



Lecture on:

## Parallelised analysis of large-scale bibliometric and text data (with Dask in Python, DuckDB and DBeaver in SQL)

Aliakbar Akbaritabar<sup>1</sup>

<sup>1</sup> Max Planck Institute for Demographic Research (MPIDR)  
Akbaritabar@demogr.mpg.de

Please tweet with hashtag  
#BiblioDemography,  
a tribute to James W. Vaupel (1945-2022).  
Thanks Ilya and Jonas for bringing up Jim's  
labeling idea!



AGENDA

- 1. What is bibliometric data?
- 2. Parallelised analysis of bibliometric data
- 3. Introduction to text analysis
- 4. Example uses and hands-on coding
- 5. Limitations of bibliometric data and final Q&A

# Materials publicly available



## Related links:

- Interview and references on using bibliometric data for demographic research:  
[https://www.demogr.mpg.de/en/news\\_events\\_6123/news\\_press\\_releases\\_4630/news/how\\_to\\_use\\_bibliometric\\_data\\_for\\_demographic\\_research\\_10784](https://www.demogr.mpg.de/en/news_events_6123/news_press_releases_4630/news/how_to_use_bibliometric_data_for_demographic_research_10784)

The screenshot shows the GitHub repository page for 'akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data'. The repository is public and contains a README.md file. The README describes the materials for a lecture on 'Parallelised analysis of large-scale bibliometric and text data (with Dask in Python, DuckDB and DBeaver in SQL)'. The repository includes a file tree with folders for code, data, presentations, outputs, and images, as well as a .gitignore file, a LICENSE file, and a README.md file. The README also lists the speaker, Aliakbar Akbaritabar, and provides a link to his GitHub profile and email address. The repository has 0 stars, 1 watching, and 0 forks.

**Repository:** akbaritabar / Parallelised-analysis-of-large-scale-bibliometric-and-text-data (Public)

**Files:**

- 0\_code (Minor edit in one URL, 10 minutes ago)
- 1\_data (Lecture materials added, 37 minutes ago)
- 2\_presentations (Lecture materials added, 37 minutes ago)
- 98\_outputs (Lecture materials added, 37 minutes ago)
- 99\_images (Lecture materials added, 37 minutes ago)
- .gitignore (Lecture materials added, 37 minutes ago)
- LICENSE (Lecture materials added, 37 minutes ago)
- Readme.md (Lecture materials added, 37 minutes ago)

**Readme.md:**

Materials for the lecture on "Parallelised analysis of large-scale bibliometric and text data (with Dask in Python, DuckDB and DBeaver in SQL)"

Speaker:

- Aliakbar Akbaritabar, research scientist at the Max Planck Institute for Demographic Research (MPIDR), GitHub: <https://github.com/akbaritabar>, ([akbaritabar@demogr.mpg.de](mailto:akbaritabar@demogr.mpg.de))

Instructions for participants

**About:** Materials for the lecture on "Parallelised analysis of large-scale bibliometric and text data (with Dask in Python, DuckDB and DBeaver in SQL)"

**Releases:** No releases published. [Create a new release](#)

**Packages:** No packages published. [Publish your first package](#)

**Languages:** Python 100.0%

**Materials:** <https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data>

# What is bibliometric data?



an open access  journal



Citation: Akbaritabar, A. (2021). A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. *Quantitative Science Studies*, 2(2), 753–777. [https://doi.org/10.1162/qss\\_a\\_00131](https://doi.org/10.1162/qss_a_00131)

DOI:  
[https://doi.org/10.1162/qss\\_a\\_00131](https://doi.org/10.1162/qss_a_00131)

Peer Review:  
[https://publons.com/publon/10.1162/qss\\_a\\_00131](https://publons.com/publon/10.1162/qss_a_00131)

Received: 2 September 2020  
Accepted: 2 April 2021

Corresponding Author:  
Aliakbar Akbaritabar  
[akbaritabar@demogr.mpg.de](mailto:akbaritabar@demogr.mpg.de)

Handling Editor:  
Ludo Waltman

## RESEARCH ARTICLE

# A quantitative view of the structure of institutional scientific collaborations using the example of Berlin

Aliakbar Akbaritabar<sup>1,2</sup> 

<sup>1</sup>Max Planck Institute for Demographic Research (MPIDR),  
Laboratory of Digital and Computational Demography, Rostock, Germany

<sup>2</sup>German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

**Keywords:** Berlin, Berlin University Alliance, bipartite community detection, coauthorship network analysis, disambiguation, internationalization


## ABSTRACT


This paper examines the structure of scientific collaborations in Berlin as a specific case with a unique history of division and reunification. It aims to identify strategic organizational coalitions in a context with high sectoral diversity. We use publications data with at least one organization located in Berlin from 1996–2017 and their collaborators worldwide. We further investigate four members of the Berlin University Alliance (BUA), as a formerly established coalition in the region, through their self-represented research profiles compared with empirical results. Using a bipartite network modeling framework, we move beyond the uncontested trend towards team science and increasing internationalization. Our results show that BUA members shape the structure of scientific collaborations in the region. However, they are not collaborating cohesively in all fields and there are many smaller scientific actors involved in more internationalized collaborations in the region. Larger divides exist in some fields. Only Medical and Health Sciences have cohesive intraregional collaborations, which signals the success of the regional cooperation established in 2003. We explain possible underlying factors shaping the intraregional groupings and potential implications for regions worldwide. A major methodological contribution of this paper is evaluating the coverage and accuracy of different organization name disambiguation techniques.


# What is bibliometric data?

## Metadata of scientific publications

- Authors' name
- Affiliation addresses
- Publication type (article, review, conference proceedings)
- Subject classification (often based on journal assignment)
- Title, abstract and keywords (for text analysis)
- Reference list (for citation analysis)
- Acknowledgements (for funding information)
- Open Access information



an open access  journal



Citation: Akbaritabar, A. (2021). A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. *Quantitative Science Studies*, 2(2), 753–777. [https://doi.org/10.1162/qss\\_a\\_00131](https://doi.org/10.1162/qss_a_00131)

DOI: [https://doi.org/10.1162/qss\\_a\\_00131](https://doi.org/10.1162/qss_a_00131)

Peer Review: [https://publons.com/publon/10.1162/qss\\_a\\_00131](https://publons.com/publon/10.1162/qss_a_00131)


Received: 2 September 2020  
Accepted: 2 April 2021

Corresponding Author:  
Aliakbar Akbaritabar  
[akbaritabar@demogr.mpg.de](mailto:akbaritabar@demogr.mpg.de)

Handling Editor:  
Ludo Waltman

RESEARCH ARTICLE

## A quantitative view of the structure of institutional scientific collaborations using the example of Berlin

Aliakbar Akbaritabar<sup>1,2</sup> 

<sup>1</sup>Max Planck Institute for Demographic Research (MPIDR),  
Laboratory of Digital and Computational Demography, Rostock, Germany  
<sup>2</sup>German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

**Keywords:** Berlin, Berlin University Alliance, bipartite community detection, coauthorship network analysis, disambiguation, internationalization

---

**ABSTRACT**

This paper examines the structure of scientific collaborations in Berlin as a specific case with a unique history of division and reunification. It aims to identify strategic organizational coalitions in a context with high sectoral diversity. We use publications data with at least one organization located in Berlin from 1996–2017 and their collaborators worldwide. We further investigate four members of the Berlin University Alliance (BUA), as a formerly established coalition in the region, through their self-represented research profiles compared with empirical results. Using a bipartite network modeling framework, we move beyond the uncontested trend towards team science and increasing internationalization. Our results show that BUA members shape the structure of scientific collaborations in the region. However, they are not collaborating cohesively in all fields and there are many smaller scientific actors involved in more internationalized collaborations in the region. Larger divides exist in some fields. Only Medical and Health Sciences have cohesive intraregional collaborations, which signals the success of the regional cooperation established in 2003. We explain possible underlying factors shaping the intraregional groupings and potential implications for regions worldwide. A major methodological contribution of this paper is evaluating the coverage and accuracy of different organization name disambiguation techniques.



## Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

## Section on more advanced data processing using parallelization





# Introductory course to R or Python?!

Please check this repository by **Vincent Traag** and others, for an introductory course and code to familiarize yourself with **Python**:

<https://github.com/vtraag/intro-python>

Or this one by Data Carpentry:

<https://datacarpentry.org/python-ecology-lesson/>

This course by Data Carpentry provides basics in **R**:

<https://datacarpentry.org/R-genomics/index.html>



# Parallelised analysis of large-scale bibliometric data (with Dask in Python, DuckDB and DBeaver in SQL); Using example of ORCID 2019 XML files



Materials under code, data and output folders of repository, **Python** users see:

[https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/0\\_code/03\\_parallelization\\_with\\_dask\\_duckdb\\_dbeaver.md](https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/0_code/03_parallelization_with_dask_duckdb_dbeaver.md)

Recorded **Video** of advanced tutorial:

<https://youtu.be/pYDVrBcluYI>

- **00:00 - 02:45**; Introduction
- **2:45 - 11:04**; Requirements and installation (skip this if you have already installed everything based on instructions in the ReadMe on repository)
- **11:05 - 38:00**; Steps in using Dask/Python and DuckDB/DBeaver and results

## Section on text analysis

Code:

[https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/0\\_code/05\\_text\\_analysis\\_exact\\_and\\_anchor\\_words.py](https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/0_code/05_text_analysis_exact_and_anchor_words.py)

[https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/0\\_code/07\\_text\\_analysis\\_noun\\_phrase\\_clauses.py](https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/0_code/07_text_analysis_noun_phrase_clauses.py)





Clarivate's Web of Science<sup>1</sup> (WOS), 1990-(end of) 2018, “article” and “review” publications

**Corpus construction:** search the lowercased “**plasticity**” in title, abstract and keywords

To further limit the usages of this concept to the areas closer (but not limited to) neurosciences, we follow **three** strategies

- 1) **Subject categories** by WOS (excluding less relevant fields, but they are covered in more precise strategies)
- 2) Specific **keyword combinations** (word pairs):
  - 'adult neurogenesis', 'synaptic plasticity', 'cortical plasticity', 'cortical map plasticity', 'receptive field plasticity', 'heterosynaptic plasticity', 'hebbian plasticity', 'structural plasticity', 'neuronal responses', 'plasticity of neuronal responses', 'bidirectional plasticity', 'functional plasticity', 'developmental plasticity', 'critical period plasticity', 'adult brain plasticity', 'ocular dominance plasticity', 'neuronanatomical plasticity', 'cross modal plasticity'

<sup>1</sup>Provided by the German competence center for bibliometrics Data is obtained from Kompetenzzentrum Bibliometrie, which is funded by the Federal Ministry for Education and Research (BMBF), Germany with grant number 16WIK2101A. <http://bibliometrie.info/>

Two previous approaches could be driven by

- a) the way WOS defines the subject categories (*journal based*)
- b) the keyword combinations highlighted previously in the literature could be **driven by the area of focus (niche naming/labeling of similar things, i.e., academic dialects)**

3) **A parallel attempt**, we try to apply a more objective criteria using text and content analysis (found words).

- **nltk** library in Python (Bird, Loper and Klein, 2009), first tokenize the words used in the title and abstract of publications. We decided to tokenize the text as our first step since we need the punctuation and sentence structure in determining the words. We then remove punctuations from these tokenized text, and lower case all the text elements. We then use plasticity as an “anchor word” and see if it is used once or more in the title and abstract. We identify which are the words pairing with it as being used before it.
- There were specific cases where the word used before plasticity was not a proper adjective or a scientific term, we excluded those from the following visualizations.

# Example 1 of text analysis using exact word pairs



PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
35583791345	PLOS COMPUTATIONAL BIOLOGY	000463877900036	2019	Article	BIOCHEMICAL RESEARCH METHODS	NATURAL SCIENCES

Exact word pairs



Title

Synaptic plasticity onto inhibitory neurons as a mechanism for ocular dominance plasticity<sup>1</sup>

Abstract

Ocular dominance plasticity is a well-documented phenomenon allowing us to study properties of cortical maturation. Understanding this maturation might be an important step towards unravelling how cortical circuits function. However, it is still not fully understood which mechanisms are responsible for the opening and closing of the critical period for ocular dominance and how changes in cortical responsiveness arise after visual deprivation. In this article, we present a theory of ocular dominance plasticity. Following recent experimental work, we propose a framework where a reduction in inhibition is necessary for ocular dominance plasticity in both juvenile and adult animals. In this framework, two ingredients are crucial to observe ocular dominance shifts: a sufficient level of inhibition as well as excitatory-to-inhibitory synaptic plasticity. In our model, the former is responsible for the opening of the critical period, while the latter limits the plasticity in adult animals. Finally, we also provide a possible explanation for the variability in ocular dominance shifts observed in individual neurons and for the counter-intuitive shifts towards the closed eye. Author summary During the development of the brain, visual cortex has a period of increased plasticity. Closing one eye for multiple days during this period can have a profound and life-long impact on neuronal responses. A well-established hypothesis is that the absolute level of inhibition regulates this period. In light of recent experimental results, we suggest an alternative theory. We propose that, in addition to the level of inhibition, synaptic plasticity onto inhibitory neurons is just as crucial. We propose a model which explains many observed phenomena into one single framework. Unlike theories considering only the level of inhibition, we can account for both the onset as well as the closure of this period. Furthermore, we also provide an explanation for the small fraction of neurons that show counter-intuitive behaviour and provide some testable predictions.

<sup>1</sup> I had a typo in ocular in my exact words definition, writing it as “ocular”, and this very accident shows how this method is prone to accidental exclusion (later my coauthor shared that “ocular” is correct!)

# Example 1 of text analysis using anchors



PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
----------	-------------	--------	---------	---------	----------------	------------------

35583791345	PLOS COMPUTATIONAL BIOLOGY	000463877900036	2019	Article	BIOCHEMICAL RESEARCH METHODS	NATURAL SCIENCES
-------------	----------------------------------	-----------------	------	---------	------------------------------------	------------------

Word used before    Anchor    Word used after

↓                      ↓                      ↓

**Synaptic plasticity onto inhibitory neurons as a mechanism for ocular dominance plasticity**

Title

Abstract

Ocular **dominance plasticity** is a well-documented phenomenon allowing us to study properties of cortical maturation. Understanding this maturation might be an important step towards unravelling how cortical circuits function. However, it is still not fully understood which mechanisms are responsible for the opening and closing of the critical period for ocular dominance and how changes in cortical responsiveness arise after visual deprivation. In this article, we present a theory of ocular **dominance plasticity**. Following recent experimental work, we propose a framework where a reduction in inhibition is necessary for ocular **dominance plasticity in** both juvenile and adult animals. In this framework, two ingredients are crucial to observe ocular dominance shifts: a sufficient level of inhibition as well as excitatory-to-inhibitory **synaptic plasticity**. In our model, the former is responsible for the opening of the critical period, while the latter limits **the plasticity in** adult animals. Finally, we also provide a possible explanation for the variability in ocular dominance shifts observed in individual neurons and for the counter-intuitive shifts towards the closed eye. Author summary During the development of the brain, visual cortex has a period of **increased plasticity**. Closing one eye for multiple days during this period can have a profound and life-long impact on neuronal responses. A well-established hypothesis is that the absolute level of inhibition regulates this period. In light of recent experimental results, we suggest an alternative theory. We propose that, in addition to the level of inhibition, **synaptic plasticity onto** inhibitory neurons is just as crucial. We propose a model which explains many observed phenomena into one single framework. Unlike theories considering only the level of inhibition, we can account for both the onset as well as the closure of this period. Furthermore, we also provide an explanation for the small fraction of neurons that show counter-intuitive behaviour and provide some testable predictions.

# Example 1 of text analysis using exact Noun-Phrase-Clauses



Relevant noun-phrase  
clauses found

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
35583791345	PLOS COMPUTATIONAL BIOLOGY	000463877900036	2019	Article	BIOCHEMICAL RESEARCH METHODS	NATURAL SCIENCES

Title

Synaptic plasticity onto inhibitory neurons as a mechanism for ocular dominance plasticity<sup>1</sup>

Abstract

Ocular dominance plasticity is a well-documented phenomenon allowing us to study properties of cortical maturation. Understanding this maturation might be an important step towards unravelling how cortical circuits function. However, it is still not fully understood which mechanisms are responsible for the opening and closing of the critical period for ocular dominance and how changes in cortical responsiveness arise after visual deprivation. In this article, we present a theory of ocular dominance plasticity. Following recent experimental work, we propose a framework where a reduction in inhibition is necessary for ocular dominance plasticity in both juvenile and adult animals. In this framework, two ingredients are crucial to observe ocular dominance shifts: a sufficient level of inhibition as well as excitatory-to-inhibitory synaptic plasticity. In our model, the former is responsible for the opening of the critical period, while the latter limits the plasticity in adult animals. Finally, we also provide a possible explanation for the variability in ocular dominance shifts observed in individual neurons and for the counter-intuitive shifts towards the closed eye. Author summary During the development of the brain, visual cortex has a period of increased plasticity. Closing one eye for multiple days during this period can have a profound and life-long impact on neuronal responses. A well-established hypothesis is that the absolute level of inhibition regulates this period. In light of recent experimental results, we suggest an alternative theory. We propose that, in addition to the level of inhibition, synaptic plasticity onto inhibitory neurons is just as crucial. We propose a model which explains many observed phenomena into one single framework. Unlike theories considering only the level of inhibition, we can account for both the onset as well as the closure of this period. Furthermore, we also provide an explanation for the small fraction of neurons that show counter-intuitive behaviour and provide some testable predictions.

<sup>1</sup> I had a typo in ocular in my exact words definition, writing it as “ocular”, and this very accident shows how this method is prone to accidental exclusion (later my coauthor shared that “ocular” is correct!)



# Example 2 of text analysis using exact word pairs



PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
----------	-------------	--------	---------	---------	----------------	------------------

Exact word pairs  
(None identified)

6980386	NEUROLOGICAL SCIENCES	000372330000006	2016	Article	NEUROSCIENCES	MEDICAL AND HEALTH SCIENCES
---------	--------------------------	-----------------	------	---------	---------------	--------------------------------

Title

Biological factors and age-dependence of primary motor cortex experimental plasticity

Abstract

To evaluate whether the age-dependence of brain plasticity correlates with the levels of proteins involved in hormone and brain functions we executed a paired associative stimulation (PAS) protocol and blood tests. We measured the PAS-induced plasticity in the primary motor cortex. Blood levels of the brain-derived neurotrophic factor (BDNF), estradiol, the insulin-like growth factor (IGF)-1, the insulin-like growth factor binding protein (IGFBP)-3, progesterone, sex hormone-binding globulin (SHBG), testosterone, and the transforming growth factor beta 1 (TGF-beta 1) were determined in 15 healthy men and 20 healthy women. We observed an age-related reduction of PAS-induced plasticity in females that it is not present in males. In females, PAS-induced plasticity displayed a correlation with testosterone ( $p = 0.006$ ) that became a trend after the adjustment for the age effect ( $p = 0.078$ ). In males, IGF-1 showed a nominally significant correlation with the PAS-induced plasticity ( $p = 0.043$ ). In conclusion, we observed that hormone blood levels (testosterone in females and IGF-1 in males) may be involved in the age-dependence of brain plasticity.

## Example 2 of text analysis using anchors



PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
----------	-------------	--------	---------	---------	----------------	------------------

6980386	NEUROLOGICAL SCIENCES	000372330000006	2016	Article	NEUROSCIENCES	MEDICAL AND HEALTH SCIENCES
---------	-----------------------	-----------------	------	---------	---------------	-----------------------------

Title

Biological factors and age-dependence of primary motor cortex **experimental** **plasticity**

Word used before

Anchor

Word used after



Abstract

To evaluate whether the age-dependence of **brain** **plasticity** **correlates** with the levels of proteins involved in hormone and brain functions we executed a paired associative stimulation (PAS) protocol and blood tests. We measured the **PAS-induced** **plasticity** **in** the primary motor cortex. Blood levels of the brain-derived neurotrophic factor (BDNF), estradiol, the insulin-like growth factor (IGF)-1, the insulin-like growth factor binding protein (IGFBP)-3, progesterone, sex hormone-binding globulin (SHBG), testosterone, and the transforming growth factor beta 1 (TGF-beta 1) were determined in 15 healthy men and 20 healthy women. We observed an age-related reduction of **PAS-induced** **plasticity** **in** females that it is not present in males. In females, **PAS-induced** **plasticity** **displayed** a correlation with testosterone ( $p = 0.006$ ) that became a trend after the adjustment for the age effect ( $p = 0.078$ ). In males, IGF-1 showed a nominally significant correlation with the **PAS-induced** **plasticity** ( $p = 0.043$ ). In conclusion, we observed that hormone blood levels (testosterone in females and IGF-1 in males) may be involved in the age-dependence of **brain** **plasticity**.

## Example 2 of text analysis using exact Noun-Phrase-Clauses



PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
----------	-------------	--------	---------	---------	----------------	------------------

Relevant noun-phrase  
clauses found

6980386	NEUROLOGICAL SCIENCES	0003723300000006	2016	Article	NEUROSCIENCES	MEDICAL AND HEALTH SCIENCES
---------	--------------------------	------------------	------	---------	---------------	--------------------------------

Title

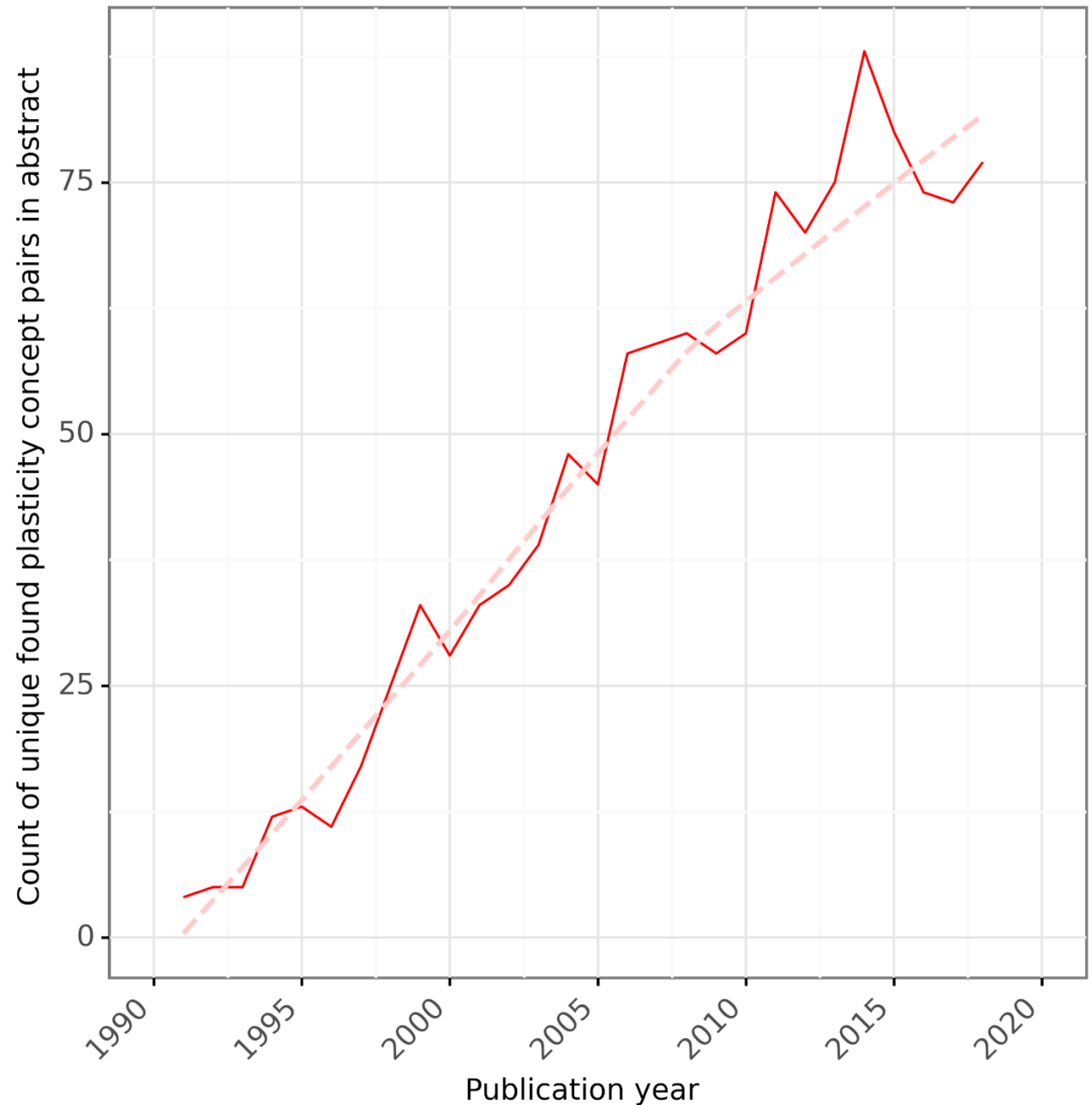
Biological factors and age-dependence of primary motor cortex experimental plasticity

Abstract

To evaluate whether the age-dependence of brain plasticity correlates with the levels of proteins involved in hormone and brain functions we executed a paired associative stimulation (PAS) protocol and blood tests. We measured the PAS-induced plasticity in the primary motor cortex. Blood levels of the brain-derived neurotrophic factor (BDNF), estradiol, the insulin-like growth factor (IGF)-1, the insulin-like growth factor binding protein (IGFBP)-3, progesterone, sex hormone-binding globulin (SHBG), testosterone, and the transforming growth factor beta 1 (TGF-beta 1) were determined in 15 healthy men and 20 healthy women. We observed an age-related reduction of PAS-induced plasticity in females that it is not present in males. In females, PAS-induced plasticity displayed a correlation with testosterone ( $p = 0.006$ ) that became a trend after the adjustment for the age effect ( $p = 0.078$ ). In males, IGF-1 showed a nominally significant correlation with the PAS-induced plasticity ( $p = 0.043$ ). In conclusion, we observed that hormone blood levels (testosterone in females and IGF-1 in males) may be involved in the age-dependence of brain plasticity.

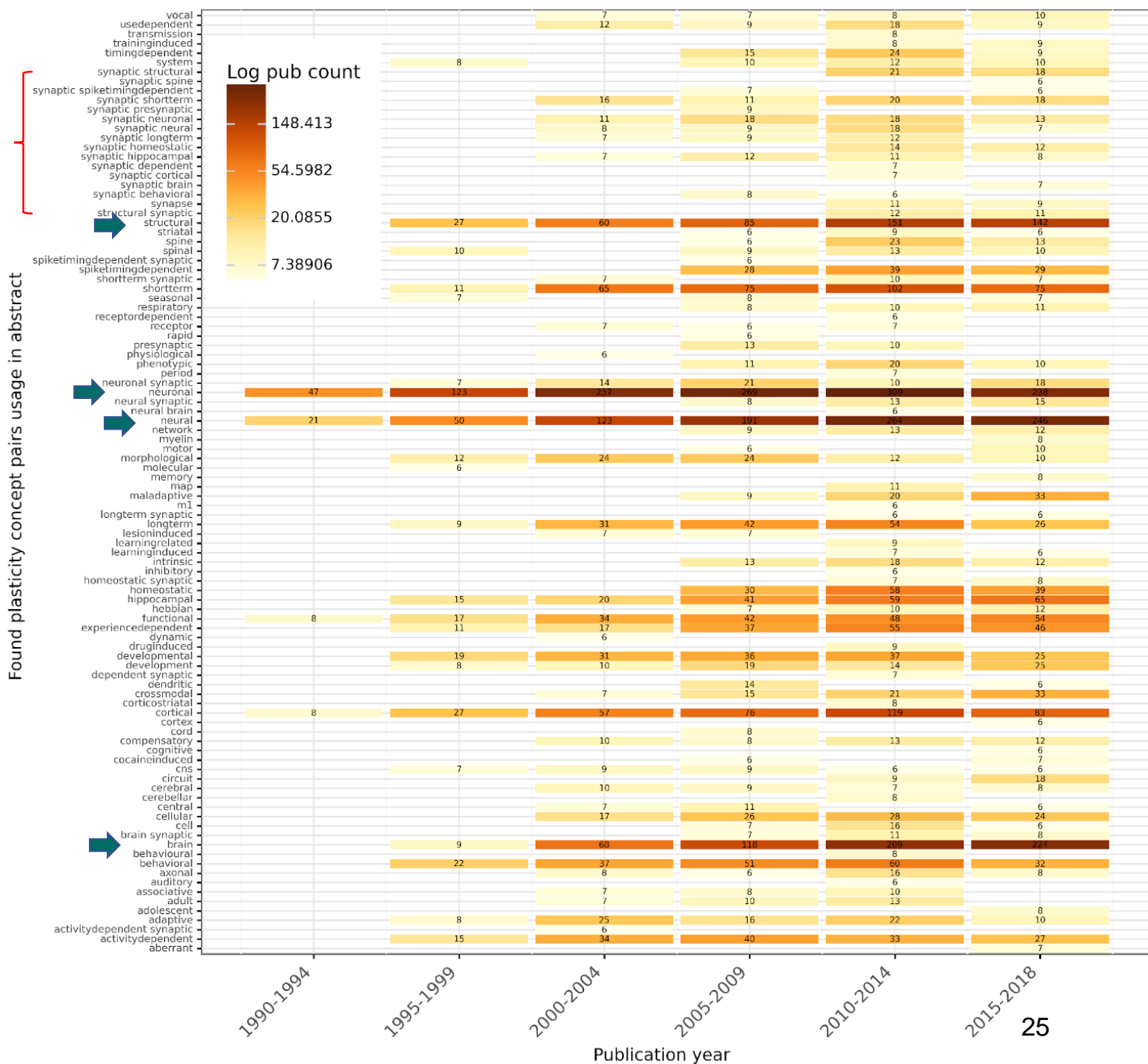
# Plasticity as a concept in the neurosciences, a word pair analysis

- Plasticity is an anchor concept
- Trend of unique word (or words in case multiple words are used in the same abstract) used before plasticity
- Red line shows the count of unique words per year and smooth trend based on mean is shown on light red dashed line.
- The count of words pairing with plasticity has been constantly on the rise since recent years. Year 2014 has the highest count of unique words (88) paired with plasticity.



# Plasticity as a concept in the neurosciences, a word pair analysis

- Synaptic plasticity is the central anchor concept
- Found words used at least 5 times before plasticity as an anchor word and temporal use of them in the abstract of scientific publications in the neurosciences (based on WOS subject categories)
- If two words are used before plasticity in the same abstract, they are presented together on the Figure. We exclude “synaptic” as it dominates the visualization (see red curly bracket for some) and it is covered before
- *Neural, neuronal, structural, and brain* are the most frequently used words before plasticity (green arrows)









# Text analysis in R?!

If you are an R user, consider [Monica Alexander's](#) writeup and RMD file below that takes **demography** papers and does cool text analysis on them.

Link to Monica's slides/files:

[https://github.com/MJAlexander/demopop-workshop/blob/main/rmd/2\\_articles.Rmd](https://github.com/MJAlexander/demopop-workshop/blob/main/rmd/2_articles.Rmd)

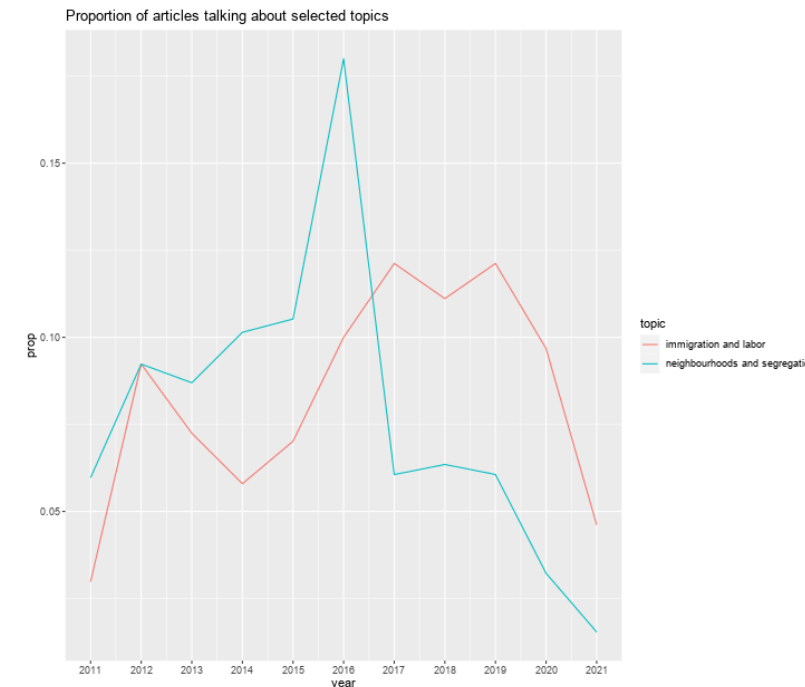
In addition, you might want to check out tm (text mining) and stm (structural topic models) in R.

```
r$> bigrams_separated <- bigrams %>%  
  separate(bigram, c("word1", "word2"), sep = " ")  
  
bigrams_filtered <- bigrams_separated %>%  
  filter(!word1 %in% stop_words$word) %>%  
  filter(!word2 %in% stop_words$word)  
  
bigrams_united <- bigrams_filtered %>%  
  unite(bigram, word1, word2, sep = " ") %>%  
  filter(bigram!="NA NA")  
  
bigrams_united %>%  
  group_by(bigram) %>%  
  tally() %>%  
  arrange(-n) %>%  
  filter(!str_detect(bigram, "table\\ "), bigram!= "online appendix",  
        !str_detect(bigram, "al "),  
        !str_detect(bigram, "resource")) %>%  
  top_n(20)
```

Selecting by n  
# A tibble: 20 x 2

bigram	n
<chr>	<int>
1 statistically significant	1925
2 fixed effects	1866
3 life expectancy	1734
4 labor market	1643
5 standard errors	1551
6 birth weight	1263
7 labor force	1193
8 family structure	1126
9 sex couples	1085
10 family size	1018
11 data set	954
12 infant mortality	928
13 race ethnicity	923
14 mortality rates	871
15 dummy variables	856
16 age specific	848
17 cross sectional	834
18 foreign born	822
19 sex ratio	800
20 low income	739

r\$> █





## Structural topic models (one avenue of further modeling text data)

- Was an introduction to “text as data”, using examples from scientific publications’ abstracts from our previous research. Examples were in nltk, python base libraries and a bit on noun-phrase-clauses identification and extraction using Apache’s OpenNLP and in python with spacy pre-trained models.
- Now, brief on modelling possibilities (e.g., structural topic models, stm package in R)
  - See example slides here:  
[https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/2\\_presentations/10\\_Structural\\_topic\\_models\\_an\\_example.pdf](https://github.com/akbaritabar/Parallelised-analysis-of-large-scale-bibliometric-and-text-data/blob/main/2_presentations/10_Structural_topic_models_an_example.pdf)
- Other topic modeling possibilities (e.g., LDA and SBM-like methods\*\*)

\*\* Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, 4(7), eaaq1360. <https://doi.org/10.1126/sciadv.aaq1360>



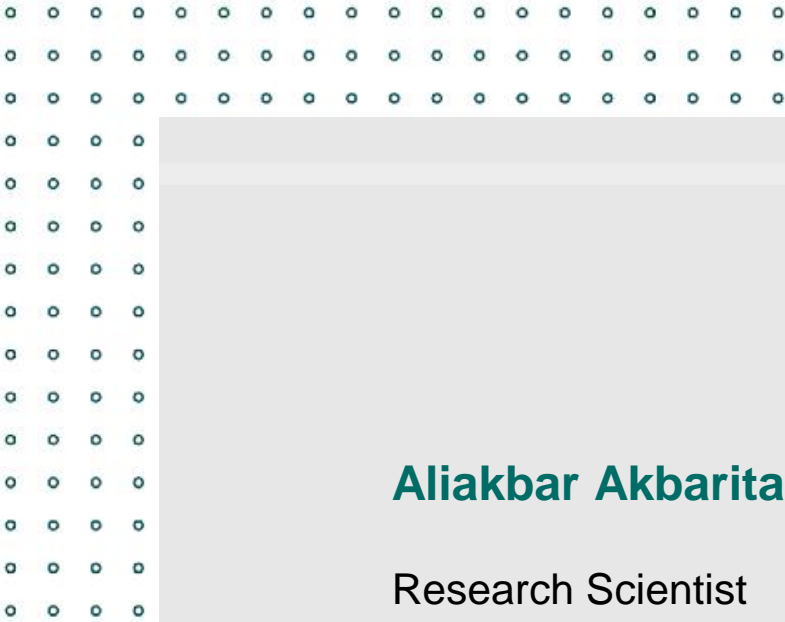
## Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

## Hands-on session with examples



Please tweet with hashtag  
#BiblioDemography,  
a tribute to James W. Vaupel (1945-2022).  
Thanks Ilya and Jonas for bringing up Jim's  
labeling idea!



**Aliakbar Akbaritabar**

Research Scientist

[Akbaritabar@demogr.mpg.de](mailto:Akbaritabar@demogr.mpg.de)

<https://akbaritabar.github.io/>

<https://twitter.com/Akbaritabar>



**THANK YOU!**

