

USPTO PatentsView: a disambiguated resource for policy research

Identifiers and Intellectual Property Workshop
OECD – ORCID – CrossRef

Evgeny Klochikhin

June 22, 2017

Agenda

- Overview
- Data
- Disambiguation
- Analysis

Overview

USPTO PatentsView (www.patentview.org)

- PatentsView is a patent data visualization and analysis platform intended to increase the value, utility, and transparency of US patent data. The initiative is supported by the [Office of Chief Economist](#) in the US Patent & Trademark Office (USPTO)
- Over 40 years of U.S. patent data – parsed, cleaned, processed, disambiguated and visualized
- Multiple tools: data visualizations, API, relational databases, search index, Query Builder
- Advanced computational techniques for inventor, assignee and location disambiguation
- Record linkages and non-USPTO patents-related data sources (e.g. government-funded awards)

Platform

PV Database



API

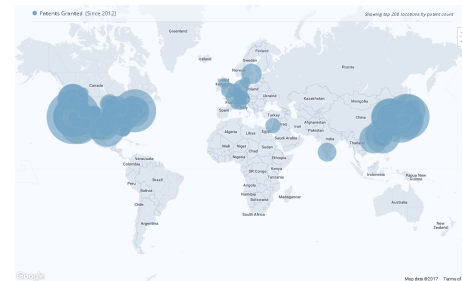
POST /api/patents/query

Bulk Downloads

- Inventors
- Assignees
- CPC
- NBER
- Etc.



Visualizations



Query Builder

Quick Search View Data Dictionary Help

Step 1: Select Category^{*}
Use the query builder to search for a subset of: Patents Inventors Assignees

Step 2: Select Search Criteria
Use the following criteria to identify a set of US patents. The output dataset will contain all primary entities (patents, inventors, or assignees) associated with those patents.

Search Summary Please select at least one search criterion to begin your search. Your selected criteria will show up here as you click to "add to search"

Patents Select Field Add to Search

Inventors Select Field Add to Search

Code Snippets

```
import requests, json
import codecs
import time
from unicode import unicode

# Defining function using Google API to fetch geo info based on inputs
def _whisk_group(**kwargs):
    kwargs['sensor'] = 'false'
    url = 'https://maps.googleapis.com/maps/api/geocode/json'
    r = requests.get(url, params=kwargs)
    return json.loads(r.text)

punctuation = "({[ ] : ; , ' \" % & ' / _ ~ * , split())
punctuation.append("")

fd = 'H:\share\Science Policy Portfolio\Patents\raw XML\Location Disambiguation Q4\''.replace("\\", '/')
mydb = sqlite3.connect(fdb+'geolocation_new.sqlite3')

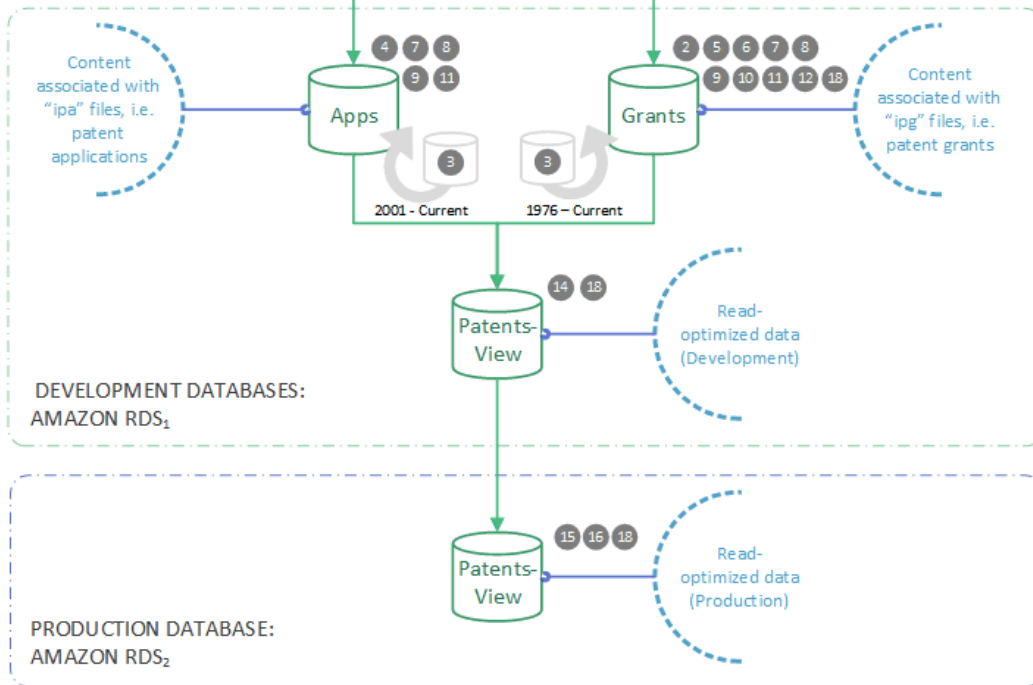
def re_fn(expr, item):
    reg = re.compile(expr, re.I)
    return reg.search(item) is not None
```

Data

PatentsView data process

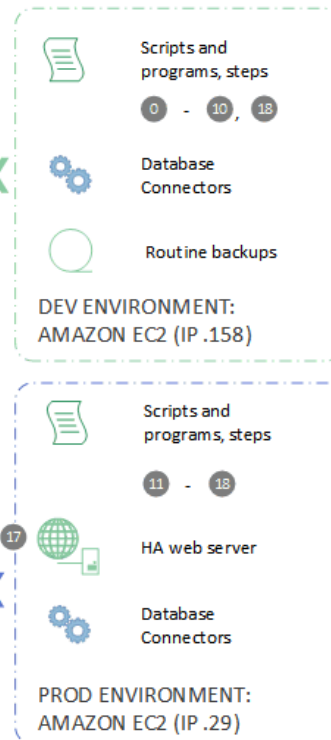
PATENTSVIEW DATA PROCESS

DATA PIPELINE



JULY 22, 2016 | v0.1

PROCESSING ENVIRONMENT



SCRIPT AND PROGRAM LEGEND

Preparing "raw" data...

- | | | |
|--|--|---|
| 0 Download new USPTO files | 4 Create application number crosswalk between apps and grants databases | 8 Run inventor disambiguation |
| 1 Process (new) files and store raw data in grant and apps databases | 5 Populate USPC and CPC tables | 9 Run location disambiguation |
| 2 Create alternative application numbers in grants database | 6 Run NBER algorithm | 10 NEW: Run government interest parsing |
| 3 Merge "new" data with prior versions of grants and apps | 7 Run assignee (apps and grants) and lawyer (grants only) disambiguation | 11 Create crosswalk tables in apps and grants raw databases |

Preparing read-optimized database...

- | | |
|---|---|
| 12 Website database generator (grants only) | 16 Generate caching database (PatentsView production) |
| 13 NOT RUN: Unicode HTML Entities | 17 Set config file to point web application to new database |
| 14 Add full text indices | 18 Remove old database on PatentsView development and production and archive raw patents and apps databases |
| 15 Promote from PatentsView development to production | |

Summary statistics

- 6,215,171 granted patents (1976-2017)
- 14,536,617 (raw) inventors
- 5,442,930 (raw) assignees
- 19,979,547 (raw) locations
- 86,184,397 U.S. patent citations
- 31,735,419 CPC classes
- 124,765 patents with government interest statement
- Etc.

Disambiguation

Disambiguated fields

- Inventors
- Assignees
- Locations
- Lawyers

USPTO Inventor Disambiguation Workshop, 2015

- Original disambiguation algorithm by UC Berkeley based on Jaro-Winkler distance (Li et al., 2014)
- 7 research teams from the United States, Europe, Australia, and China
- Evaluation strategy based on the training (“gold standard”) data shared by Manuel Trajtenberg (2009), Ivan Png (2016), and Pierre Azoulay (2012)
- UMass Amherst is the winner with a hierarchical coreferencing algorithm (F1 score=0.98)

* More information at www.patentsview.org/workshop

Results

Table 2. Disambiguated inventor cluster size in the PatentsView databases.

	Average inventor cluster size	Minimum cluster size	Maximum cluster size
Before workshop	3.4686	1	4,723
After workshop	4.5037	1	4,758

Table 3. Disambiguation of eight National Hall of Fame inventors.

	Last Name	First Name	# of patents
Before workshop	Arnold	Frances	1
	Arnold	Frances H.	42
	Arnold	Frances T.	1
	Bowerman	William J.	7
	Boykin	Otis F.	3
	DiMarchi	Richard D.	65
	DiMarchi	Richard D.	1
	Dresselhaus	Mildred	4
	Dresselhaus	Mildred S.	10
	Gadgil	Ashok	6
	Gadgil	Ashok J.	8
	Hull	Charles W.	76
	Hull	Charles W.	3
	Whitfield	Willis J.	1
After workshop	Arnold	Frances H.	48
	Bowerman	William J.	7
	Boykin	Otis F.	3
	DiMarchi	Richard D.	85
	Dresselhaus	Mildred S.	15
	Gadgil	Ashok J.	14
	Hull	Charles W.	78
	Hull	Charles W.	1
	Whitfield	Willis J.	1

14,536,617 raw inventors

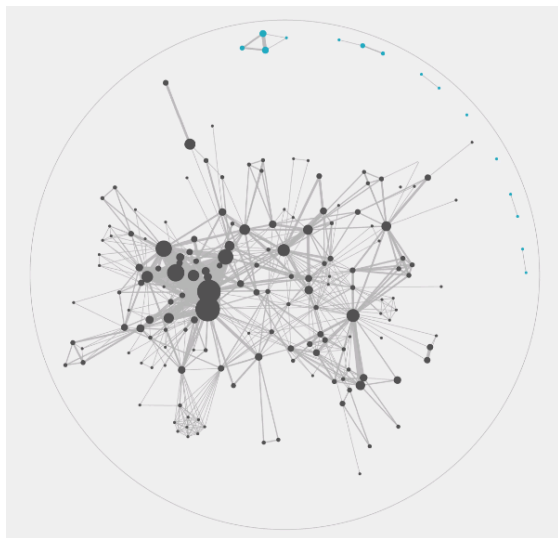


3,414,473 disambiguated inventors

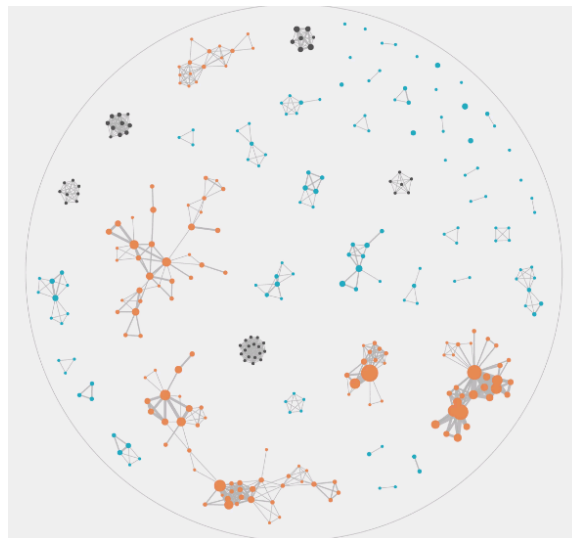
Analysis

Companies innovate differently (Scientific American, 2015)

Tesla



Intrexon

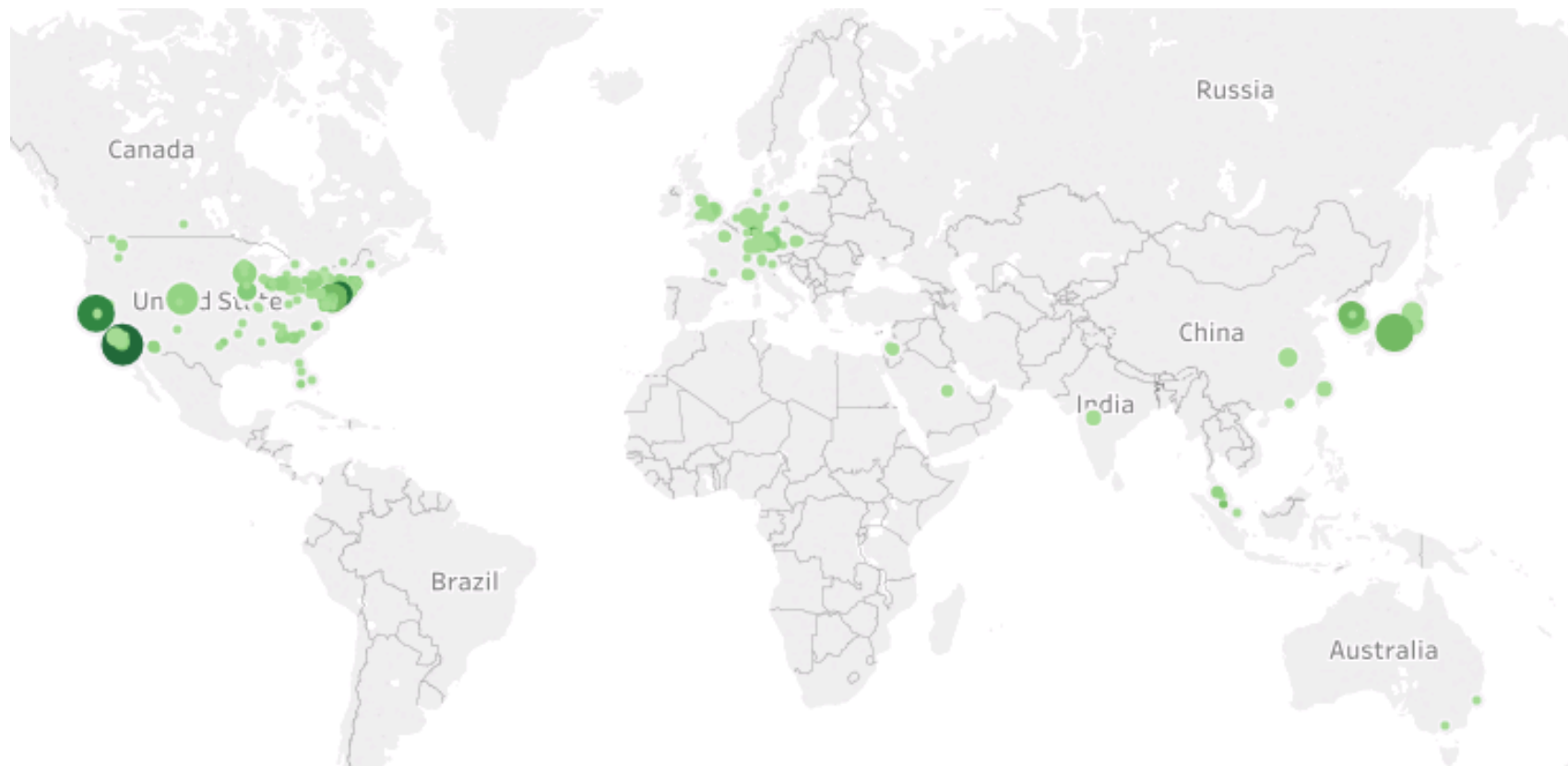


Facebook

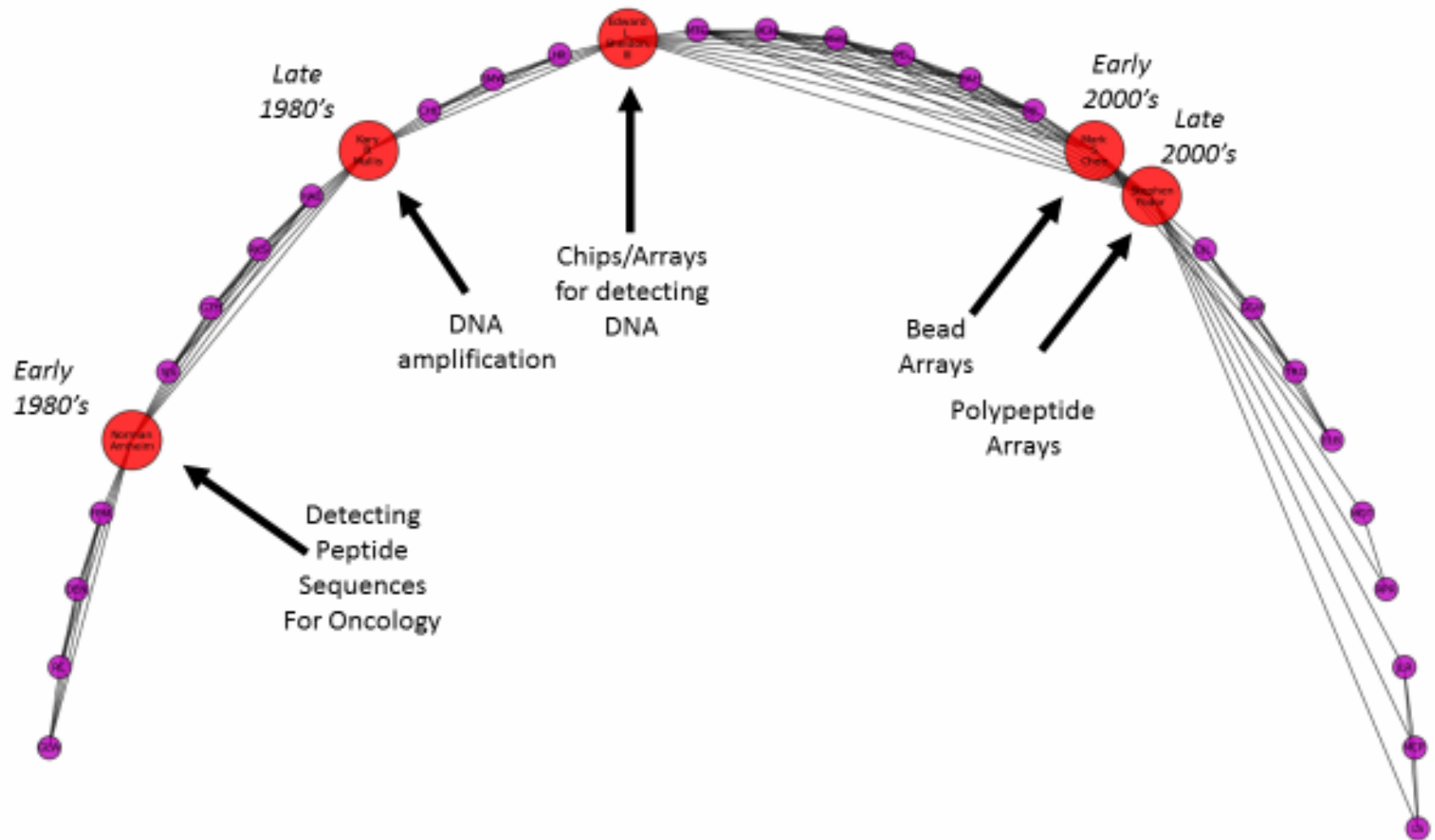


Credit: PERISCOPIC

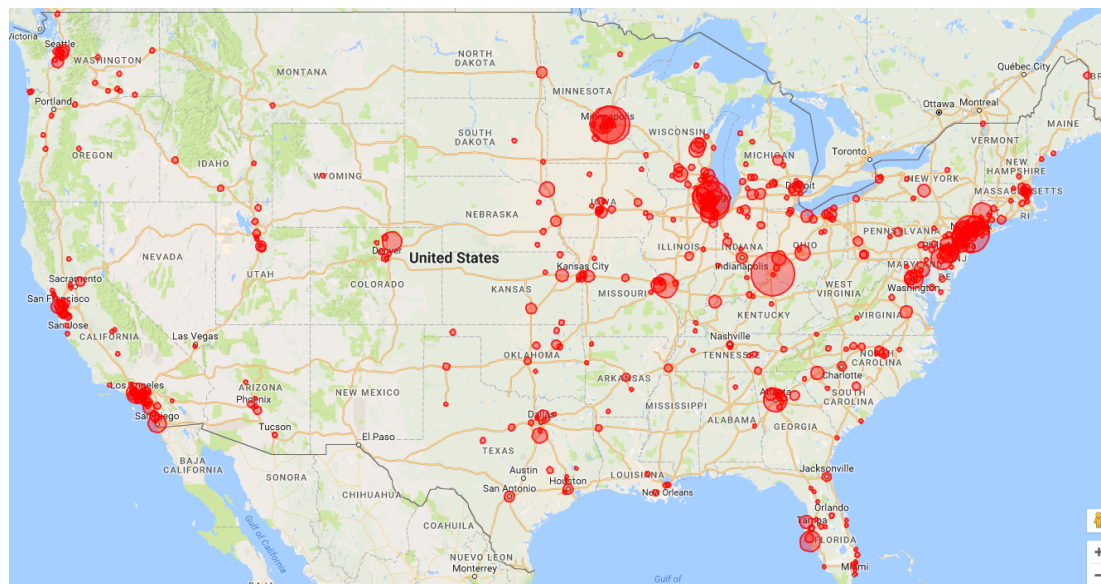
Cancer Moonshot data



Co-inventor networks for patents with high citation network pagerank



Food safety patents



Government-funded patents

Agency	No. of patents
Army	3
Department of Agriculture	14
Department of Energy	3
National Institutes of Health	9
National Science Foundation	4
Centers for Disease Control and Prevention	1
U.S. Government (as a whole)	5
Total	39

Questions?

Evgeny Klochikhin
eklochikhin@air.org