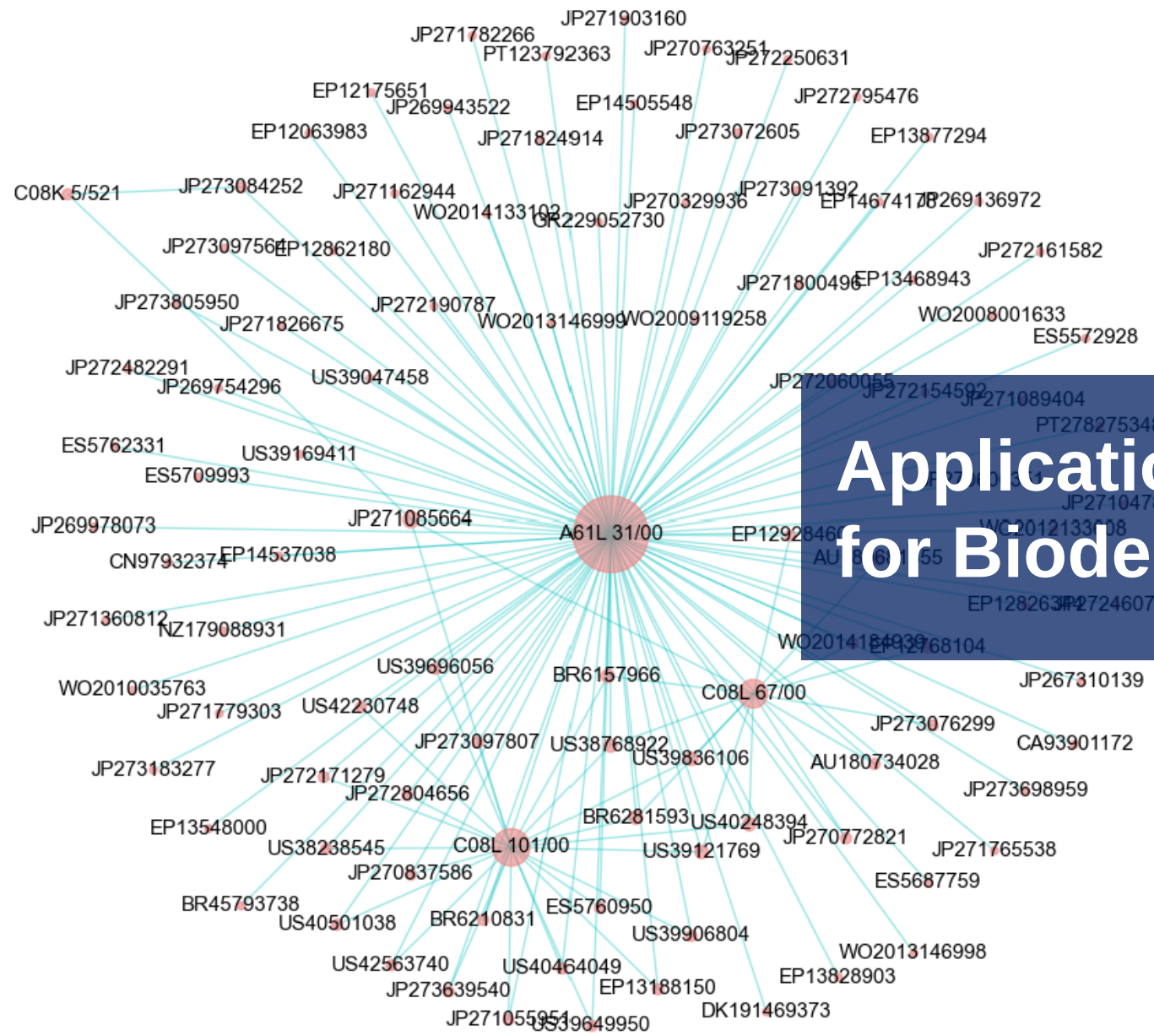


Application Recommendation System for Biodegradable Polymer

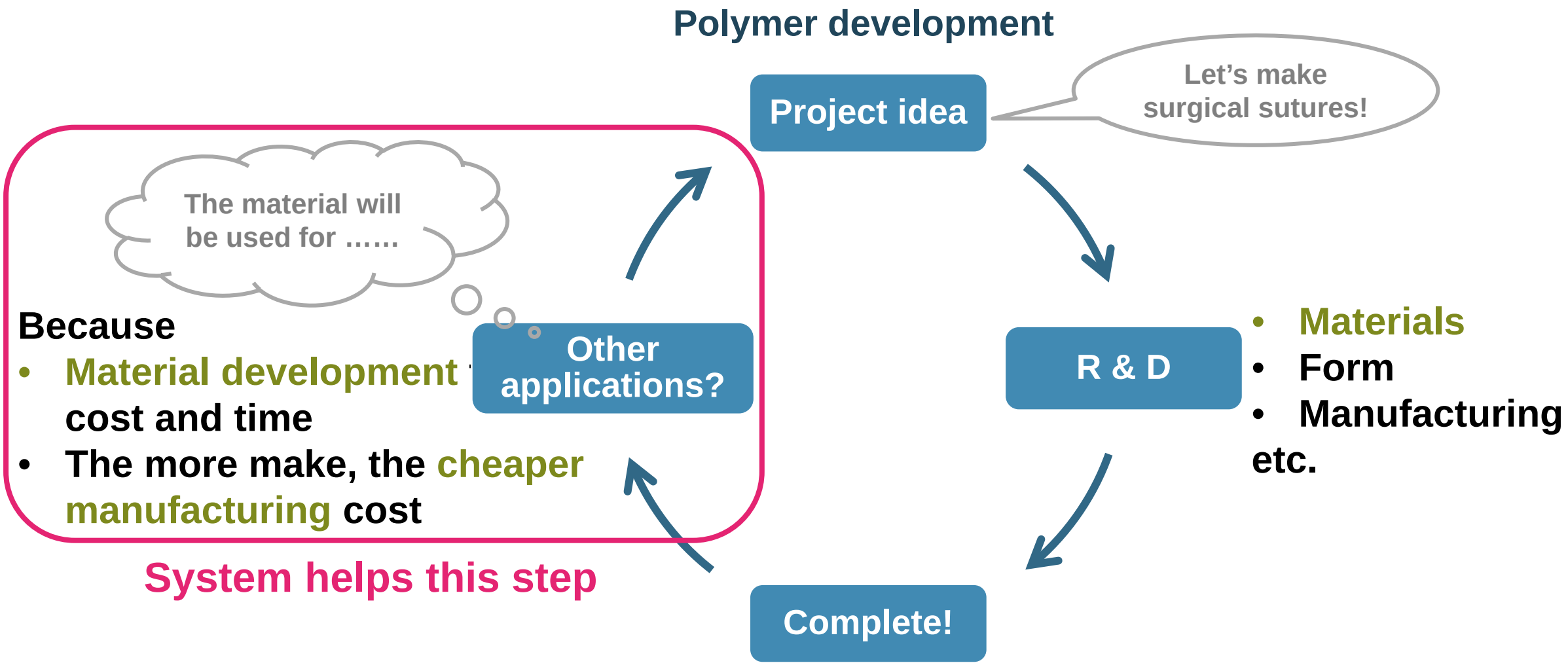
Namiko Nakashima

- This system is what I wanted when I was a polymer scientist -

Table of Contents

1. Problem Statement

1.1. Problem



Goal

Provide an application recommendation system:

- Giving some hints about potential applications
- Interactive (accept user input and return the result)
- About biodegradable polymers

What are they?

Polymers degrading:

- in natural environment
- in our body

Why are they chosen?

They are needed:

- raising environmental awareness
- advancing in medical technology

1.2. Client

1. Chemical companies

- Find other applications of their polymer
- Search competitors and their technologies

2. Research institutions

- Find other applications of their polymer
- Propose their polymer to commercial companies

  (Which company will be interested in their polymer?)

2. Exploratory Data Analysis
2.1. Data

Data Source:

From World Intellectual Property Organization (WIPO)

with a search word "IC:(C08L 101/16)"

IPC code

- gathers patent information from 193 member countries
- coverage is not exhaustive, but wide

Data:

- **8,182 patents**
- **from 1970 to Jun.2.2020**
- **registered in 37 countries and organizations**
- **19 languages**

will be used to build the system!

	Patent_Id	Application_Number	Application_Date	Country	Title	Abstract	IPC	Applicants	Inventors
0	AR192047768	P150101734	01.06.2020	AR	COMPOSICIÓN POLIMÉRICA RELLENA CON UNA MEZCLA DE MATERIAL DE CARGA INORGÁNICO	La presente se refiere a una composición polimérica que comprende por lo menos 20,0% en peso, en...	C08L 67/02; C08L 67/04; C08L 101/16	OMYA INTERNATIONAL AG	NaN

What is IPC code?

International Patent Classification Code

- **All technical fields are divided into 8 "sections" from A to H:**
  - A: Human Necessities
  - B: Performing Operations, Transporting
  - C: Chemistry, Metallurgy
  - D: Textiles, Paper
  - E: Fixed Constructions
  - F: Mechanical Engineering, Lighting, Heating, Weapons, Blasting
  - G: Physics
  - H: Electricity
- **Each IPC code: material, application, or technology to manufacture it**
- **Hierarchical structure**

e.g.) **C08L 101/16**

(section) Chemistry, Metallurgy

(class) Organic macromolecular compounds, preparation or chemical working-up

(subclass) Macromolecular compounds

(maingroup) Unspecified macromolecular compounds

(subgroup) Biodegradable macromolecular compounds

2.2. Frequency of each IPC code in data

- **4,344 kinds of IPC codes**

They appear…

min	25%	50%	75%	95%	max
1	1	2	6	35	2830

Less informative
(cannot use them to extract recommendations)

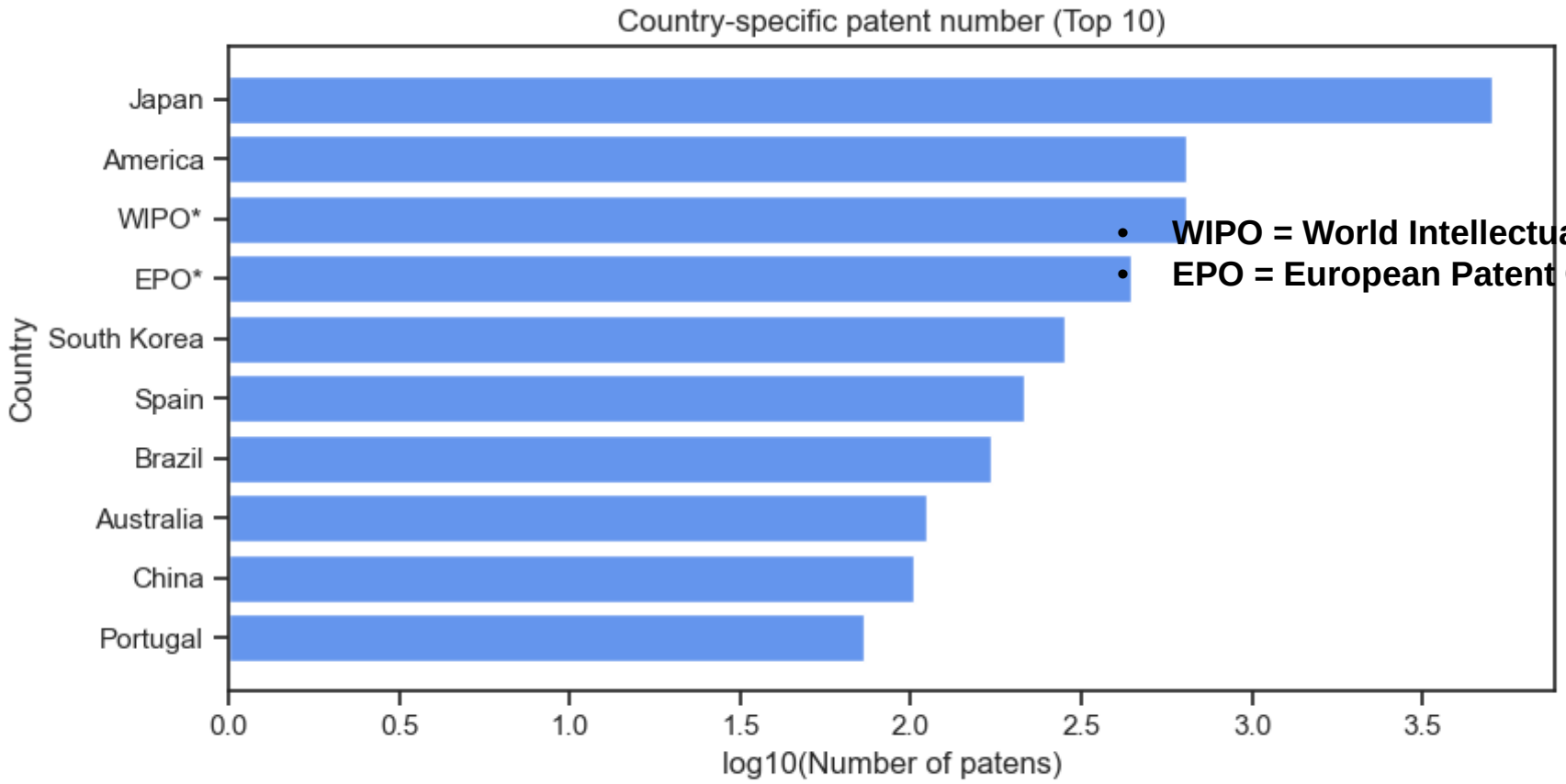
Too informative
(need to know which ones are more important)

- **Use IPC codes and abstracts to build the system**

more prospective recommendations

2.3. Countries where patents were filed

37 countries and organizations

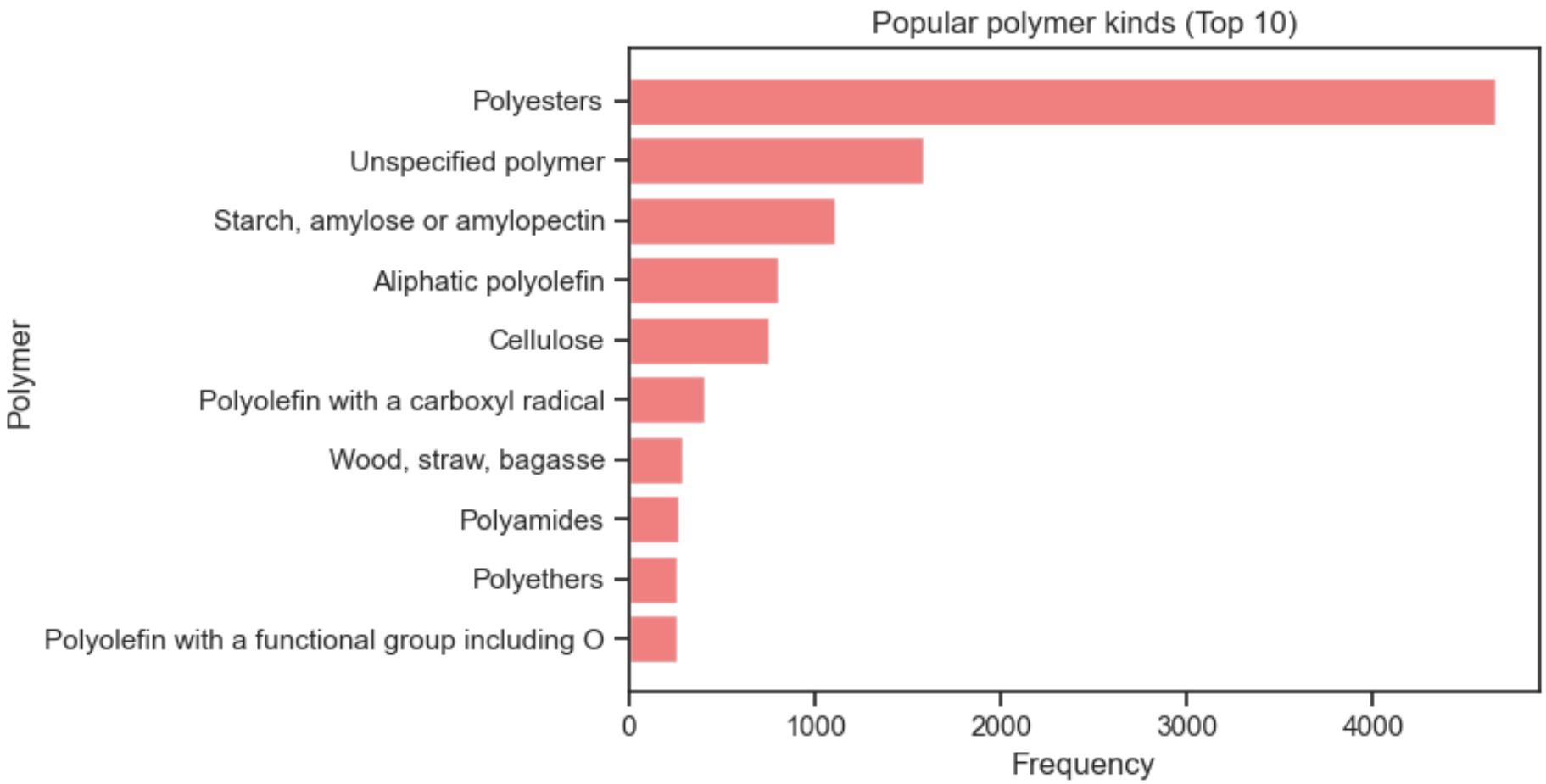


- **WIPO = World Intellectual Property Organization**
- **EPO = European Patent Office**

- **Top 2: Japan, America**
- **Influence on the prediction results**

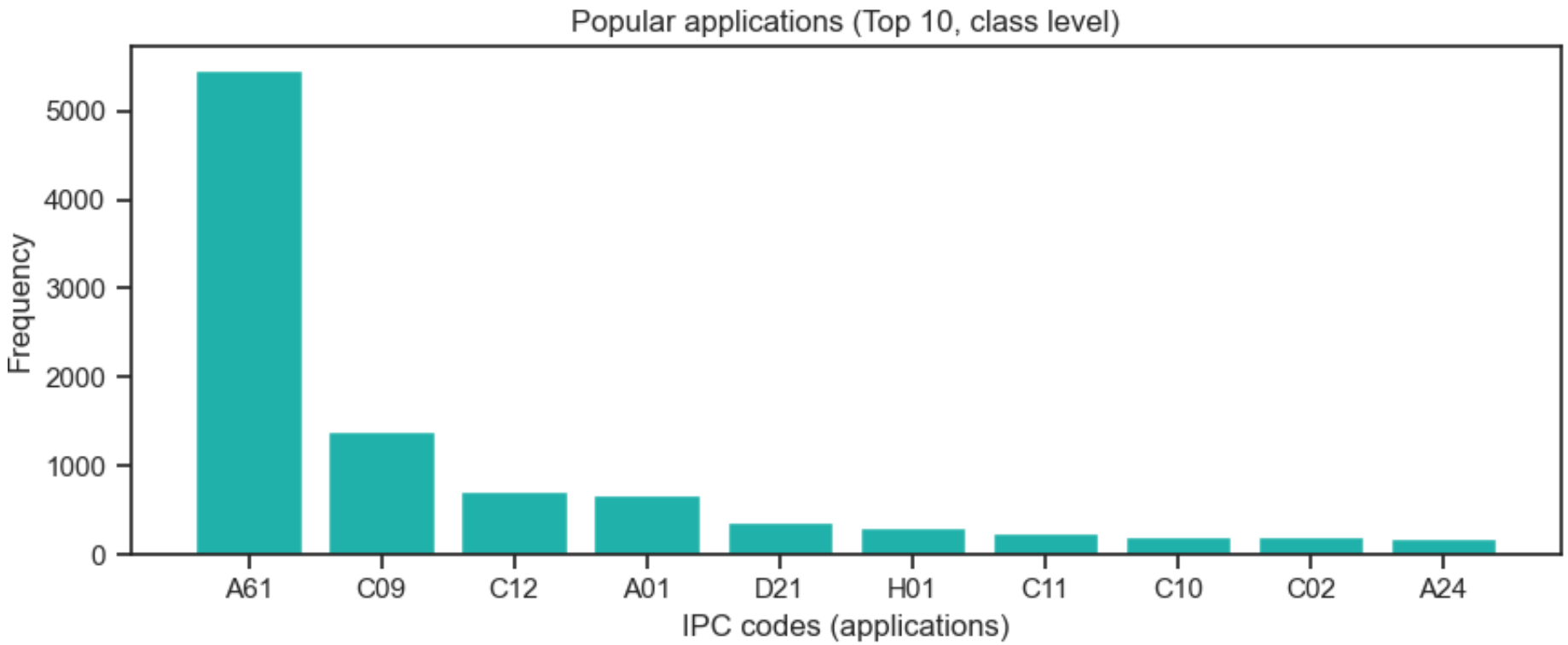
2.4. Popular polymers

49 kinds of polymers



Top: polyesters

2.5. Popular applications



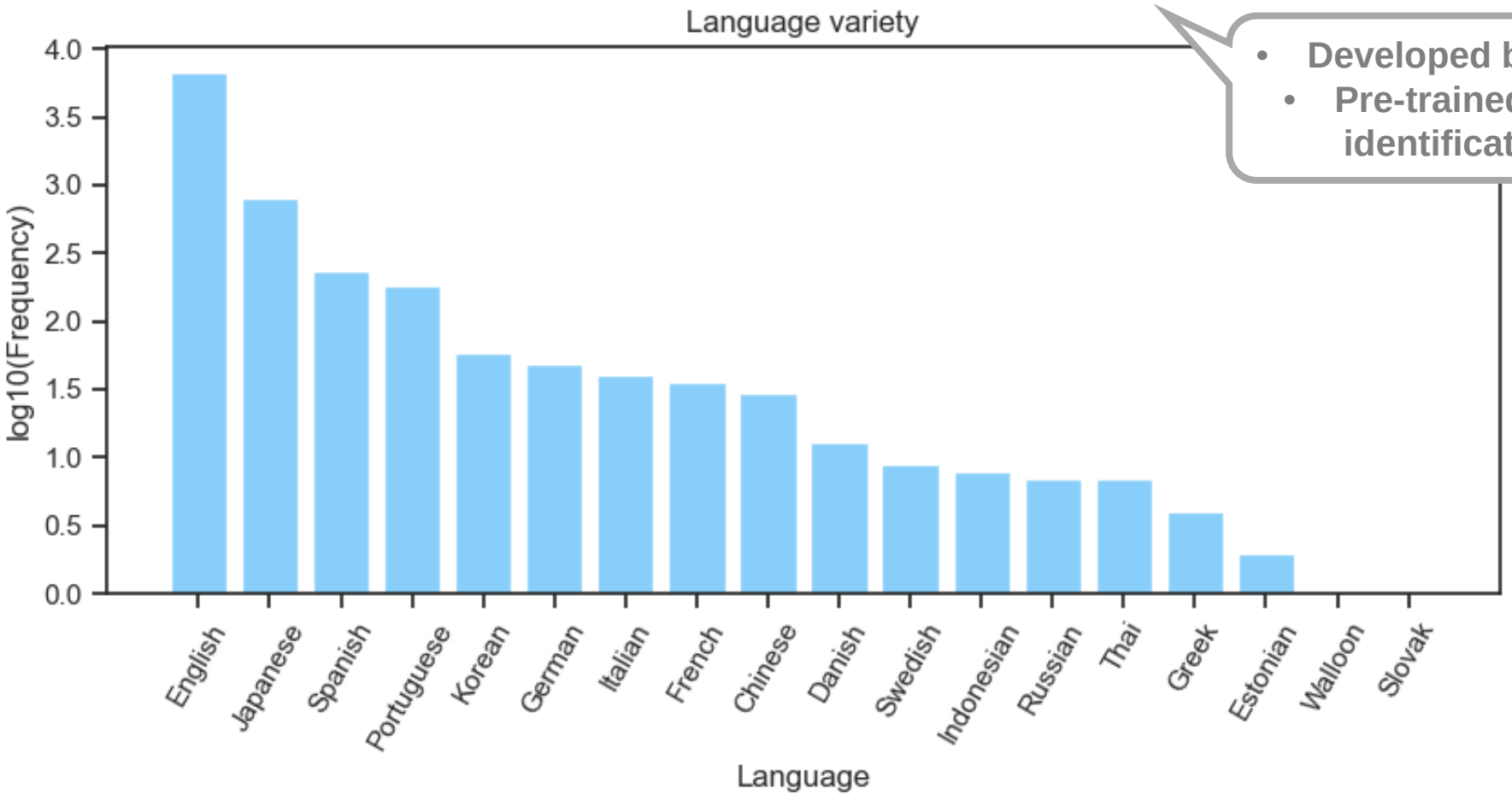
A61: medical or veterinary Science, hygiene
C09: dyes, paints, polishes, natural resins, adhesive
C12: Biochemistry, Microbiology, Enzymology, mutation or genetic engineering
A01: agriculture, forestry, animal husbandry, hunting, trapping, fishing
D21: paper-making, production of cellulose

H01: basic electric elements
C11: animal or vegetable oils, fats, fatty substances or waxes, detergents, candles
C10: petroleum, gas or coke industries
C02: treatment of water, waste water, sludge
A24: tobacco, cigarettes, simulated smoking devices

Top: medical use

2.6. Language variety

19 kinds of languages (by fastText)



- Developed by **Facebook**
- Pre-trained language identification model

Top 2: English (82%), Japanese (9.8%)

➡ **Use these abstracts for sentence similarity**

3. Data Wrangling
3.1. Preprocess the abstracts

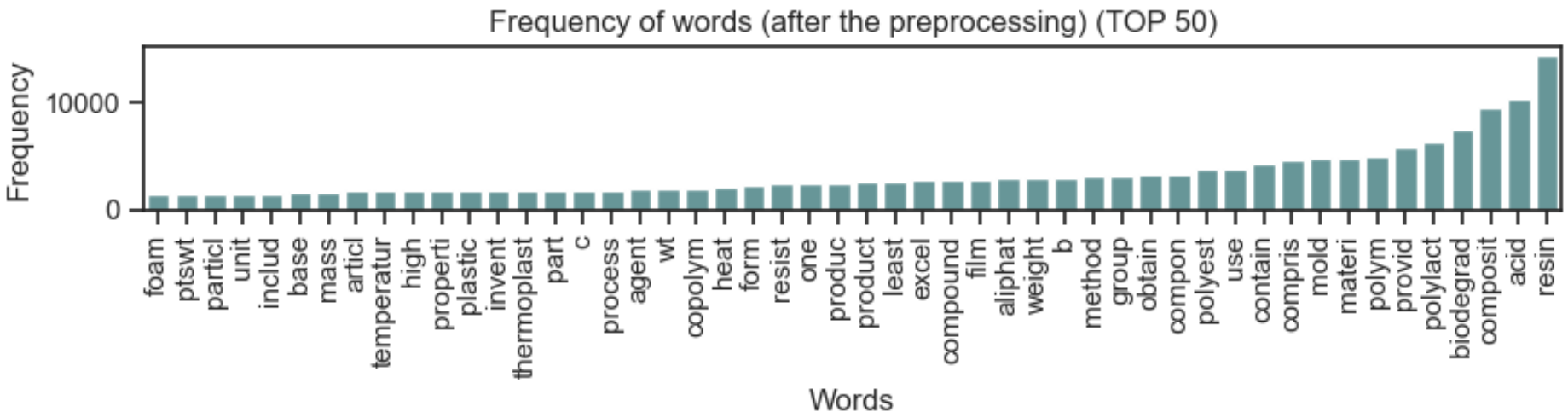
1. Translated Japanese abstracts into English (by googletrans)
2. Removed punctuations

   removed here to keep compound names

   e.g.) poly (1,5-dioxepan-2-one) → poly 15dioxepan2one

- Developed by Google
- Auto language detection

3. Tokenization
4. Removed uninformative tokens (stop words, numbers)
5. Stemming



14,487 unique tokens

4. Modeling

Two recommendation systems:

- System 1: accepts one patent as input
- System 2: accepts two IPC codes as input

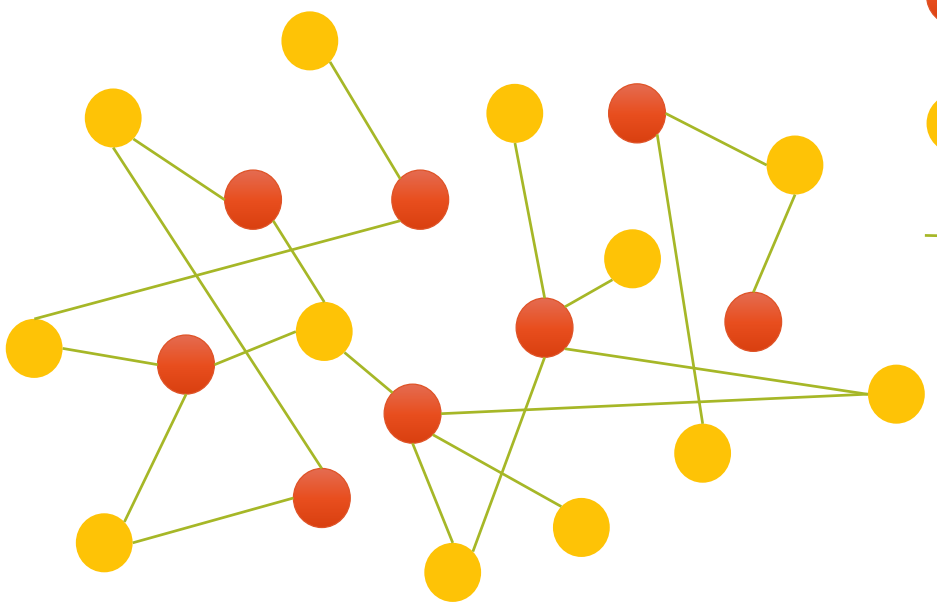
➡ Return 10 potential applications and reference patents

Their two features:

- Network: patents, IPC codes, the connections
  To extract neighbor patents and IPC codes
- Sentence similarity: abstracts
  To prioritize them

4.1. Build a network

Network Image



Patent nodes

IPC nodes

Edges (connections)

- By NetworkX
- Edges: 46,012
- Patent nodes: 8,182
- IPC nodes: 4,344

2,153 application IPC codes

10 applications will be chosen from 2,153 by the system

e.g.)

	Patent_Id	Application_Date	Country	Title	Abstract	IPC
0	AR192047768	01.06.2020	AR	COMPOSICIÓN POLIMÉRICA RELLENA CON UNA MEZCLA DE MATERIAL DE CARGA INORGÁNICO	La presente se refiere a una composición polimérica que comprende por lo menos 20,0% en peso, en...	C08L 67/02; C08L 67/04; C08L 101/16

(C08L 101/16 was excluded)

4.2. Sentence similarity

Calculated between 7,355 patents (90%)

having an English or Japanese (EN/JP) abstract

1. Tf-idf vectorizer

- **unigram bigram, and trigram**

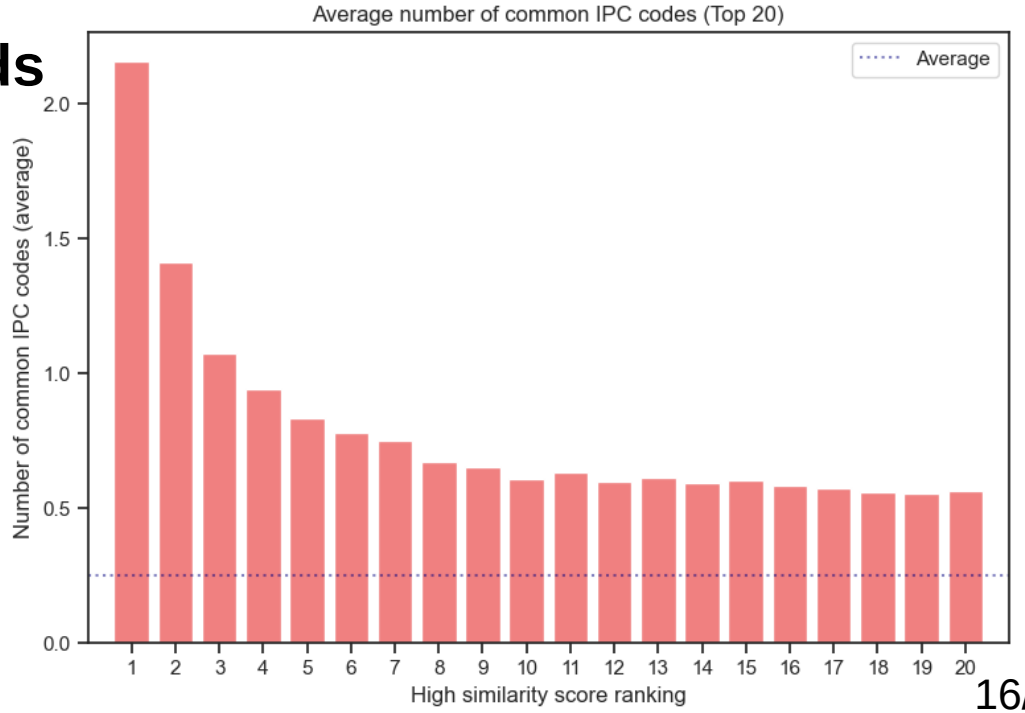
Compound names often consist of several words

e.g.) methyl methacrylate,
poly (ethylene -co- acrylic acid)

2. Cosine similarity

3. Evaluation

Average number of common IPC codes (Top 20)



4.3. Create recommendation system 1

System workflow:

* neighbor patents = patents sharing IPC codes with the input patent

1. User input: one patent
2. Get the neighbor patents*
3. Sort them by the sentence similarities
4. Get the application IPC codes of the neighbor patents

   (from the most similar patent to least until getting 10 applications)
5. Print the result

   (10 application IPC codes)

No neighbor patent → warning 1: Use the sentence similarities to get similar patents

Input patent: no EN/JP abstract → warning 2: Use the neighbor patents without sorting

Less than 10 applications → warning 3: Get the neighbor patents of the neighbor patents, go back to 3

Input patent has

- neighbor patents
- EN/JP abstract

→ Higher recommendation accuracy

4.4. Create recommendation system 2

System workflow:

1. User input: two IPC codes
2. Get the neighbor patents separately
3. Get the intersection patents → No intersection patent → Get the neighbor patents of the neighbor patents, go back to 3
4. Get the application IPC codes of the intersection patents
   (from the most frequent IPC code to least until getting 10 applications) → Less than 10 applications → warning 3: Get the neighbor patents of the intersection patents, get the new intersection patents, and go back to 4
5. Print the result
   (10 application IPC codes)

Two IPC codes inputted have intersection patents

→ Higher recommendation accuracy

4.5. Limitations of the systems

- **Input patent and IPC codes need to be in the data**
- **Recommendation accuracy can be low when:**
  - **Input patent or IPC codes have no neighbor patent**
  - **Input patent doesn't have an English or Japanese abstract**

- **Popular applications will appear more in the result**
- **Recommended applications are extracted from the applications (2,153 kinds) in the data.**

Some of them can be improved by future work
(See 7. Further work)

5. Recommend Applications (using the systems)

5.1. Example 1: use system 1

Example patent: “JP274783873”

	Patent_Id	Application_Number	Application_Date	Country	Title	Abstract	IPC	Applicants	Inventors
100	JP274783873	2018097974	22.05.2018	JP	POLYMER COMPOSITION CONTAINING PLLA AND PDLA	PROBLEM TO BE SOLVED: To provide a polymer having high heat resistant shape stability, biodegradability, and based on an organism-derived raw material and a molded article obtained from the polymer. SOLUTION: There is provided a polymer composition containing following components, a. 15 to 70 wt.% of PLLA, b. 0.1 to 15 wt.% of PDLA, c. 5 to 40 wt.% of polyester and d. 5 to 40 wt.% of an organic or inorganic filler, based on total weight of the polymer composition. Such kind of polymer composition can be biodegradated, mainly can contain bio-based carbon and can have high heat resistance. Further such kind of polymer composition can be used in a special method for manufacturing a molded article, a film or a fiber, and the molded article, the film or the fiber can be used as a container for a coffee service system because they are high in heat resistant shape stability. SELECTED DRAWING: Figure 1 COPYRIGHT: (C)2018,JPO&INPIT	C08L 67/04; C08K 3/013; C08L 67/02; C08J 5/00; C08J 5/18; C08L 101/16; D01F 6/62	BIO-TEC BIOLOGISCHE NATURVERPACKUNGEN GMBH & CO KG; バイオーテック ビオローギッシュ ナチューフェアパックンゲン ゲーエムベーハー ウント コンパニ カーゲー	SCHMIDT HARALD; ハラルド シュミット; CHRISTOPH HESS; クリストフ ヘース; WOLFGANG FRIEDEK; ウルフガング フリーデク; BECKMANN RALF; ラルフ ベックマン

- **Polylactic acid (PLA) composition**
- **Heat-resistance property**
- **Can be used for coffee capsules for a coffee service system**

Suppose

The company wants to make a new product using their PLA (other than food containers.)

Example 1

1. Execute system 1 and input "JP274783873"

```
# Accept input from a user
user_input = input('Input a patent ID (e.g., JP273590701):')

# Input 'JP274783873'

# Run the recommendation system 1
messages, appIpc_refPatents_dict = recommend_app_from_patent(user_input)
df_applications, df_references = make_dataframes(appIpc_refPatents_dict)
show_result(messages, df_applications, df_references, user_input1=user_input)
```

```
Input a patent ID (e.g., JP273590701): JP274783873
```

2. The system returns the result

```
==========================================================================================
User input: JP274783873 ,

Recommended application codes and the reference patents:
------------------------------------------------------------------------------------------
  Application IPC Code Reference Patent
0            A61J 3/07      JP274736750
1           A61L 27/00      JP271388842
2            A61F 2/84      JP271388842
3           A61L 17/00      JP271388842
4           A01G 13/02      JP270713333
5            A01G 9/14      JP270713333
6            B27N 5/00       EP13094850
7            B27N 3/02       EP13094850
8            A01G 9/10       EP13094850
9            A61K 6/10       US39070606
------------------------------------------------------------------------------------------

The reference patents:
------------------------------------------------------------------------------------------
        Patent_Id Application_Date                                    Title
4691  JP270713333       19.06.2003                ポリ乳酸系重合体組成物、その成形品、および、フィルム
6569   EP13094850       25.05.1998           Biodegradable molding material
3220  JP271388842       30.11.2006  -HYDROXY ACID POLYMER COMPOSITION AN...
439   JP274736750       23.12.2015                                成形部品の生産方法
6753   US39070606       10.03.1997  Cross-linkable or curable polylacton...
------------------------------------------------------------------------------------------
==========================================================================================
```

About PLA

0. Capsules for oral use medicines,
1. Materials for prostheses,
2. Devices providing patency to tubular structures of the body,
3. Materials for surgical sutures or ligaturing blood vessels,
4. Protective coverings for plants,
5. Greenhouses,
6. Manufacture (dry processes) of non-flat articles made from wood particles or fibers,
7. Manufacture of boards from wood particles,
8. Receptacles for seedlings,
9. Compositions for taking dental impressions.

Based on my experience…

#1 to 8: prospective

(similar material, Heat-resistance property can be a plus)

8 out of 10: prospective

Great!

5.2. Example 2: use system 2

Suppose

My company sells cellulose nanofibers as a material, and wants to develop an end-product using the cellulose nanofibers to increase profitability.

1. **Google some patents using cellulose nanofibers**
   (if you find a great competitor patent → input it on system 1)

2. **Search the IPC codes listed on them here (https://www.wipo.int), and get two IPC codes**
   - **"C08L 1/00" (cellulose)**
   - **"B82Y 30/00" (nanotechnology for materials)**

3. **Input the two IPC codes on system 2**

Example 2

1. Execute system 2 and input "C08L 1/00" and "B82Y 30/00"

```
# Accept input from a user
user_input_1 = input('Input the first IPC code (e.g., A61J 3/07):')
user_input_2 = input('Input the second IPC code (e.g., A61J 3/07):')

# Input 'C08L 1/00'
# Input 'B82Y 30/00'

# Run the recommendation system 2
messages, appIpc_refPatents_dict = recommend_app_from_2ipcs(user_input_1, user_input_2)
df_applications, df_references = make_dataframes(appIpc_refPatents_dict)
show_result(messages, df_applications, df_references, user_input1=user_input_1, user_input2=user_input_2)
```

```
Input the first IPC code (e.g., A61J 3/07):C08L 1/00

Input the second IPC code (e.g., A61J 3/07): B82Y 30/00
```

warning 3

- #0 to 2: extracted at the first cycle of the system
- #3 to 9: from the second cycle

#0 to 2 could be more useful

2. The system returns the result

```
==========================================================================
User input: C08L 1/00 , B82Y 30/00

** Rough estimate **
The second-tiers after 3 [warning3]

Recommended application codes and the reference patents:
--------------------------------------------------------------------------
  Application IPC Code Reference Patent
0          A61K 47/38    {JP289824828}
1          A61K 47/36    {JP289824828}
2           A23L 5/00    {JP289824828}
3           A24D 3/10     {EP12753518}
4           D21C 3/00     {EP13533016}
5         C09D 201/00     {EP13533016}
6           D21B 1/36     {EP13533016}
7           A01C 1/06     {EP13533016}
8           D21B 1/04     {EP13533016}
9         C09D 101/02     {EP13533016}
--------------------------------------------------------------------------

The reference patents:
--------------------------------------------------------------------------
        Patent_Id Application_Date                                Title
6984   EP12753518       01.12.1995  Cellulose ester compositions and sha...
6545   EP13533016       24.07.1998  CELLULOSE FIBER BASED COMPOSITIONS A...
48    JP289824828       05.12.2018                 セルロースナノファイバー及び澱粉を含む組成物
--------------------------------------------------------------------------
==========================================================================
```

About cellulose fibers

0. Medicinal preparations using cellulose as the non-active ingredient (e.g., carriers),
1. Medicinal preparations using polysaccharides as the non-active ingredient (e.g., carriers),
2. Preparation or treatment of foods,
3. Tobacco smoke filters using cellulose,
4. Pulping cellulose-containing materials,
5. Coating compositions based on unspecified macromolecular compounds,
6. Fibrous raw materials by dividing raw materials into small particles (e.g., fibers) by defibrating by explosive disintegration by sudden pressure reduction
7. Seed coating or dressing,
8. Fibrous raw materials or their mechanical treatment by dividing raw materials into small particles (e.g., fibers), and
9. Coating compositions based on cellulose

Recommended applications covered a broad range

Try this patent on system 1!

Example 2

Execute system 1 and input “JP289824828”

```
==========================================================================================
User input: JP289824828 ,

Recommended application codes and the reference patents:
------------------------------------------------------------------------------------------
  Application IPC Code Reference Patent
0            A61K 9/70      JP274091683
1           A61L 27/00      JP274091683        New!
2            A61K 9/06      JP274091683
3            D21C 3/00       EP13533016
4          C09D 201/00       EP13533016
5            D21B 1/36       EP13533016
6            A01C 1/06       EP13533016
7            D21B 1/04       EP13533016
8          C09D 101/02       EP13533016
9            A23L 1/00       EP13533016        New!
------------------------------------------------------------------------------------------

The reference patents:
------------------------------------------------------------------------------------------
        Patent_Id Application_Date                                    Title
6545   EP13533016       24.07.1998  CELLULOSE FIBER BASED COMPOSITIONS A...
534   JP274091683       08.05.2015  BIODEGRADABLE CELLULOSE NANOFIBER MI...  New!
------------------------------------------------------------------------------------------
==========================================================================================
```

- 4 new applications
- One new reference patent

System 1 X System 2

Get more information in the field:

- **competitors**
- **their technologies**
- **their scopes of patent claims**

6. Conclusion

- **Two recommendation systems for biodegradable polymers**
- **Accept one patent or two IPC codes as input**
- **Recommend 10 potential applications from 2,153 options**
- **Systems work in the combination of network analysis and sentence similarity**
- **Combine system 1 and 2 to get more information about the field**

7. Further work

- **Show the application names instead of IPC codes:**

  IPC codes need to look up to understand. There is a sheet including IPC code-category name. It can be used for this once the complexity is taken care of.

- **Increase the coverage of languages:**

  Not only English or Japanese, but also other languages. The accuracy will increase.

- **Try sentence embedding:**

  For example, if Doc2Vec is used instead of tf-idf vectorizer, the accuracy might increase.

- **Increase the coverage of polymer kinds:**

  To all kinds of polymers (not only biodegradable polymers)

Acknowledgment

This project was conducted as a part of Data Science Career Track Course at Springboard.

My deepest appreciation goes to them, especially to my mentor, who gave me constructive comments and warm encouragement.

Thank you

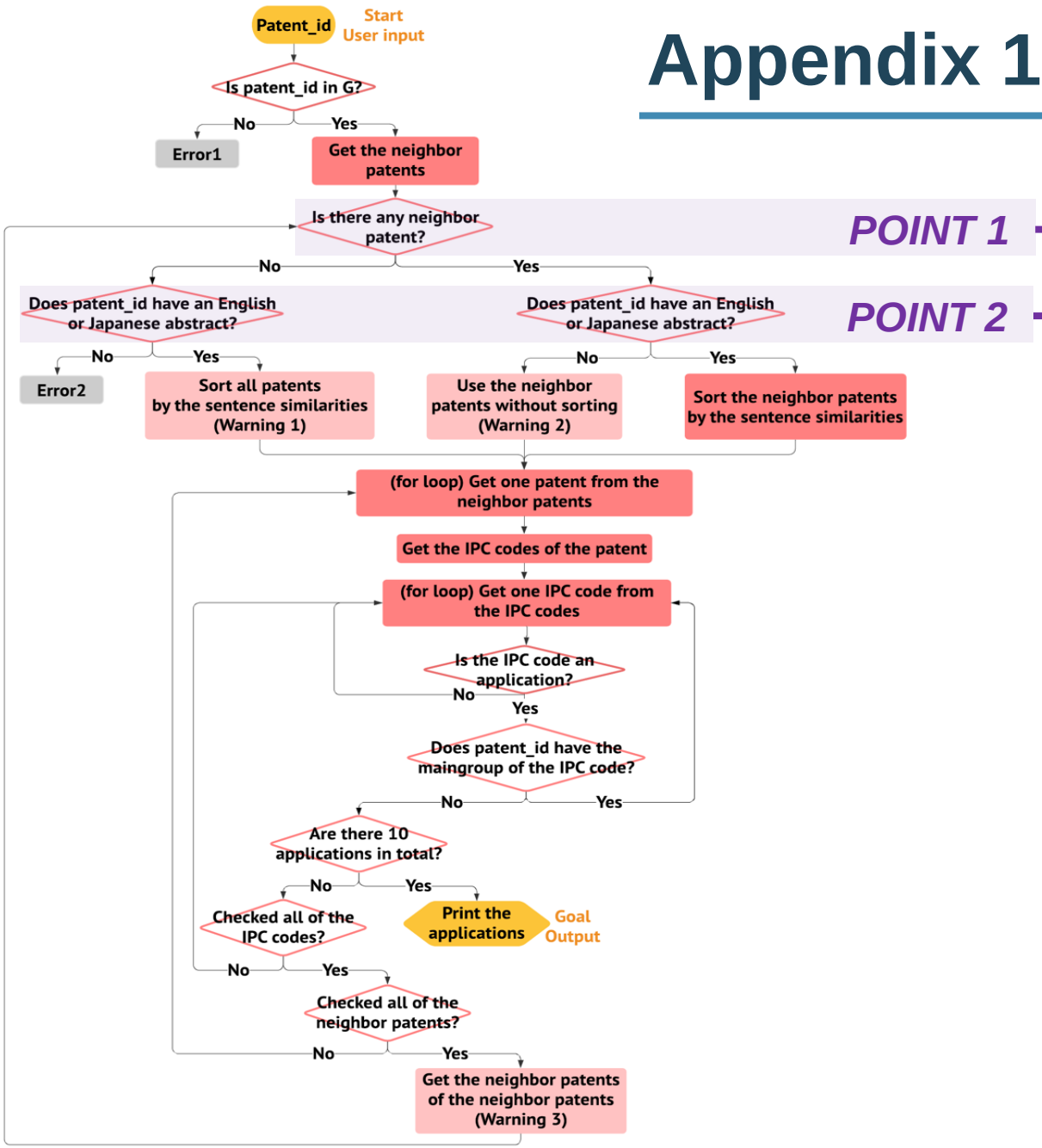
Namiko Nakashima

If you're interested in learning more about me or this project, please reach out!

LinkedIn: https://www.linkedin.com/in/namikonakashima/
GitHub: https://github.com/NamikoNa
Email: namiko.nakash@gmail.com

Appendix

Appendix 1: Recommendation system1



- Search from one patent
- Error 1, 2 → Exit the system
- Warning 1, 2, 3

  → Show warning and continue

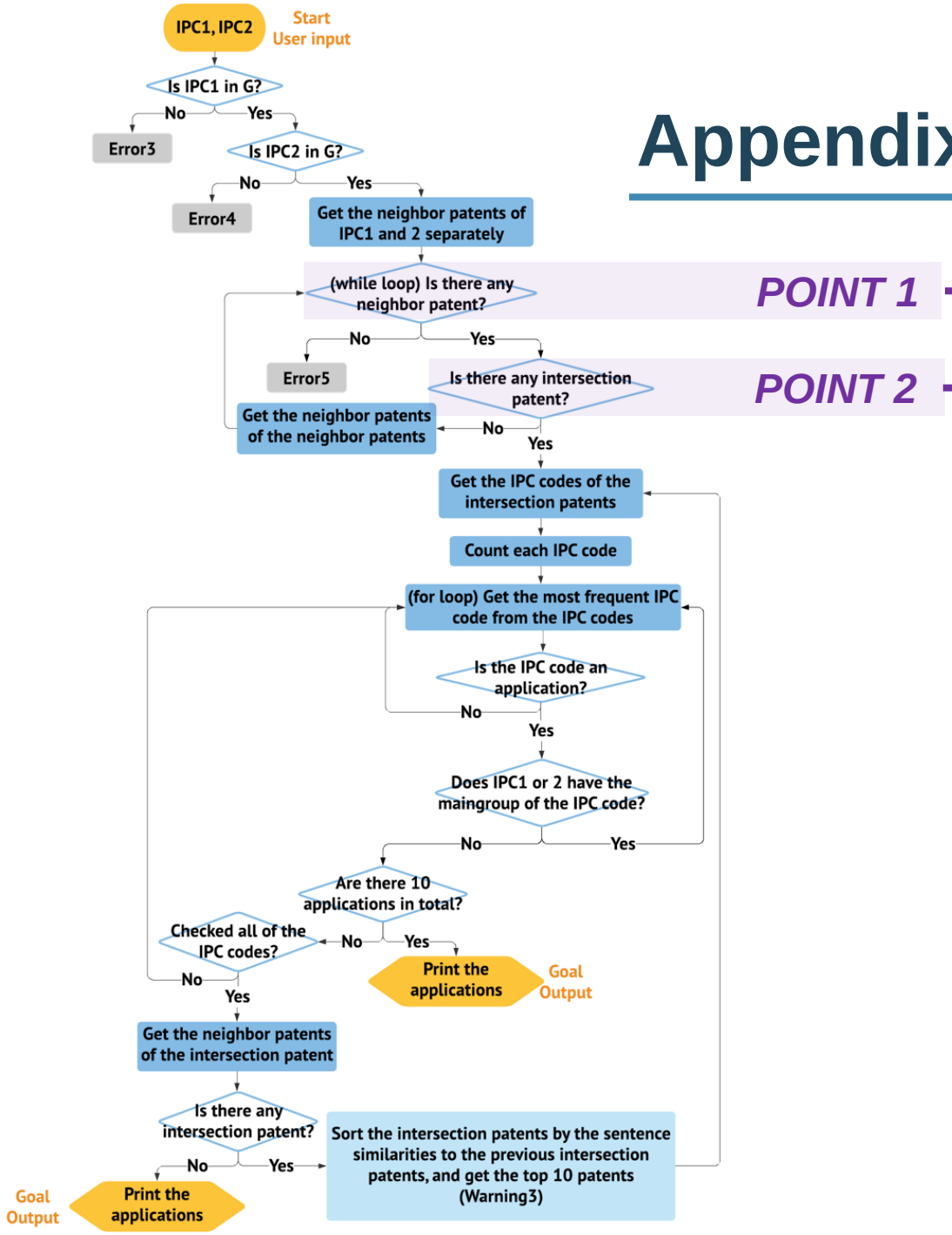
- Input patent has neighbor patents

  → Higher recommendation accuracy

- Input patent has an EN/JP abstract

  → Higher recommendation accuracy

Appendix 2: Recommendation system2



- Search from two IPC codes
- Error 3, 4, 5 → Exit the system
- Warning 3
  → Show warning and continue
- Input IPC codes have neighbor patents
  → Continue
- There are intersection patents
  → Higher recommendation accuracy