Mosty w nauce mosty

2 marca 2020

sciencebridges.umk.pl

# Niepewność wyników wizualizacji danycH

Veslava Osińska

# Sondaże a słupki


late poll z 2018 r.

# Błędy/manipulacje (2)

**DANE**

Stężenie w Krakowie przekroczyło **10**x dawkę uznaną przez WHO za bezpieczną

# Plan wystąpienia

1. Wizualizacje sieci złożonych
2. Bigdata a wizualizacje – widzenie makroskopowe
3. Algorytmy machine learning w wizualizacji danych
4. Mapa nauki Polskich czasopism
5. Problem reprodukowalności
6. Niepewność w reprezentacji wiedzy

**28 slajdów**

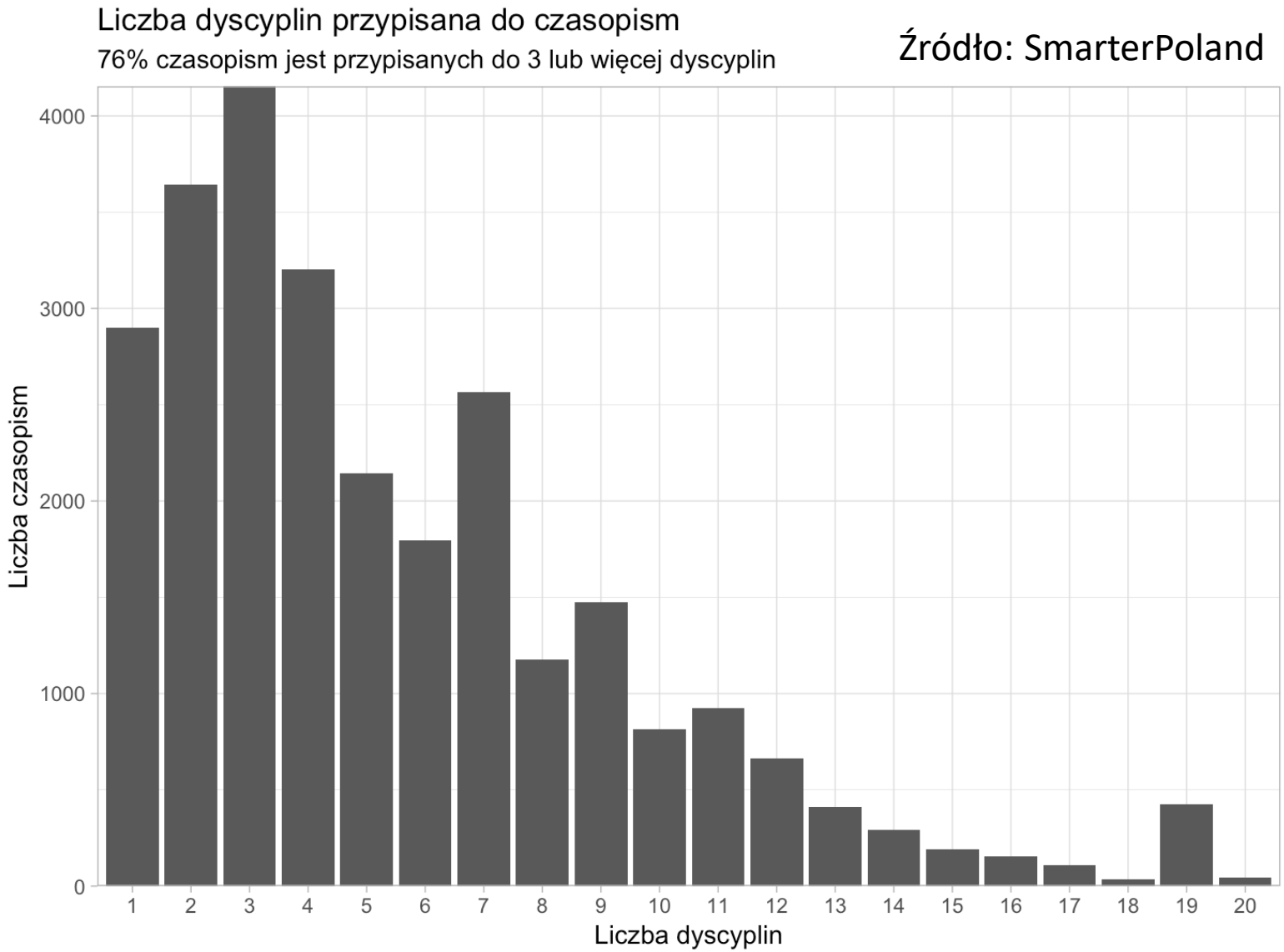
# Wykaz czasopism MNiSW z 2018

| L.p. | Tytuł w bazie Web of Science | Tytuł w bazie Scopus | Dyscypliny |
|---|---|---|---|
| 1 | 2D Materials | 2D Materials | inżynieria biomedyczna; inżynieria chemiczna; inżynieria lądowa i transport; inżynieria materiałowa; inżynieria mechaniczna; inżynieria środowiska, górnictwo i energetyka; nauki farmaceutyczne; rolnictwo i ogrodnictwo; technologia żywności i żywienia; nauki chemiczne; nauki fizyczne; |
| 2 | 3 Biotech | 3 Biotech | inżynieria biomedyczna; inżynieria środowiska, górnictwo i energetyka; nauki farmaceutyczne; nauki leśne; rolnictwo i ogrodnictwo; technologia żywności i żywienia; weterynaria; zootechnika i rybactwo; nauki biologiczne; nauki o Ziemi i środowisku; |
| 3 | 3C Empresa | | ekonomia i finanse; nauki o zarządzaniu i jakości; |
| 4 | 3c Tecnologia | | architektura i urbanistyka; automatyka, elektronika i elektrotechnika; informatyka techniczna i telekomunikacja; inżynieria biomedyczna; inżynieria chemiczna; inżynieria lądowa i transport; inżynieria materiałowa; inżynieria mechaniczna; inżynieria środowiska, górnictwo i energetyka; |
| 5 | 3C Tic | | informatyka techniczna i telekomunikacja; informatyka ; |
| 6 | 3D Printing and Additive Manufacturing | 3D Printing and Additive Manufacturing | automatyka, elektronika i elektrotechnika; informatyka techniczna i telekomunikacja; inżynieria biomedyczna; inżynieria chemiczna; inżynieria materiałowa; inżynieria mechaniczna; nauki chemiczne; nauki fizyczne; |
| 7 | 3D Research | 3D Research | automatyka, elektronika i elektrotechnika; informatyka techniczna i telekomunikacja; inżynieria biomedyczna; informatyka ; |
| 8 | 3L-Language Linguistics Literature-The Southeast Asian Journal of English Language Studies | 3L: Language, Linguistics, Literature | językoznawstwo; literaturoznawstwo; nauki o komunikacji społecznej i mediach; |
| 9 | 452 F-Revista de Teoria de la Literatura y Literatura Comparada | | literaturoznawstwo; |
| 10 | 4OR-A Quarterly Journal of Operations Research | 4OR | informatyka techniczna i telekomunikacja; nauki o komunikacji społecznej i mediach; nauki o zarządzaniu i jakości; informatyka ; |
| 11 | A & A Practice | A & A case reports | inżynieria biomedyczna; nauki farmaceutyczne; nauki medyczne; nauki o kulturze fizycznej; nauki o zdrowiu; zootechnika i rybactwo; psychologia; |
| 12 | A + U-ARCHITECTURE AND URBANISM | A + U-Architecture and Urbanism | nauki o kulturze i religii; nauki o sztuce; architektura i urbanistyka; geografia społeczno-ekonomiczna i gospodarka przestrzenna; nauki socjologiczne; |
| 13 | A&C-Revista de Direito Administrativo & Constitucional | | nauki o bezpieczeństwie; nauki prawne; prawo kanoniczne; |
| 14 | AAA-ARBEITEN AUS ANGLISTIK UND AMERIKANISTIK | AAA, Arbeiten aus Anglistik und Amerikanistik | językoznawstwo; literaturoznawstwo; nauki o komunikacji społecznej i mediach; |
| 15 | AACA Digital | | nauki o sztuce; |
| 16 | AACN Advanced Critical Care | AACN Advanced Critical Care | inżynieria biomedyczna; nauki farmaceutyczne; nauki medyczne; nauki o kulturze fizycznej; nauki o zdrowiu; zootechnika i rybactwo; psychologia; |
| 17 | AAPG BULLETIN | AAPG Bulletin | inżynieria chemiczna; inżynieria lądowa i transport; inżynieria mechaniczna; inżynieria środowiska, górnictwo i energetyka; geografia społeczno-ekonomiczna i gospodarka przestrzenna; nauki chemiczne; nauki o Ziemi i środowisku; |
| 18 | AAPS Journal | AAPS Journal | nauki farmaceutyczne; |
| 19 | AAPS PHARMSCITECH | AAPS PharmSciTech | inżynieria biomedyczna; inżynieria środowiska, górnictwo i energetyka; nauki farmaceutyczne; nauki medyczne; nauki o kulturze fizycznej; nauki o zdrowiu; nauki leśne; rolnictwo i ogrodnictwo; technologia żywności i żywienia; weterynaria; zootechnika i rybactwo; geografia społeczno-ekonomiczna i gospodarka przestrzenna; psychologia; nauki biologiczne; nauki o Ziemi i środowisku; |

# Tabela.Statystyki

44

| Dyscyplina | Liczba |
|---|---|
| astronomia | 591 |
| architektura i urbanistyka | 993 |
| nauki o sztuce | 1581 |
| matematyka | 1607 |
| nauki prawne | 1733 |
| nauki teologiczne | 1908 |
| nauki fizyczne | 2135 |
| językoznawstwo | 2175 |
| automatyka, elektronika i elektrotechnika | 2206 |
| inżynieria lądowa i transport | 2328 |
| literaturoznawstwo | 2354 |
| pedagogika | 2494 |
| informatyka | 2510 |
| archeologia | 2544 |
| nauki o komunikacji społecznej i mediach | 2649 |
| ekonomia i finanse | 2669 |
| historia | 2710 |
| inżynieria materiałowa | 2721 |
| nauki chemiczne | 2898 |
| nauki o kulturze i religii | 2929 |
| prawo kanoniczne | 2934 |
| nauki o Ziemi i środowisku | 3074 |
| inżynieria chemiczna | 3133 |



# Tabela.Relacje

- Relacje pomiędzy czasopi
  dyscyplin
- Metryka podobieństwa: l
- Rozmiar macierzy: 44x44
- Im większa liczba tym ści
- Waga relacji (liczba) decy

| | astronomia | architektura i urbanistyka | nauki o sztuce | matematyka | nauki prawne | nauki teologiczne | nauki fizyczne | językoznawstwo | automatyka, elektronika i elektrotechnika | inżynieria lądowa i transport | literaturoznawstwo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **astronomia** | 591 | 175 | 138 | 196 | 137 | 138 | 568 | 138 | 305 | 244 | 138 |
| **architektura i urbanistyka** | 175 | 993 | 205 | 220 | 161 | 149 | 297 | 147 | 681 | 806 | 147 |
| **nauki o sztuce** | 138 | 205 | 1581 | 144 | 602 | 916 | 148 | 1030 | 156 | 156 | 1070 |
| **matematyka** | 196 | 220 | 144 | 1607 | 155 | 158 | 491 | 146 | 362 | 300 | 145 |
| **nauki prawne** | 137 | 161 | 602 | 155 | 1733 | 659 | 145 | 634 | 169 | 164 | 626 |
| **nauki teologiczne** | 138 | 149 | 916 | 158 | 659 | 1908 | 141 | 934 | 145 | 147 | 986 |
| **nauki fizyczne** | 568 | 297 | 148 | 491 | 145 | 141 | 2135 | 144 | 915 | 614 | 144 |
| **językoznawstwo** | 138 | 147 | 1030 | 146 | 634 | 934 | 144 | 2175 | 164 | 146 | 1801 |
| **automatyka, elektronika i ele** | 305 | 681 | 156 | 362 | 169 | 145 | 915 | 164 | 2206 | 943 | 147 |
| **inżynieria lądowa i transport** | 244 | 806 | 156 | 300 | 164 | 147 | 614 | 146 | 943 | 2328 | 146 |
| **literaturoznawstwo** | 138 | 147 | 1070 | 145 | 626 | 986 | 144 | 1801 | 147 | 146 | 2354 |

**Wykaz czasopism naukowych**

| Lp. | Tytuł 1 | archeologia | filozofia | historia | językoznawstwo | literaturoznawstwo | nauki o kulturze i religii | nauki o sztuce | architektura i urbanistyka | automatyka, elektronika i elektrotechnika | informatyka techniczna i telekomunikacja | inżynieria biomedyczna | inżynieria chemiczna | inżynieria lądowa i transport | inżynieria materiałowa | inżynieria mechaniczna | inżynieria środowiska, górnictwo i energetyka | nauki farmaceutyczne | nauki medyczne | nauki o kulturze fizycznej | nauki o zdrowiu | nauki leśne | rolnictwo i ogrodnictwo | technologia żywności i żywienia | weterynaria | zootechnika i rybactwo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 201 | 202 | 203 | 204 | 205 | 206 | 207 | 208 | 209 | 301 | 302 | 303 | 304 | 401 | 402 | 403 | 404 | 405 |
| 1 | 2D Materials | | | | | | | | x | x | | x | | x | x | x | | x | | | | | | | | |
| 2 | 3 Biotech | | | | | | | | x | | | x | x | | | | x | x | | | x | | x | | x | |
| 3 | 3C Empresa | | | | | | | | x | | | | | | | | | | | | | | | | | |
| 4 | 3c Tecnologia | | | | | | | | | | | x | | x | | x | x | | | | | | | | | |
| 5 | 3C Tic | | | | | | | | x | | | x | | | | | | | | | | | | | | |
| 6 | 3D Printing and Additive Manufacturing | | | | | | | | x | x | | x | x | | x | x | | | | | | | | | | |
| 7 | 3D Research | | | | | | | | x | x | x | x | | | | x | | | | | | | | | | |
| 8 | 3L-Language Linguistics Literature-The Southeast Asian Journal of English Language Studies | | | | x | x | | | | | | | | | | | | | | | | | | | | |
| 9 | 452 F-Revista de Teoria de la Literatura y Literatura Comparada | | | | | x | | | | | | | | | | | | | | | | | | | | |
| 10 | 4OR-A Quarterly Journal of Operations Research | | | | | | x | | | | x | | | | | | | | | | | | | | | |
| 11 | A & A Practice | | | | | | | | | | | x | | | | | | | x | | x | | | | x | |
| 12 | A + U-ARCHITECTURE AND URBANISM | | | | | | x | x | x | | | | | | | | | | | | | | | | | |
| 13 | A&C-Revista de Direito Administrativo & Constitucional | | | | | | | | | | | | | | | | | | | | | | | | | |
| 14 | AAA-ARBEITEN AUS ANGLISTIK UND AMERIKANISTIK | | | | x | x | | | | | | | | | | | | | | | | | | | | |
| 15 | AACA Digital | | | | | | | x | | | | x | | | | | | | | | | | | | | |
| 16 | AACN Advanced Critical Care | | | | | | | | | | | x | | | | | | x | x | | x | | | | x | |
| 17 | AAPG BULLETIN | | | | | | | | | | | | | x | | | | | | | | | | | | |
| 18 | AAPS Journal | | | | | | | | | | | | | | | | | x | x | | | | | | x | |
| 19 | AAPS PHARMSCITECH | | | | | | | | | | | x | | | x | | | x | x | | x | | | | | |
| 20 | AATCC Journal of Research | | | | | | | | | | | x | | | x | x | | x | | | | | | | | |

Czasopisma | Konferencje

# Wizualizacja

Dyscypliny (44) grupuje się w dziedziny (8)

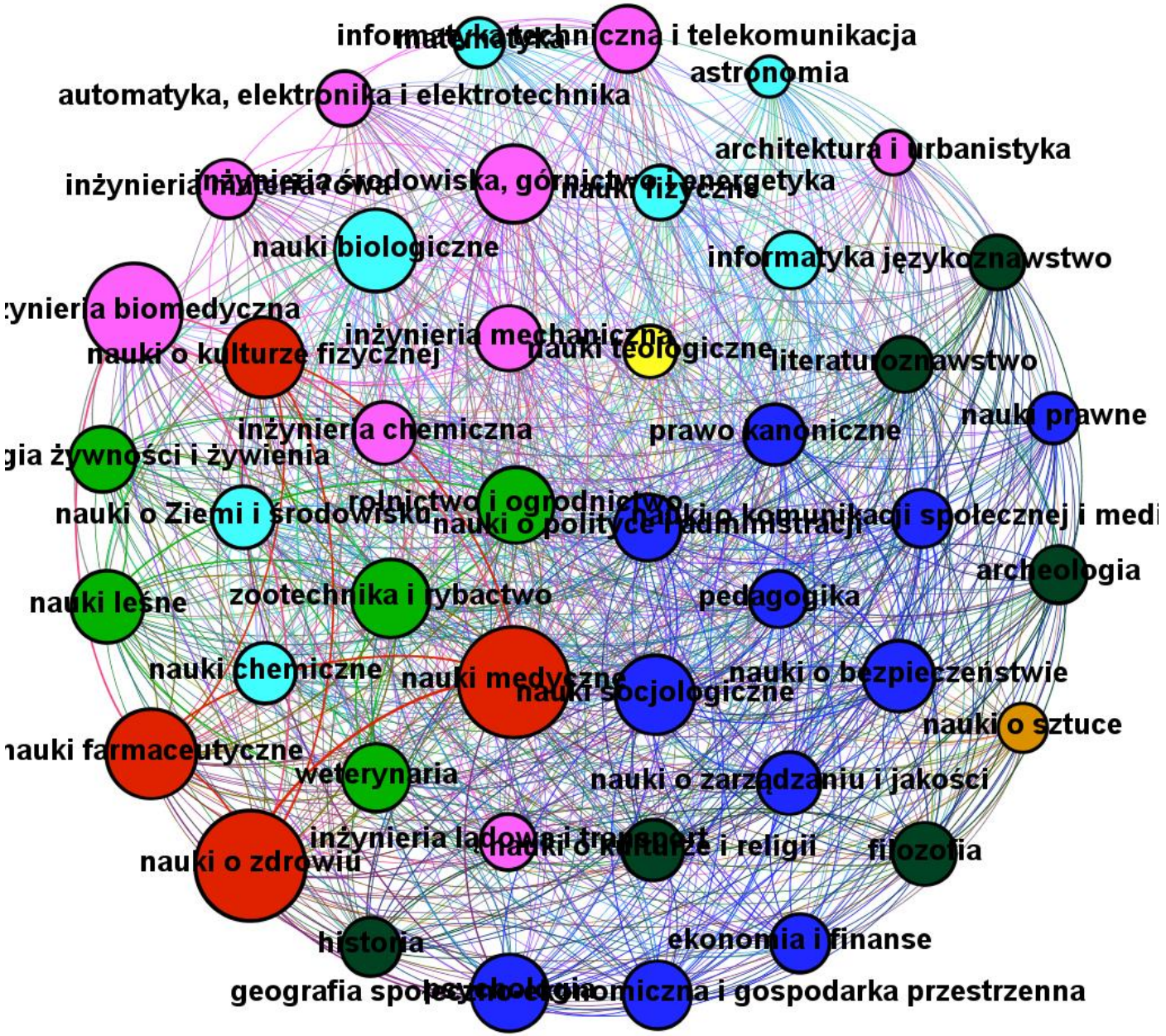
**Kodowanie:**

Dziedzina –> kolor

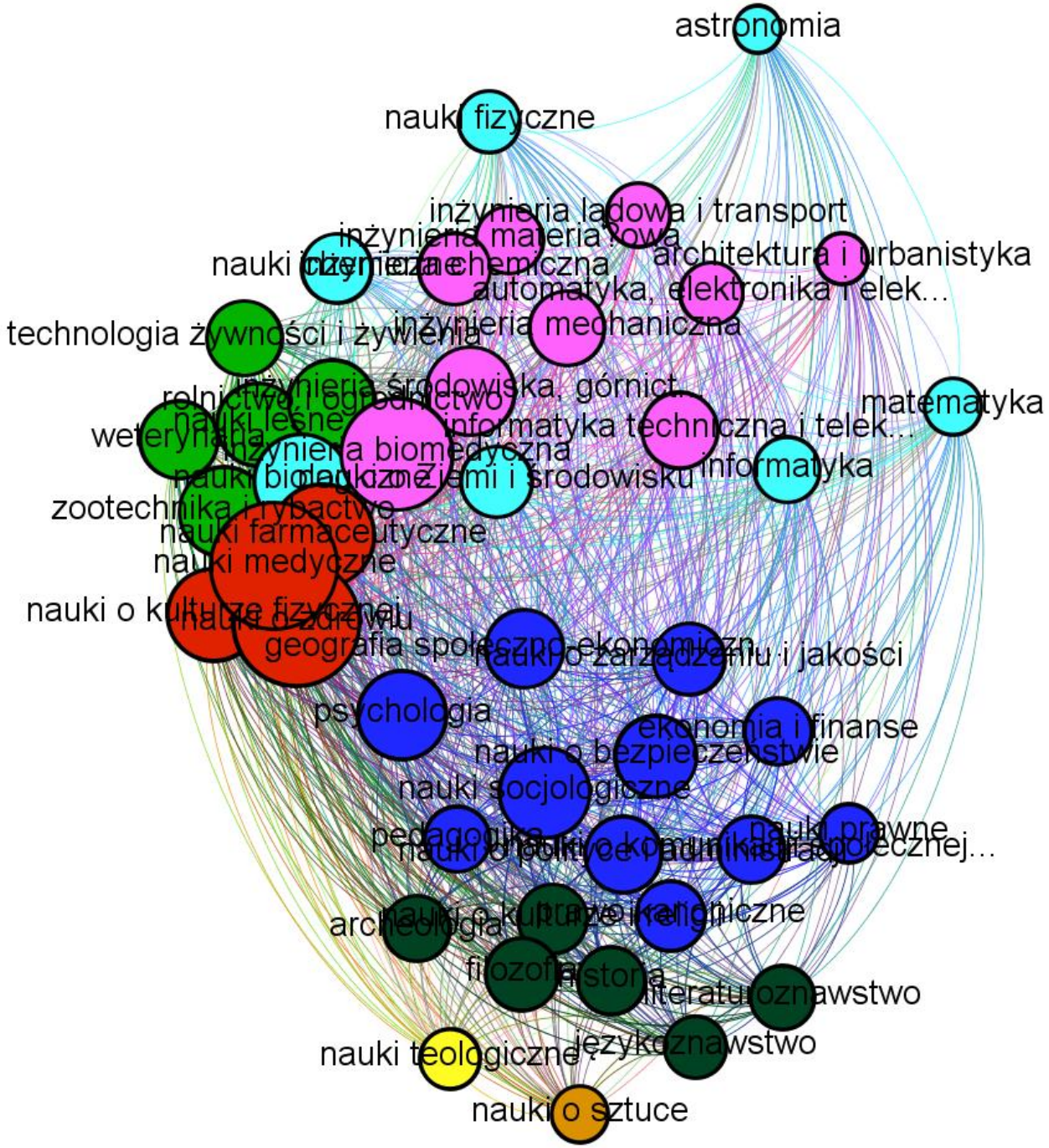
Dyscyplina –> node

Liczba czasopism w danej dyscyplinie –> rozmiar nodu

# Frusherman Layout



# Spring Layout



Dane **takie same**

**ale**

**różne** wyniki wizualizacji sieciowej

Paradoks ?

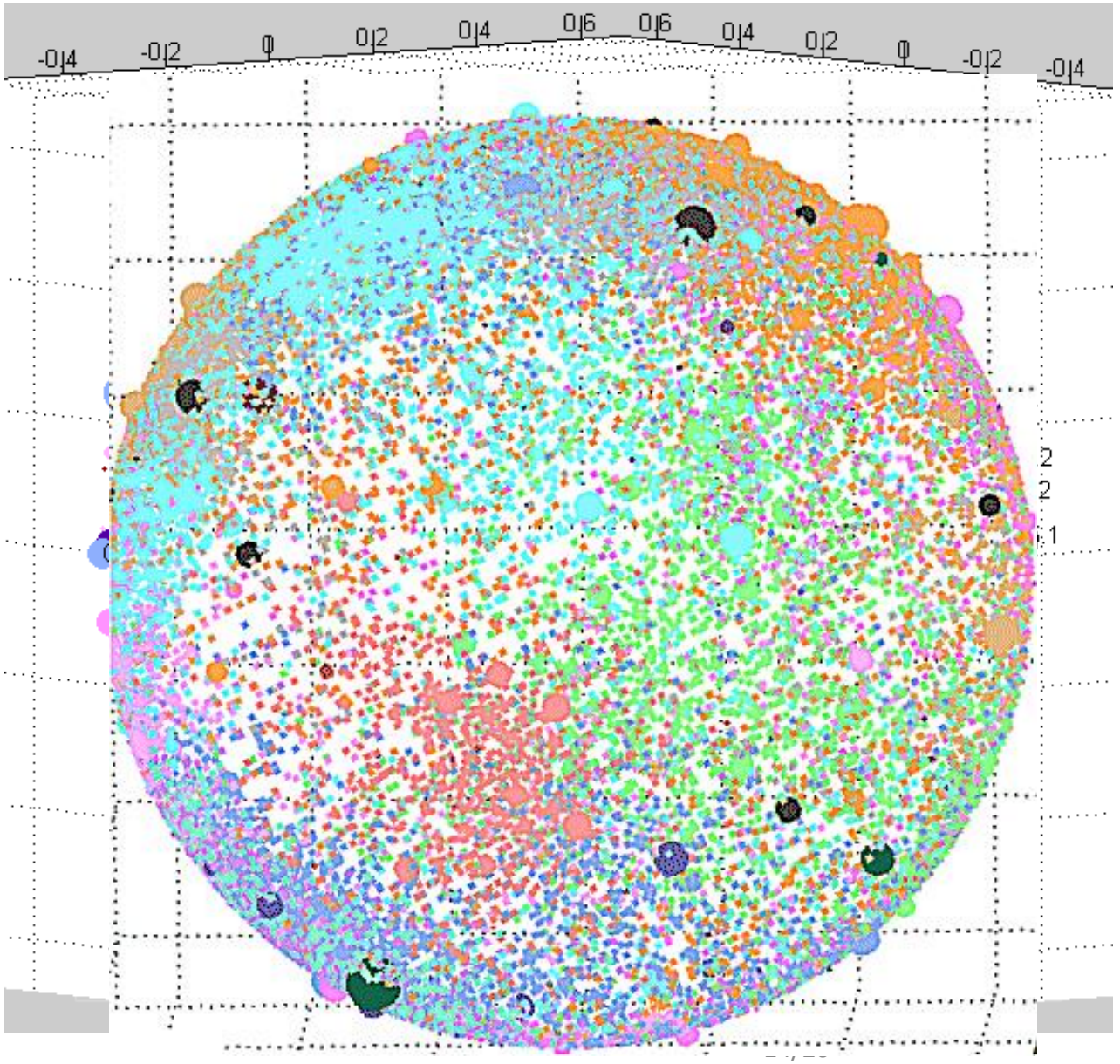
Paradygmat bigdata?

Layout (wizualizacja) ma służyć:

- odkrywaniu wiedzy o danych
- znalezieniu lub nie potencjalnych grup
- wykryciu relacji

**dane z tabeli definiują związki**

# Pobyt w DANS, Amsterdam (2009 r.)





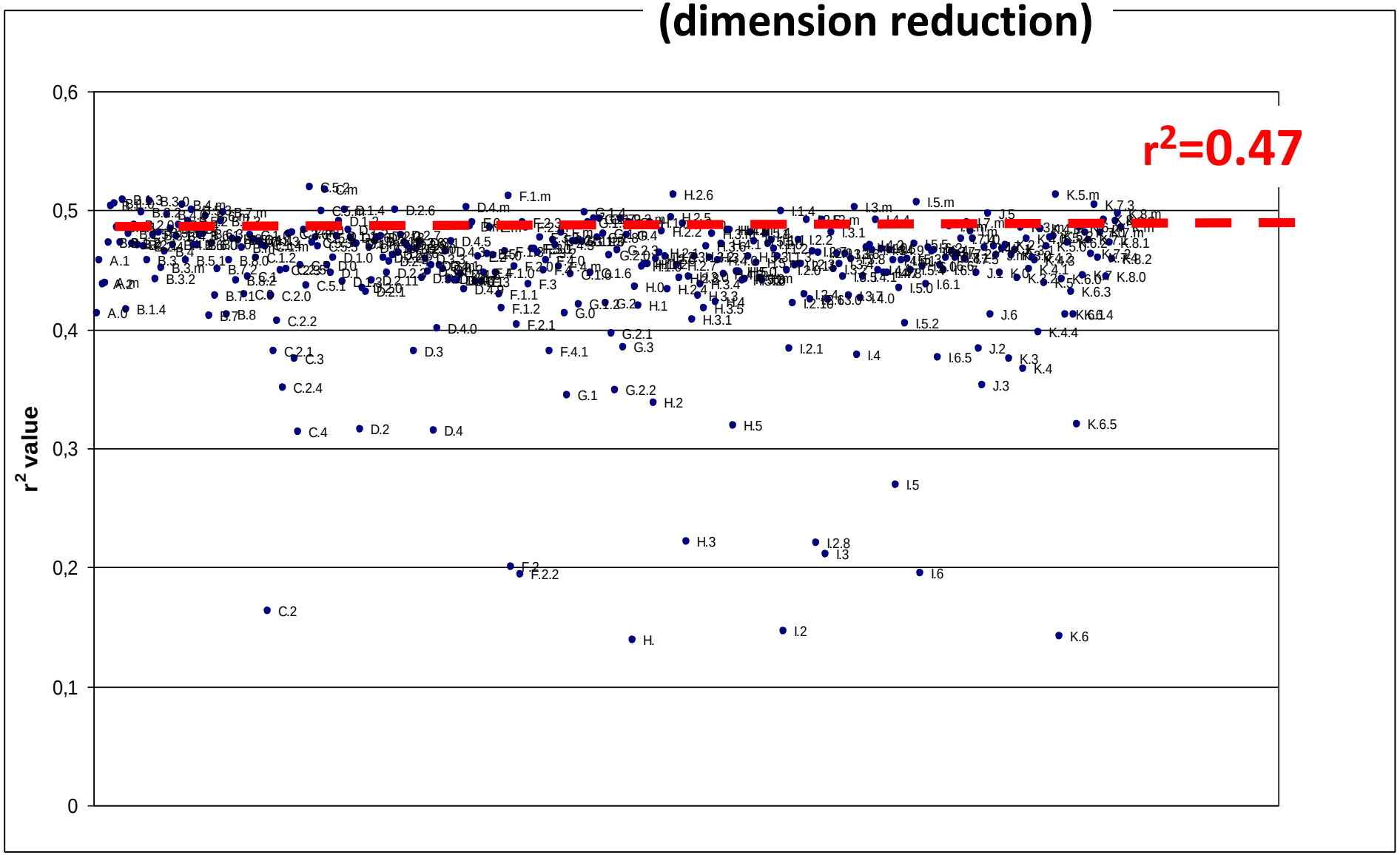
ACM
Digital Library

**Atributes:**

1. **main class (color) 11**
2. **level 1,2,3 (luminosity)**
3. **population (size)**

# The distribution $r^2$ **MDS** algorithm

**(dimension reduction)**

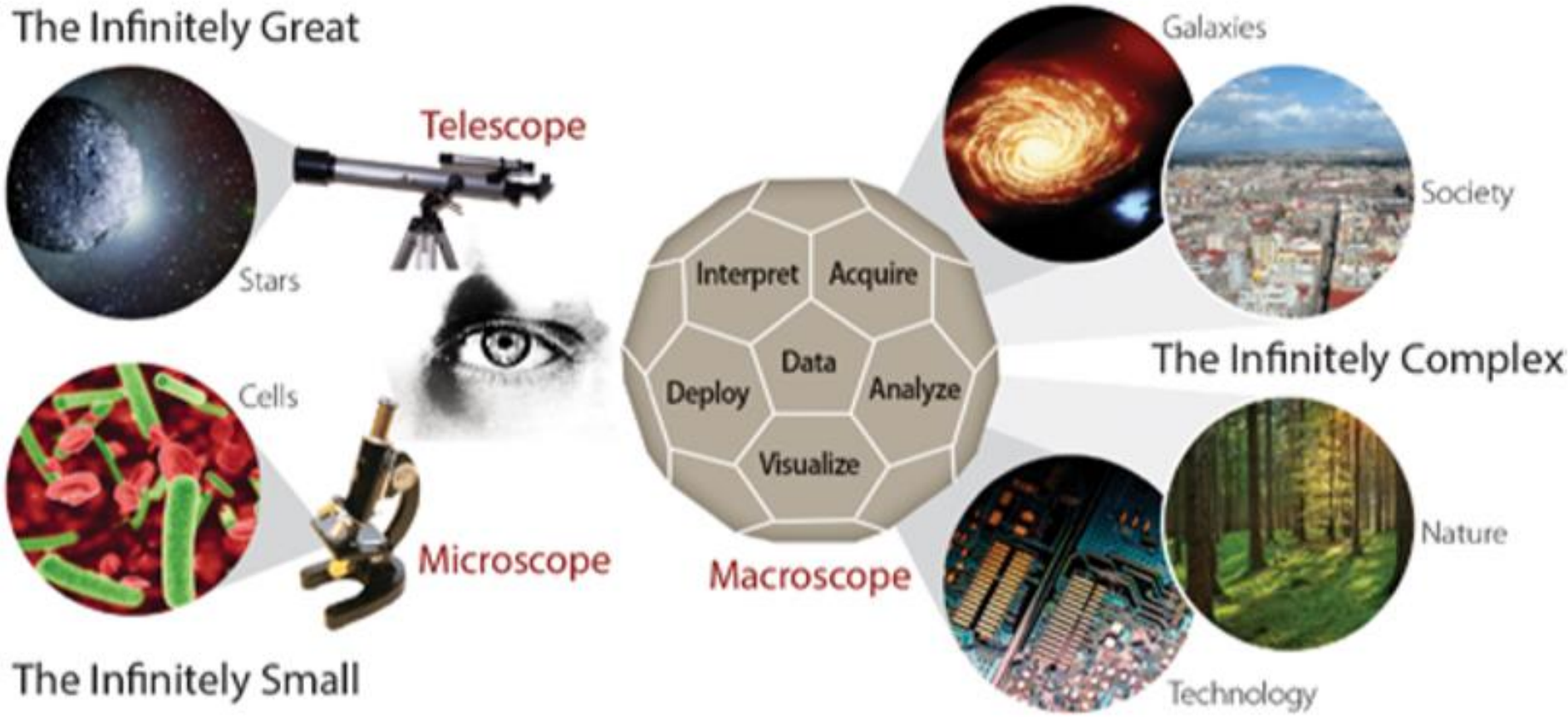


# Bigdata. Cechy

- Zrezygnowanie z **precyzji** (kosztem zwiększenia próby losowej)
- Zaniechanie poszukiwania **przyczynowości** zjawisk (na rzecz szacowania **prawdopodobieństwa**)
- Specjalistyczna **wiedza ekspercka** traci na znaczeniu i jest zastępowana przez: **prawdopodobieństwo i korelację** – wymusza to dostosowanie się do analiz big data

# Makroskopowa perspektywa (macroscopic view)

## 1979, Joël de Rosnay *The Macroscope: A New World Scientific System*



# Machine learning algorithms

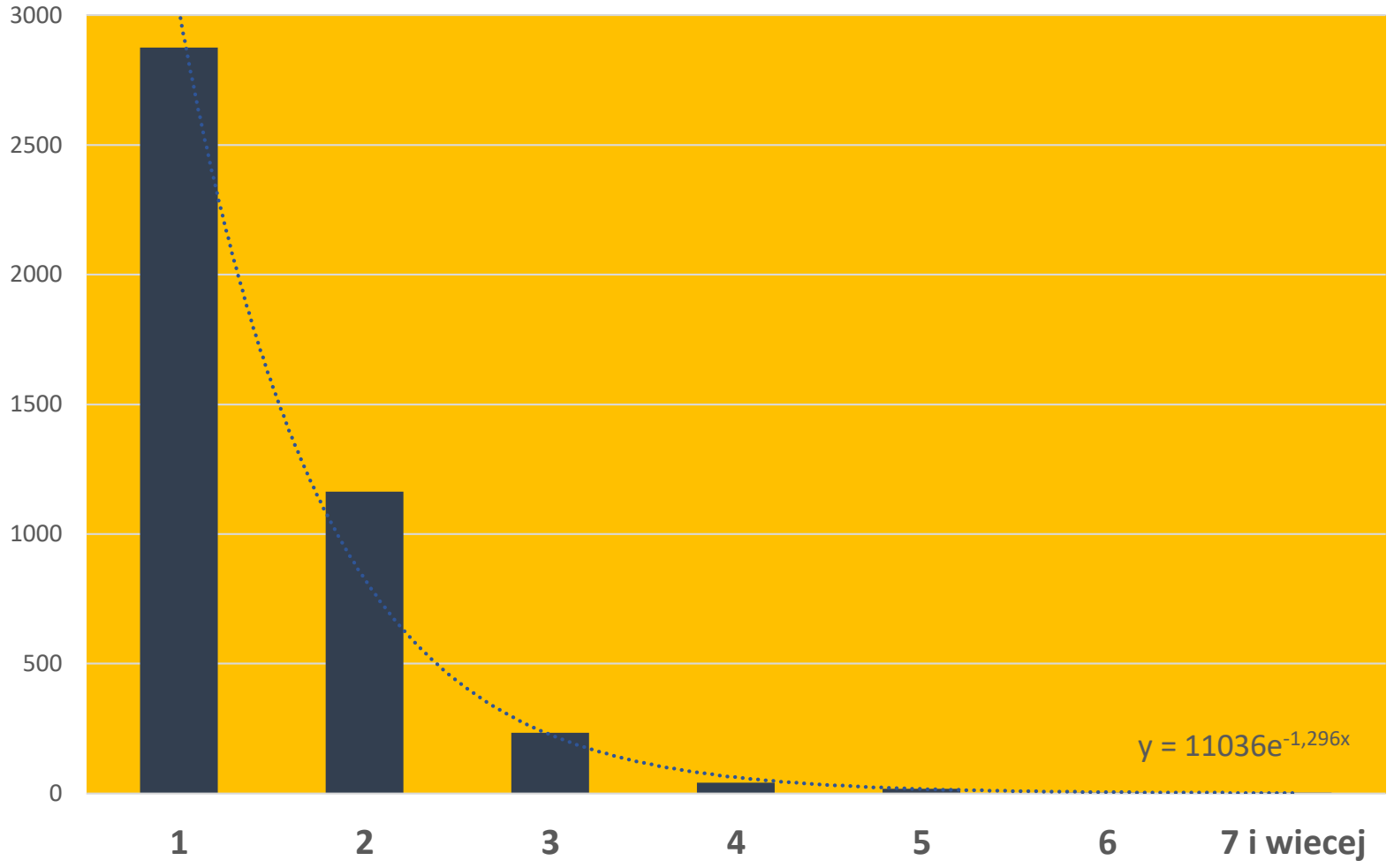
- Linear regression
- Logical regression
- Classification and regression trees
- K-nearest neighbor (KNN)
- Naïve Bayes
- **t-SNE**

**5 najpopularniejszych**

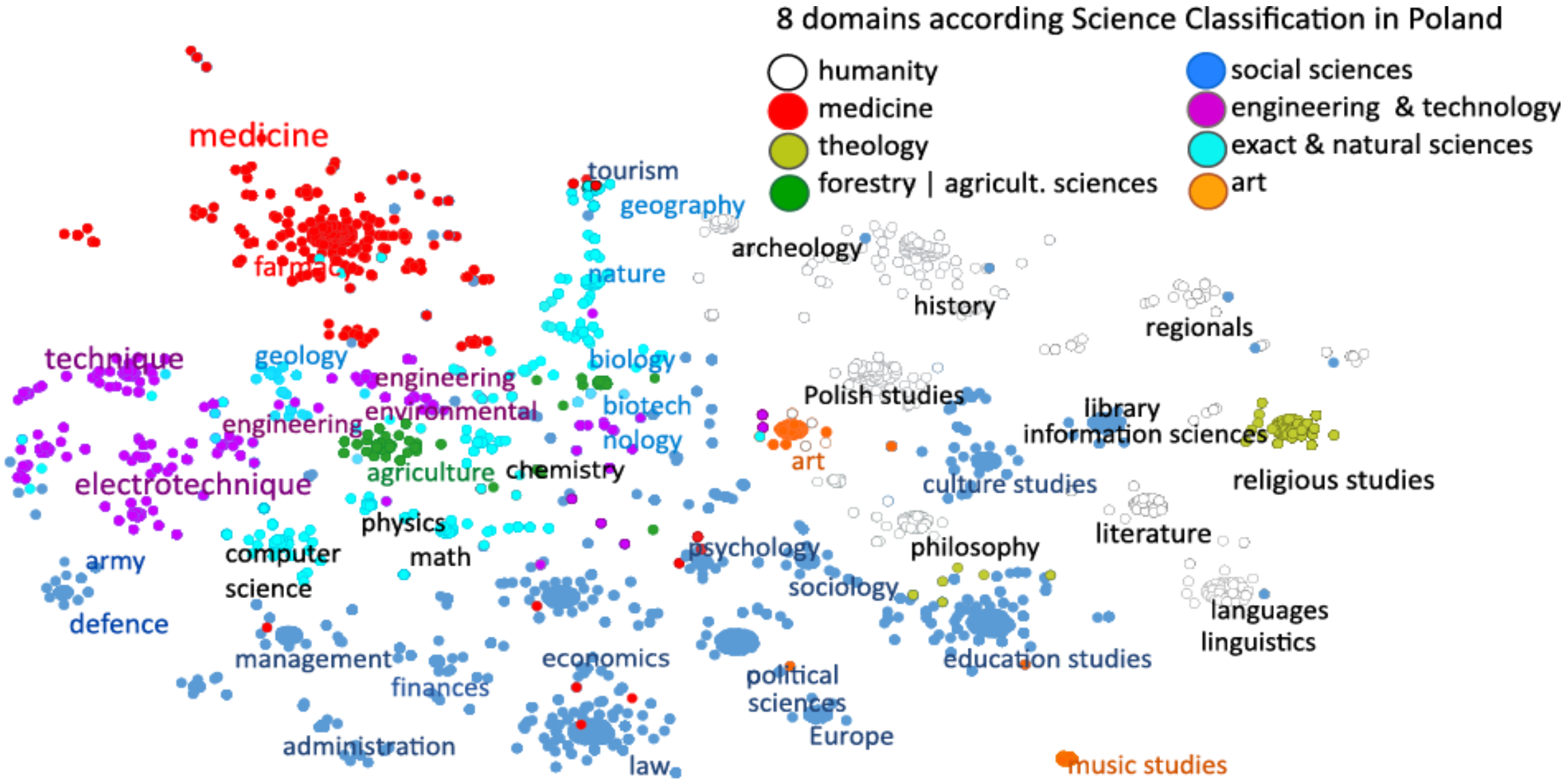
**Porównałam t-SNE względem 5 innych algorytmów**

# Charakterystyka danych

- Lista Polskich czasopism Arianta (sprzed 2018 r.)
- N = 4338
- $L_{disc}$ = 171
- S: 4338 x 171



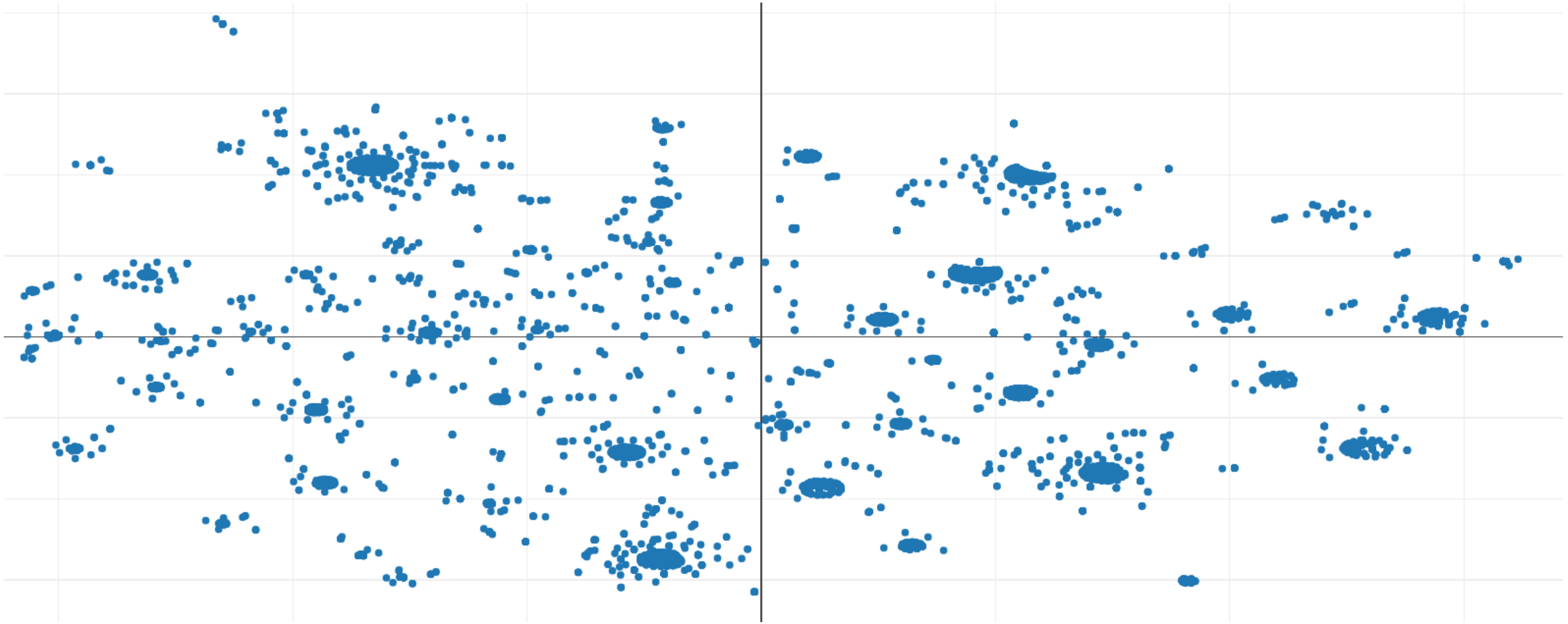

V. Osińska. A qualitative–quantitative study of science mapping by different algorithms: The Polish journals landscape. *Journal of Information Science* 1–14, 2020.

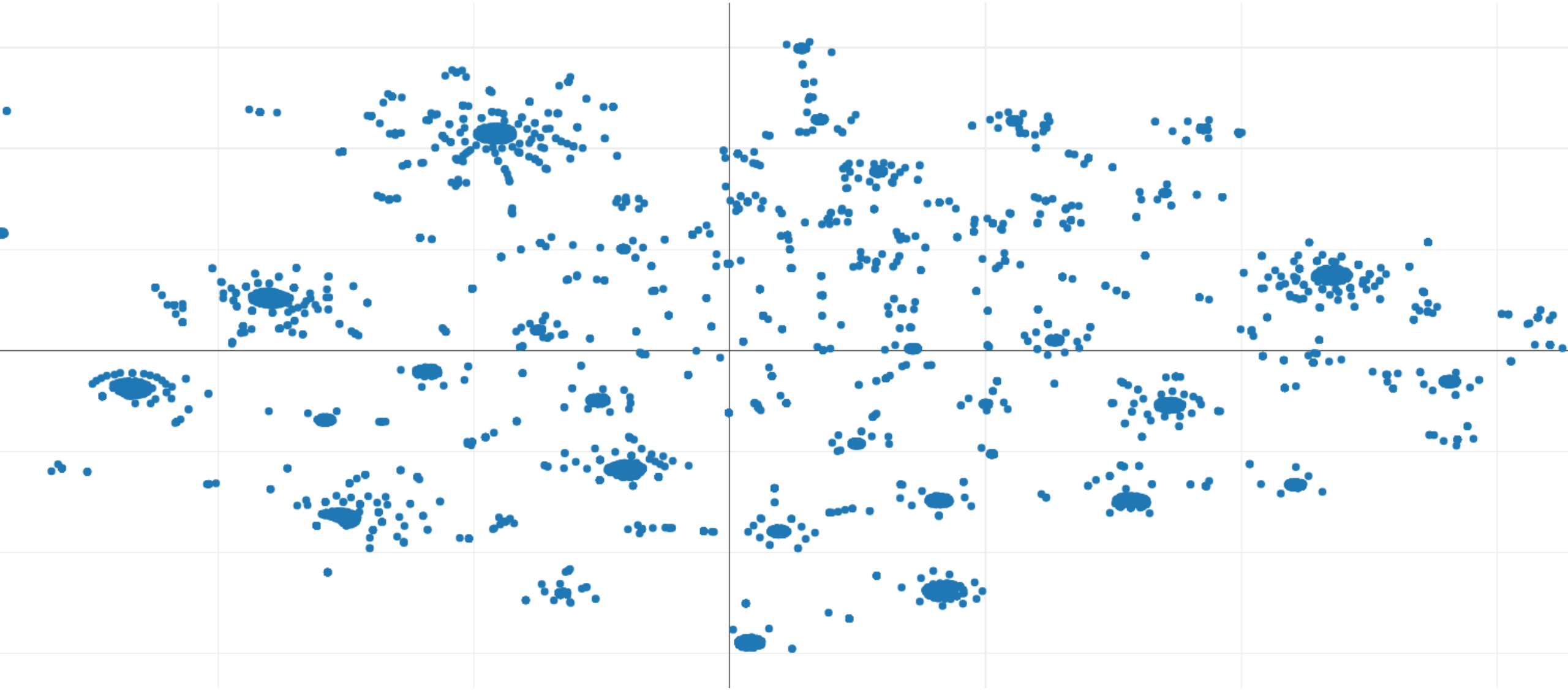



V. Osińska. A qualitative–quantitative study of science mapping by different algorithms: The Polish journals landscape. *Journal of Information Science* 1–14, 2020.

# Mapy interaktywne

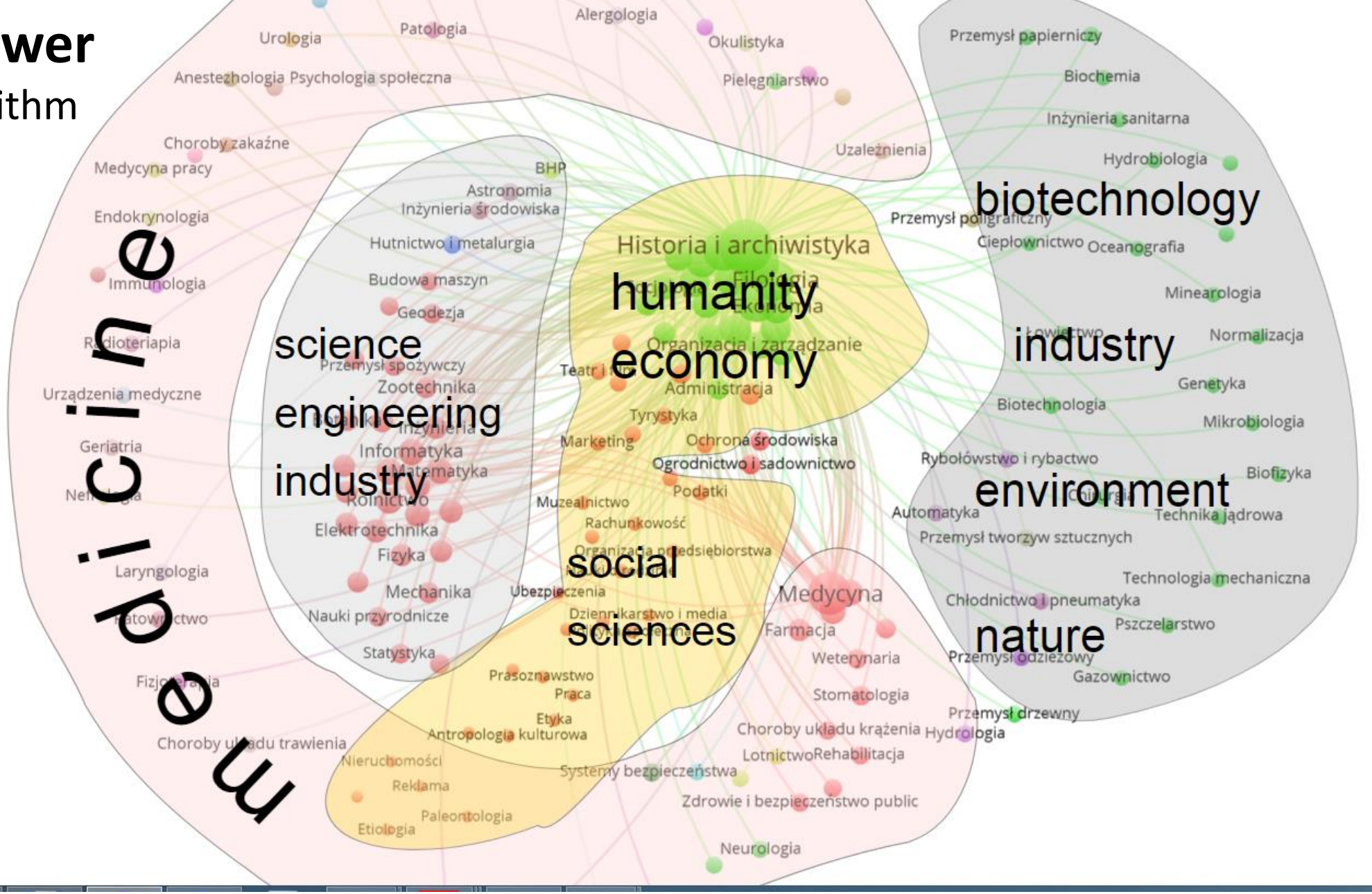
- http://www.wizualizacjainformacji.pl/journals_maps
- Iteracja mapy 1
- Iteracja mapy 2





# VOSViewer

VOS algorithm



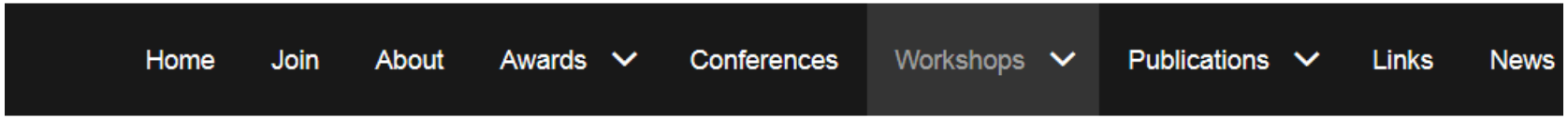
# 16th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

**16-20 October, 2017**

**WUHAN UNIVERSITY • WUHAN • CHINA**

# Problem reprodukowalności wyników





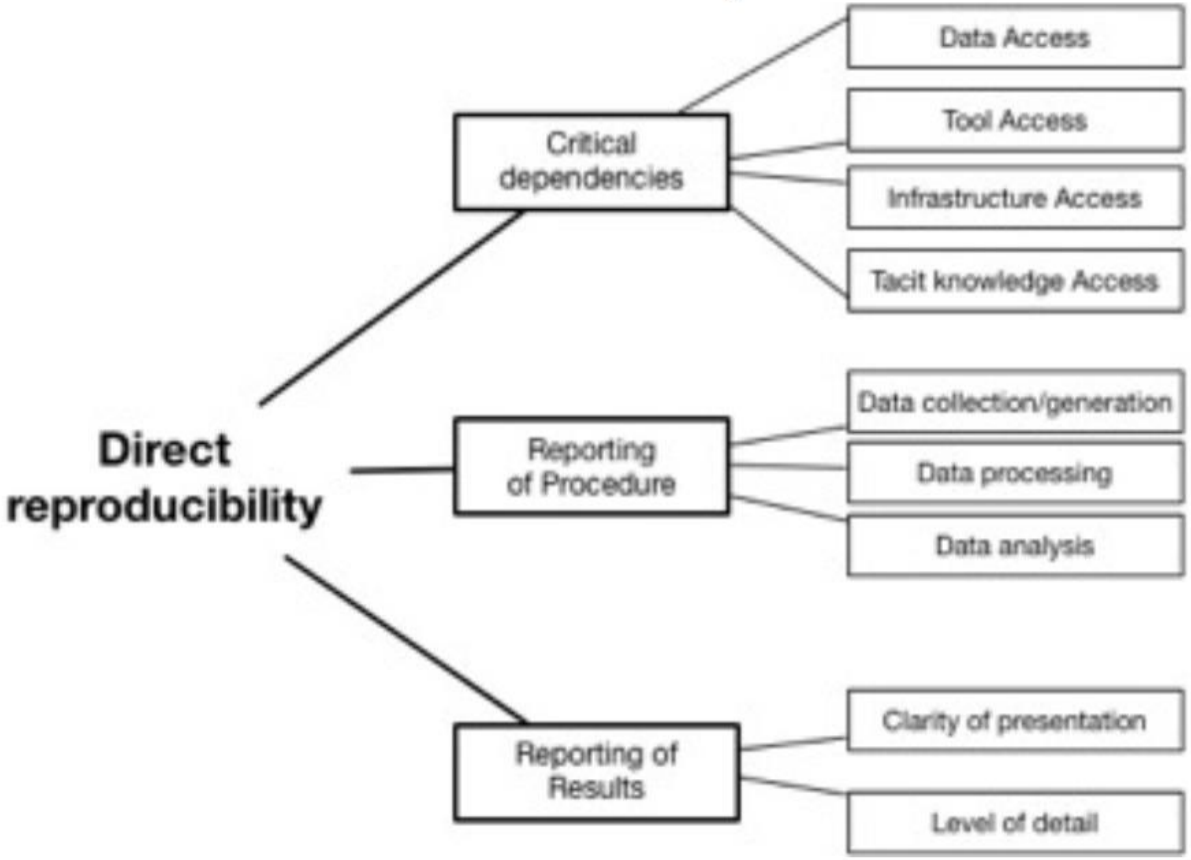


Workshop "Reproducible Scientometrics Research: Open Data, Code, and Education" (ISSI2017)

# Reprodukowalność: 2 definicje

- **Direct reproducibility** is located at the ‘greatest similarity’ end of the spectrum where the **same data, tools and methods** are used to reproduce and verify a study with the expectation of obtaining identical or **very similar empirical** results…
- **Conceptual reproducibility** is located at the other end of the spectrum where a study is reproduced using **different data, tools and methods** with the aim of testing the **robustness** of the fundamental knowledge…

Figure 1: Taxonomy to identify potential threats to direct reproducibility of a published scientometric study.




Ludo Waltman et al. (2018). Exploration of reproducibility issues in scientometric researchPart 1: Direct reproducibility. Proceedings of Science, Technology and Innovations indicators in Transition STI 2018, Leiden, The Netherlands.

# Porównywanie metod

Niepewność może dotyczyć:

- dokładności i kompletności wykorzystanych surowych danych
- wiarygodności mapowania przez wybór modelu danych i algorytmu grupowania
- istnienia niewykrytych błędów w kodowaniu,
- interpretacji wyniku ekstrakcji tematów

**Stopień niepewności zależy od postawionego celu**

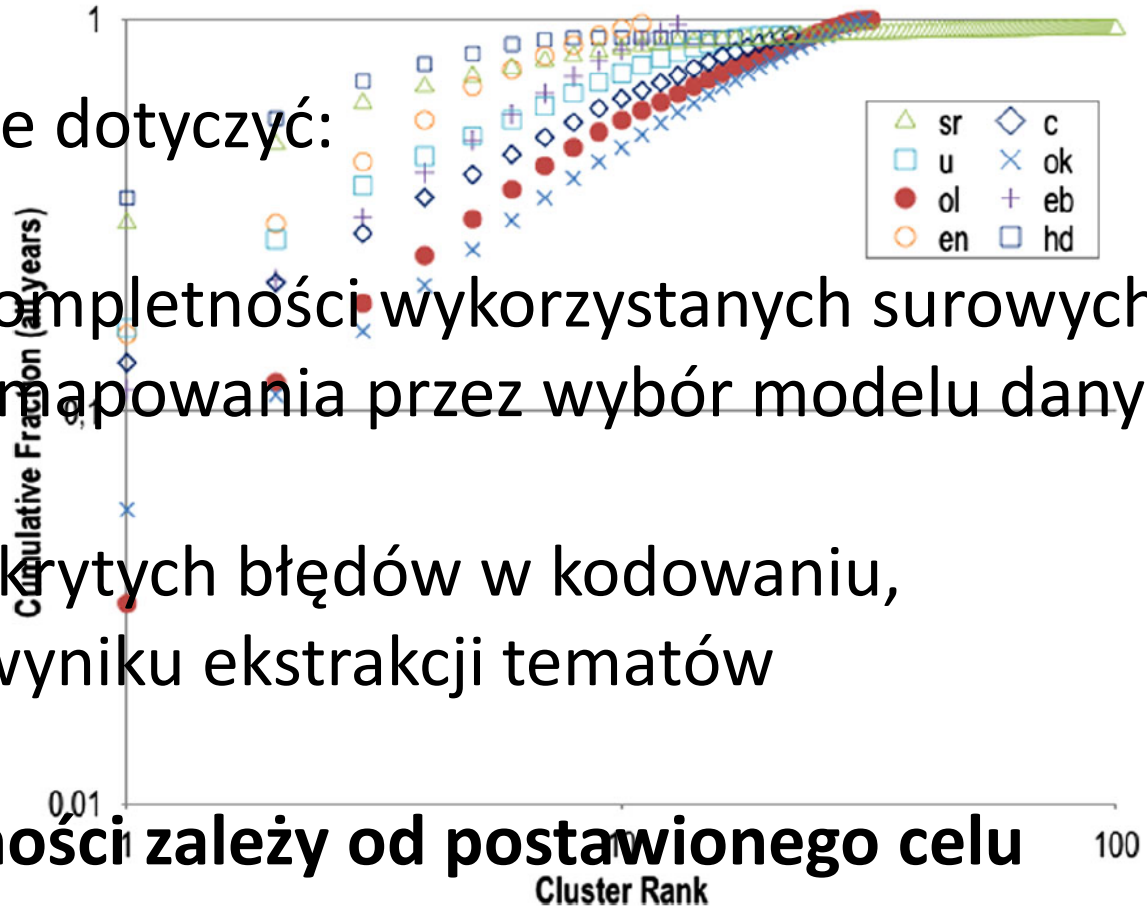



**Fig. 2** Accumulative fractional size distribution of clusters in each solution. The y-axis indicates w

Theresa Velden et all (2017). **Comparison of topic extraction approaches and their results**. Scientometrics, 111:1169–1221. DOI 10.1007/s11192-017-2306-1.

# Pytania wśród naukometrów

1. W jakim stopniu struktury wyłaniające się z mapowania nauki są rzeczywiście **reprezentacją tematyczną struktury** w nauce lub

   **artefaktami** wytwarzanymi przez same metody?

1. W jakim stopniu są wyniki identyfikacji tematów określają właściwości struktury wiedzy i jak ich zakres zależy od stosowanych metod?
2. Czy produkujemy więcej niż artefakty?

Jochen Glaser et all (2017). **Same data—different results? Towards a comparative approach to the identification of thematic structures in science**. Scientometrcis, 111:981–998, DOI 10.1007/s11192-017-2296-z.

Google Scholar

# Chaomei Chen

FOLLOW

GET MY OWN PROFILE

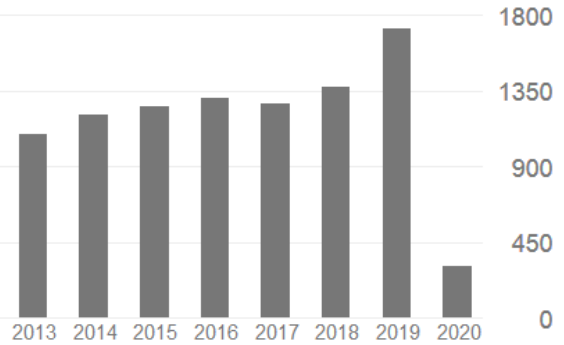
Professor of Informatics, College of Computing and Informatics, Drexel University
Verified email at drexel.edu - Homepage

Information visualization  visual analytics  scientometrics  bibliometrics  scholarly communication

| TITLE | CITED BY | YEAR |
|---|---|---|
| CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature<br>C Chen<br>Journal of the American Society for Information Science and Technology 57 (3 … | 2583 | 2006 |
| Visualizing knowledge domains<br>K Börner, C Chen, KW Boyack<br>Annual review of information science and technology 37 (1), 179-255 | 1299 | 2003 |
| Searching for intellectual turning points: Progressive knowledge domain visualization<br>C Chen<br>Proceedings of the National Academy of Sciences 101 (suppl 1), 5303-5310 | 999 | 2004 |
| Information visualization: Beyond the horizon<br>C Chen<br>Springer Science & Business Media | 698 | 2006 |
| The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis<br>C Chen, F Ibekwe-SanJuan, J Hou<br>Journal of the American Society for information Science and Technology 61 (7 | 688 | 2010 |

Cited by  VIEW ALL

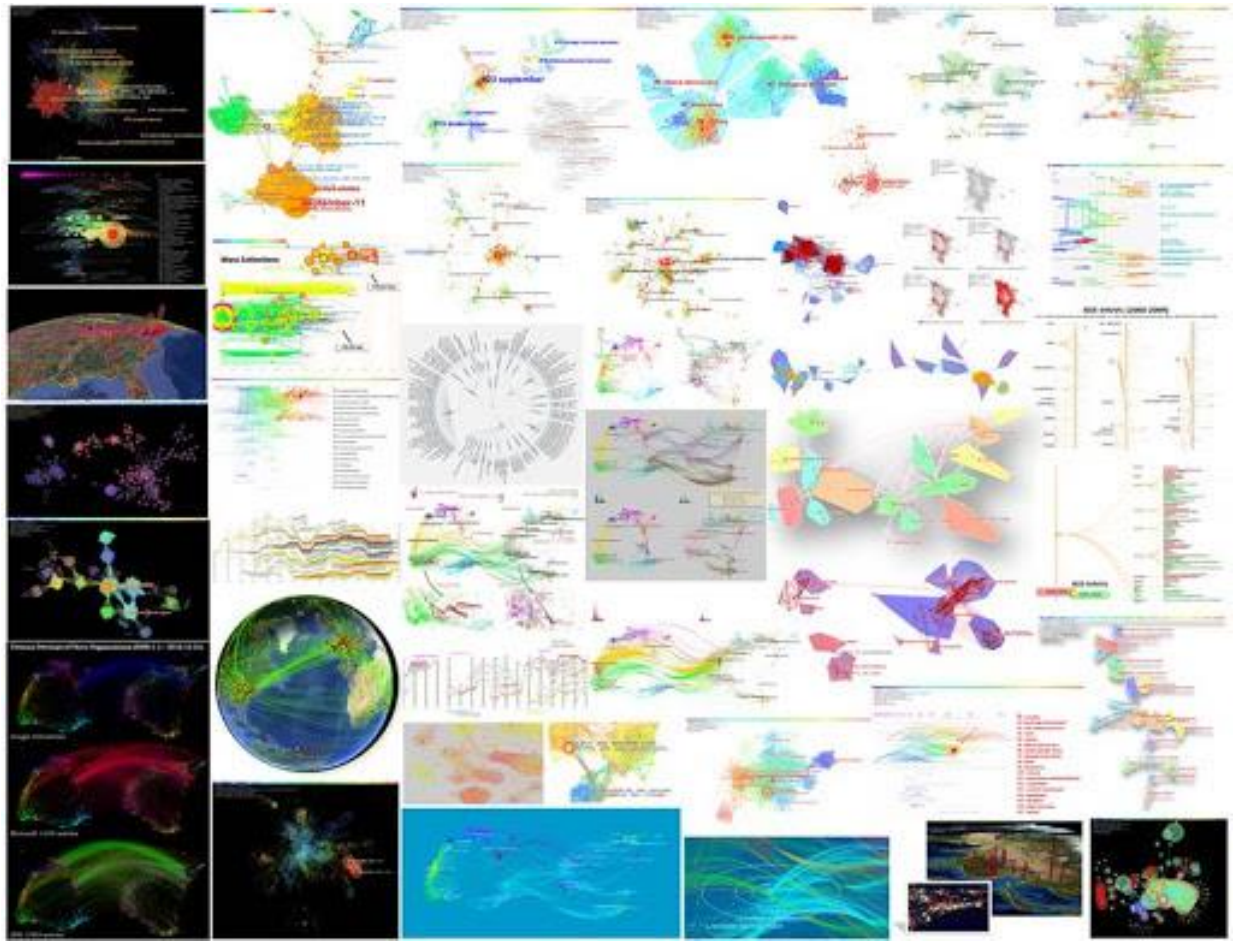
| | All | Since 2015 |
|---|---|---|
| Citations | 17151 | 7263 |
| h-index | 54 | 38 |
| i10-index | 141 | 92 |

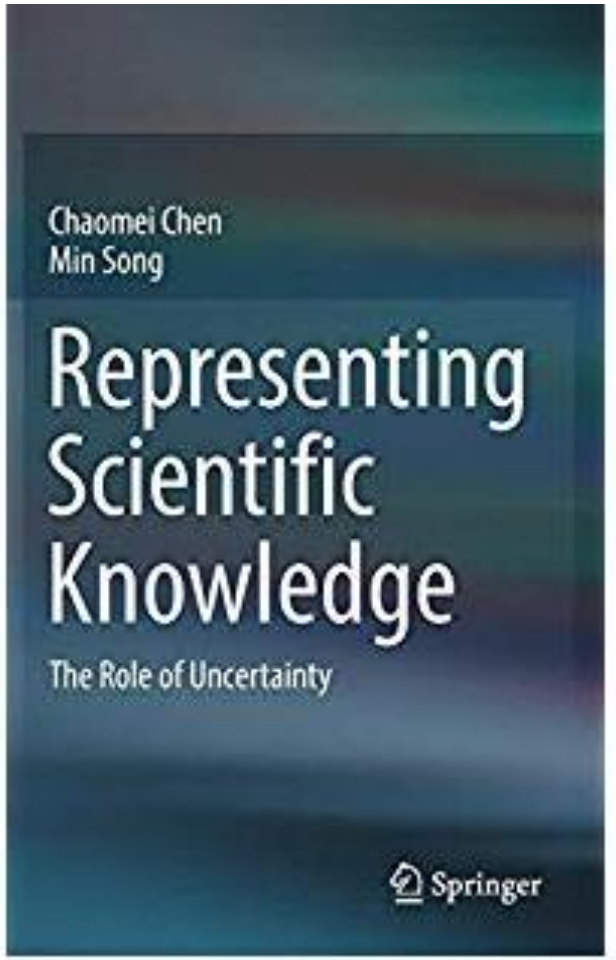



Co-authors  VIEW ALL

Katy Borner

# CiteSpace – aplikacja Ch. Chena





# Wnioski o wizualizacji

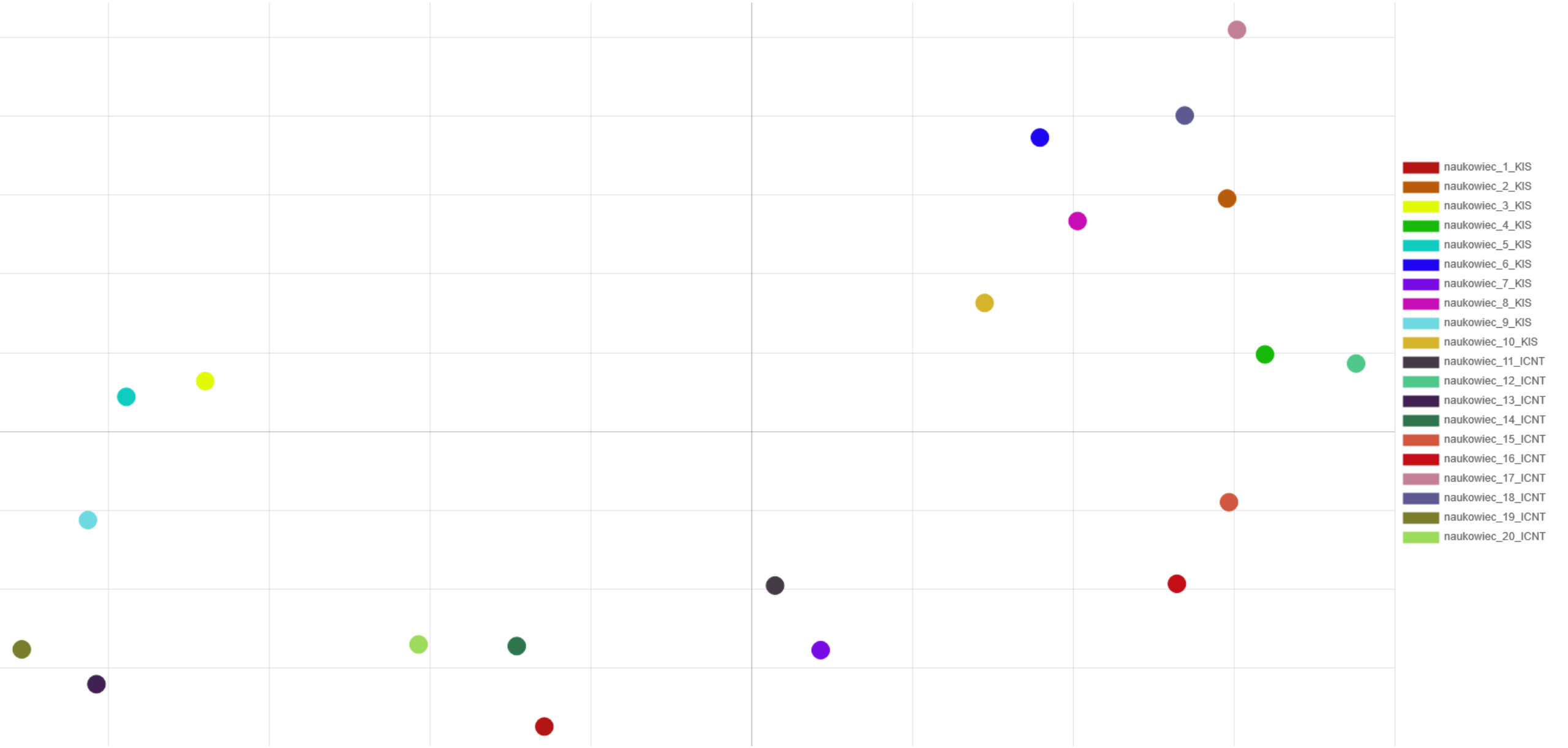
1. Niepewność jest wpisana w studia nad mapowaniem nauki
2. Artefakty generują niepewność replikacji
3. Maksymalne odległości definiują topologię mapy
4. Czy zawsze i w jakiej postaci istnieje oś symetrii (software-hardware)?
5. Należy wykorzystać jest widzenie makroskopowe (wg analogii bigdata)

# Dyskusja

- Niepewność: sąsiedztwo lokalne a globalne
- Definicja sąsiedztwa globalnego (oś symetrii)
- Jednostki na mapach nauki nieistotne (?)

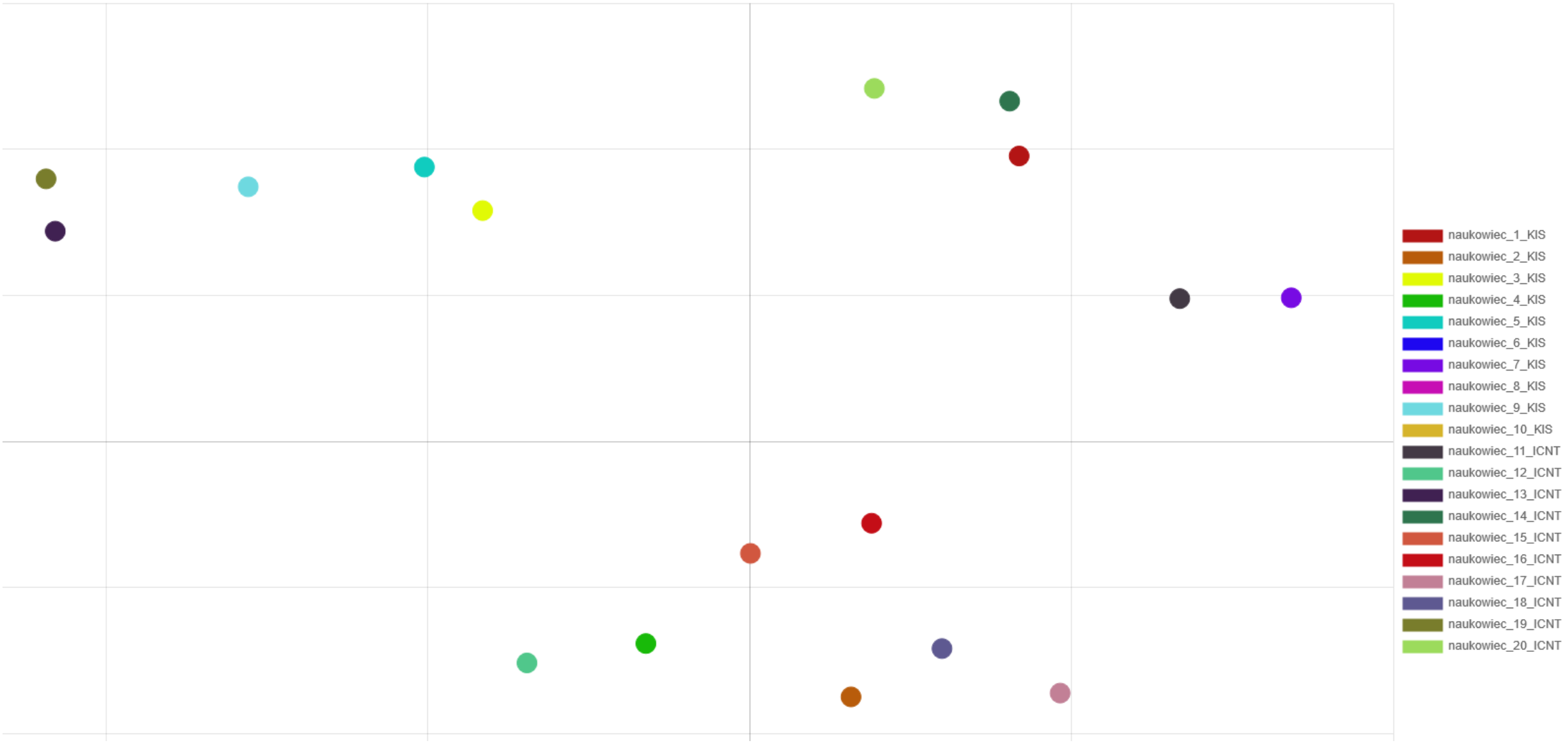
# Porównać rozkłady t-SNE

- https://sciencefuzz.azurewebsites.net/Calculator


naukowiec_1_KIS
naukowiec_2_KIS
naukowiec_3_KIS
naukowiec_4_KIS
naukowiec_5_KIS
naukowiec_6_KIS
naukowiec_7_KIS
naukowiec_8_KIS
naukowiec_9_KIS
naukowiec_10_KIS
naukowiec_11_ICNT
naukowiec_12_ICNT
naukowiec_13_ICNT
naukowiec_14_ICNT
naukowiec_15_ICNT
naukowiec_16_ICNT
naukowiec_17_ICNT
naukowiec_18_ICNT
naukowiec_19_ICNT
naukowiec_20_ICNT

ScienceFuzz

KALKULATOR T-SNE K-MEANS



# Informacje tekstowa a niepewność ☺

Seminaria odbywają się w:

- Ostatni **wtorek** miesiąca
- **3** marca **2019**

Dedykuję wykład …

Dziękuję za uwagę ☺