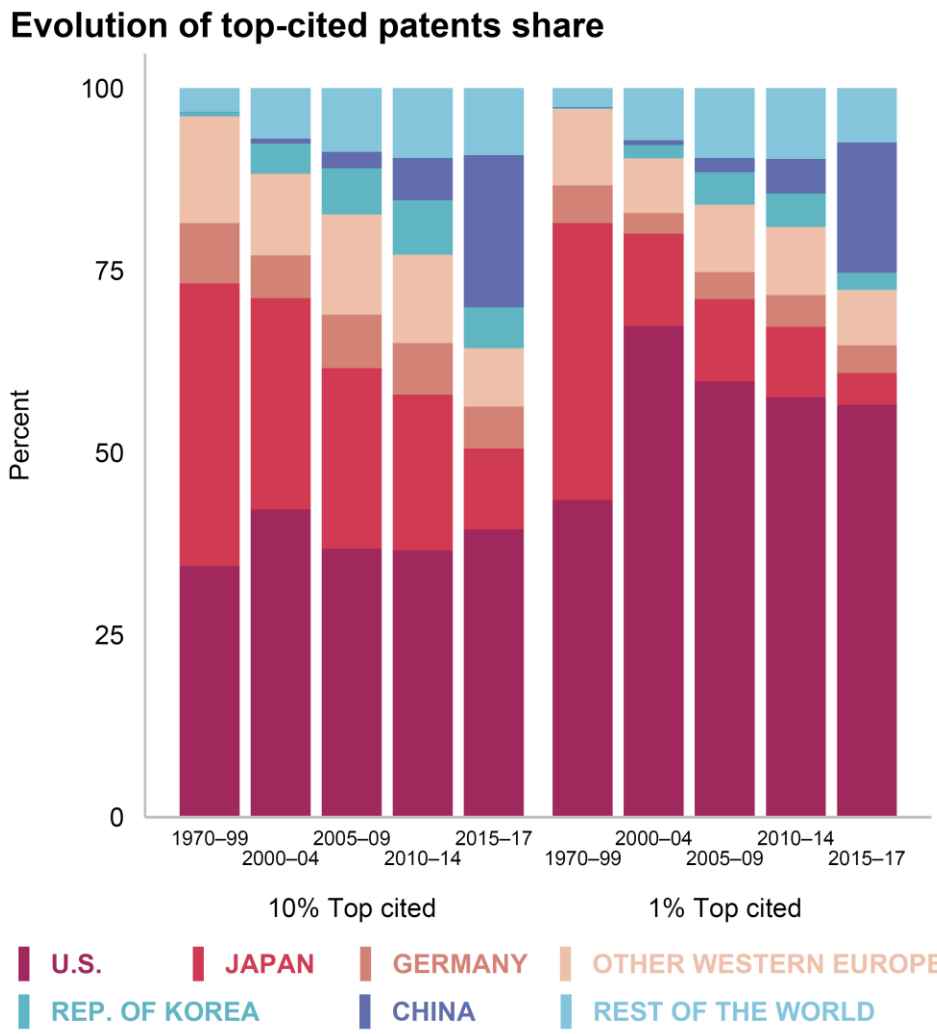
Methodologies

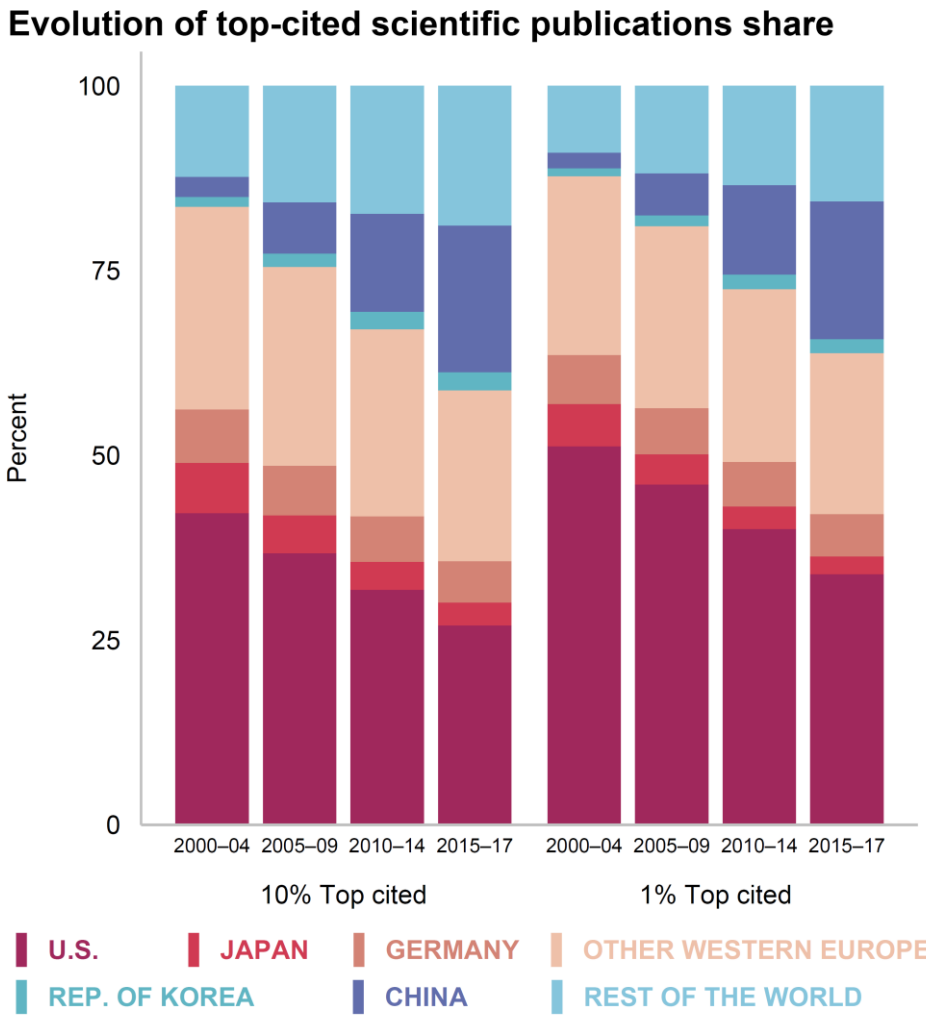
- **are internationally comparable**, i.e. the same scientific publication or patent (specialized) density would have determined the same hotspot (cluster) anywhere in the world;
- **have non-predefined boundaries**, i.e. hotspots and niche clusters can have different sizes and include more than one city, state/province or country.
- **can have different scientific and technological density**, i.e. hotspots and niche clusters need only scientific publication or patent high concentration, but not necessary both;
- **have different specialization density**, i.e. niche clusters are defined with lower density thresholds than hotspots;
- **are distinct geographical areas**, i.e. the polygons are non-overlapping within and across hotspots and niche clusters; and,

Descriptive Statistics

2.1 The two sides of global knowledge production

Figure 2.2: More the value, more the concentration





2.1 The two sides of global knowledge production

Table 2.2 Shares of top innovation subnational regions within countries

Top three large administrative areas in patent and scientific publication concentration by period, selected countries

Country (level)	Patents					Publications				
	1991-95	%	2011-15	%		2001-05	%	2011-15	%	
China (provinces)	Beijing Guangdong Shanghai	42.3	Guangdong Beijing Jiangsu	60.3	↑	Beijing Shanghai Jiangsu	45.5	Beijing Shanghai Jiangsu	39.4	↓
Germany (states)	Baden-Württemberg Bayern Nordrhein-Westfalen	63.8	Bayern Baden-Württemberg Nordrhein-Westfalen	65.0	↑	Bayern Nordrhein-Westfalen Baden-Württemberg	49.4	Nordrhein-Westfalen Baden-Württemberg Bayern	50.0	→
France (regions)	Île-de-France Auvergne-Rhône-Alpes Grand Est	64.1	Île-de-France Auvergne-Rhône-Alpes Occitanie	59.9	↓	Île-de-France Auvergne-Rhône-Alpes Occitanie	63.1	Île-de-France Auvergne-Rhône-Alpes Occitanie	62.7	↑
United Kingdom (counties)	Greater London Hertfordshire Cambridgeshire	17.9	Greater London Cambridgeshire Oxfordshire	23.9	↑	Greater London Cambridgeshire Oxfordshire	35.8	Greater London Oxfordshire Cambridgeshire	38.7	↑
India (states)	Maharashtra Karnataka Telangana	51.6	Karnataka Maharashtra Telangana	60.1	↑	Maharashtra Tamil Nadu NCT of Delhi	36.4	Tamil Nadu Maharashtra NCT of Delhi	36.1	→
Japan (prefecture)	Tokyo Kanagawa Osaka	51.5	Tokyo Kanagawa Osaka	56.3	↑	Tokyo Osaka Ibaraki	35.8	Tokyo Osaka Aichi	35.4	→
United States (states)	California New York Texas	30.8	California New York Texas	36.5	↑	California New York Massachusetts	28.2	California Massachusetts New York	28.7	→

2.1 The two sides of global knowledge production

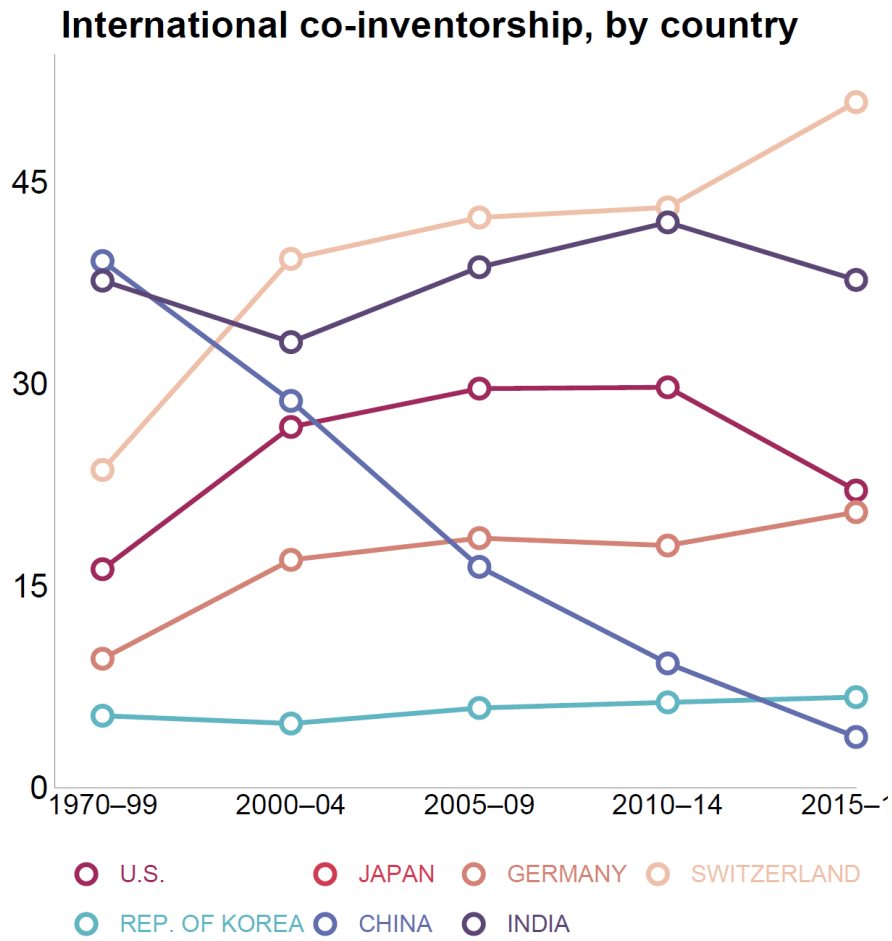
Table 2.4 Persistent concentration of innovation in a few hotspots

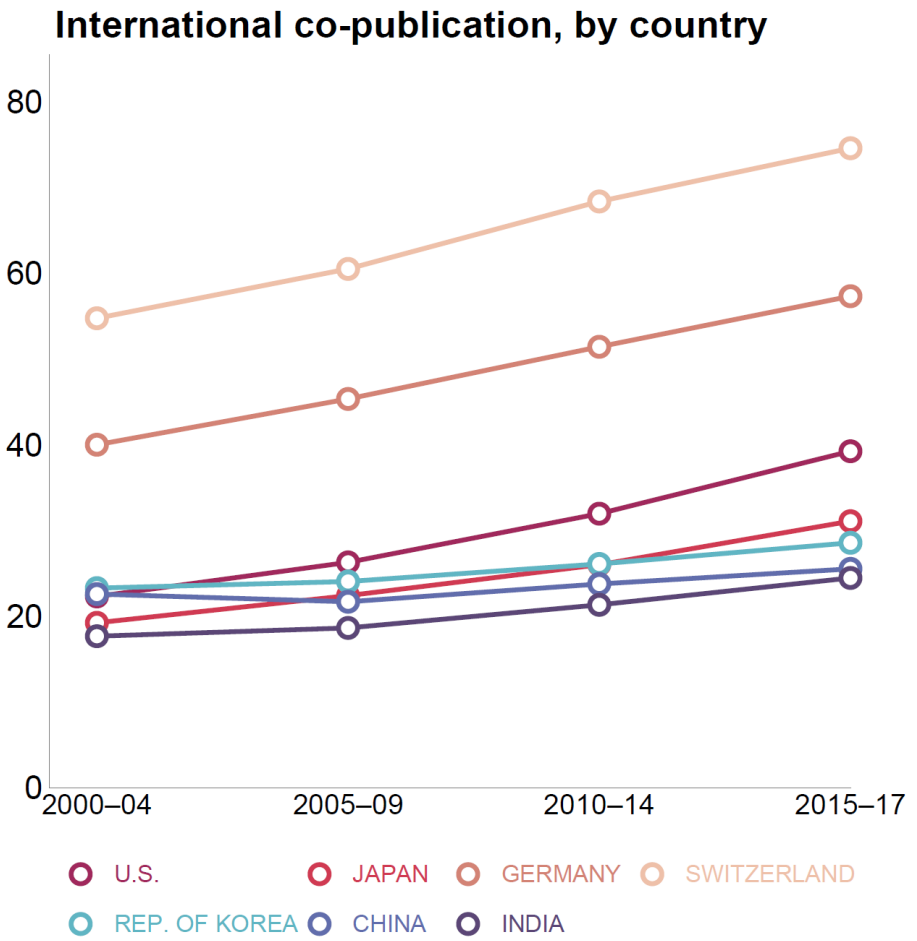
Top three GIH concentration, patents

Country	Patents				Publications				
	1991-95	%	2011-15	%	2001-05	%	2011-15	%	
China	Beijing Shanghai Shenzhen-Hong Kong	**36.5**	Shenzhen-Hong Kong Beijing Shanghai	**52.2**	Beijing Shanghai Nanjing	**43.9**	Beijing Shanghai Nanjing	**35.8**	↓
Germany	Frankfurt Köln-Dusseldorf Stuttgart	**37.4**	Frankfurt Stuttgart Köln-Dusseldorf	**29.4**	Frankfurt Köln Berlin	**34.4**	Frankfurt Köln Berlin	**34.2**	→
France	Paris Lyon Grenoble	**47.1**	Paris Grenoble Lyon	**42.8**	Paris Lyon Grenoble	**51.0**	Paris Lyon Toulouse	**49.4**	↓
United Kingdom	London Manchester Cambridge	**30.0**	London Cambridge Oxford	**35.0**	London Cambridge Oxford	**39.8**	London Oxford Cambridge	**41.8**	↑
India	Bengaluru Mumbai Delhi	**41.9**	Bengaluru Hyderabad Delhi	**46.2**	Delhi Mumbai Bengaluru	**27.7**	Delhi Mumbai Kolkata	**24.6**	↓
Japan	Tokyo Osaka Nagoya	**80.5**	Tokyo Osaka Nagoya	**83.4**	Tokyo Osaka Nagoya	**64.3**	Tokyo Osaka Nagoya	**64.8**	→
United States	New York City San Jose-San Francisco Boston	**19.4**	San Jose-San Francisco New York City Boston	**23.4**	New York DC-Baltimore Boston	**21.2**	Boston New York DC-Baltimore	**21.4**	→

2.2 – Global networks of collaboration and sourcing

Figure 2.8
Large economies are highly internationalized





2.3 – Local innovation and global networks of innovative hubs
Global Innovation Network: Country-level

More international collaboration between countries

International co-invention

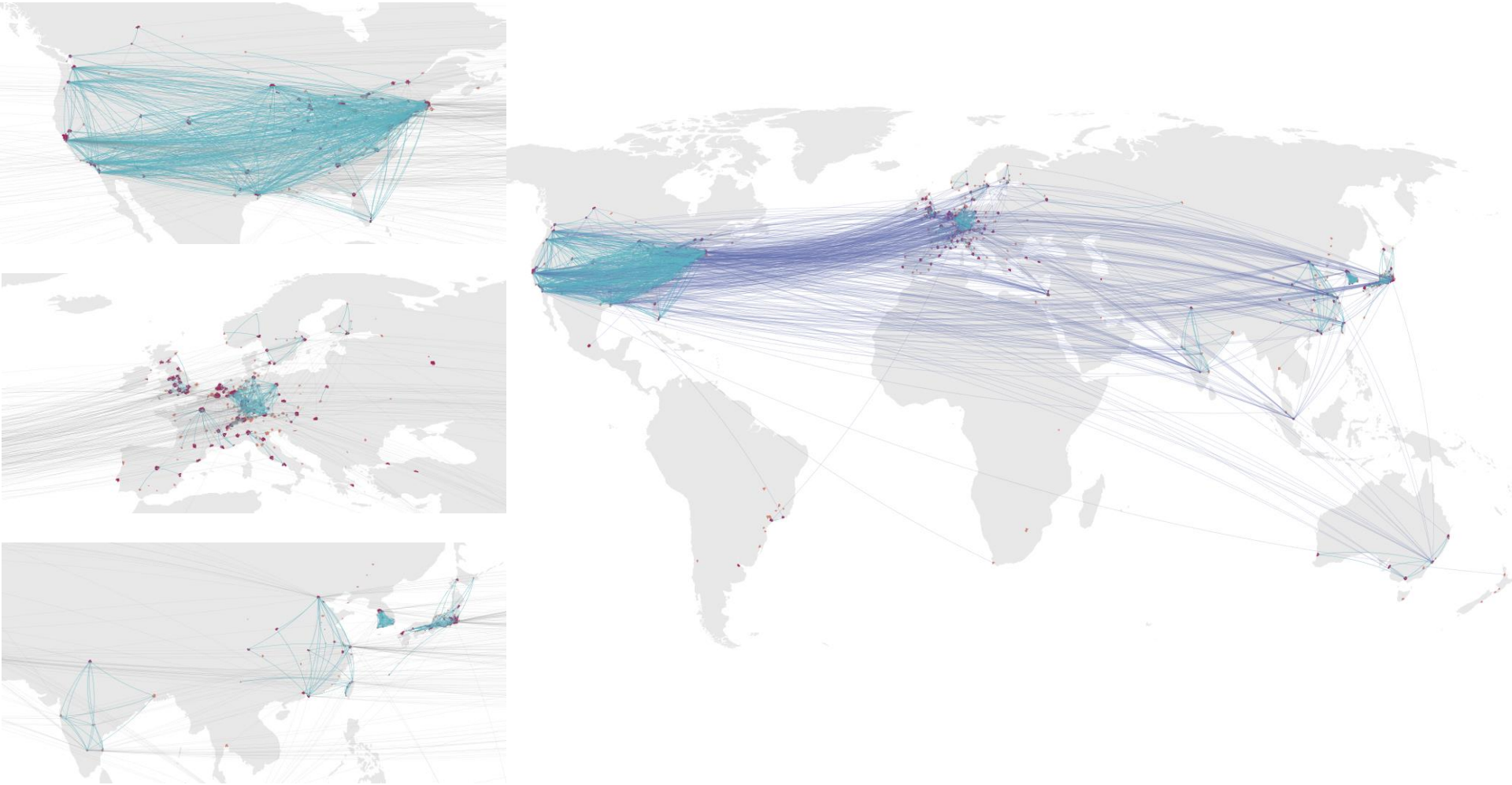
International co-publication

1998-2002

2011-2015

Who collaborates with whom?

Top 10 percent co-invention ties among GIHs and SNCs, 2011-2015

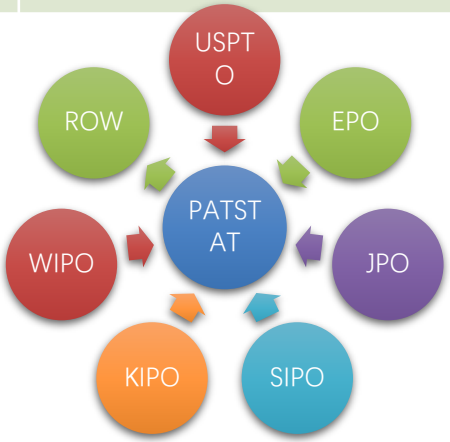


Concluding remarks
Global Patent database: an overview

	Who?	Who?		Where?
	Applicants Harmonization	**Inventor Disambiguation**	**Gender identification**	**Geocoding**
USPTO	NBER; PatentsView	NBER; PatentsView	PatentsView	NBER; PatentsView
PATSTAT	OECD HAN	Pezzoni et al., 2014	OECD, 2015	de Rassenfosse et al. (2019),
WIPO	Hao Zhou	Penner et al., 2019	Martinez et al., 2016	Bergquist et al., 2017
JPO	IIP database	Ikeuchi et al., 2018	Easy	Ikeuchi et al., 2018
CNIPA	He et al., 2017, 2018	Yin et al., 2018	Yin et al., 2019	Yin et al., 2018
A unified?		Under-processing	Under-processing	Under-processing

Some ideas about Innovation Information Initiatives:

- Fragmented, different rules ➔ A unified framework, rules and algorithms, and global patent database (PATSTAT)
- A daunting task! ➔ Need international collaboration



Thank you!

Q&A

deyun.yin@wipo.int