

# Построение ландшафта области знаний

## для курса 'Количественные методы в гуманитарных науках: критическое введение' (2024, НИУ ВШЭ)

Г. А. Мороз (Международная лаборатория языковой  
конвергенции)

13.04.2024

*Как все люди Библиотеки, в юности я путешествовал. Это было паломничество в поисках книги, возможно каталога каталогов...*

Х. Л. Борхес, “Вавилонская библиотека”

# “Информационный гриб”

- Данные вокруг гуманитария?
- Данные вокруг гуманитария...
- Данные вокруг гуманитария!
- Данные вокруг гуманитария:

- Данные вокруг гуманитария?
- Данные вокруг гуманитария...
- Данные вокруг гуманитария!
- Данные вокруг гуманитария: Можем быть мы можем попробовать хотя бы обозреть эти самые данные с высоты птичьего полета?

# Научных публикаций очень много: желаемое

Google Академия



Стоя на плечах гигантов

# Научных публикаций очень много: желаемое

## Google Академия



Стоя на плечах гигантов

- “...Мы подобны карликам, усевшимся на плечах великанов; мы видим больше и дальше, чем они, не потому, что обладаем лучшим зрением, и не потому, что выше их, но потому, что они нас подняли и увеличили наш рост собственным величием”  
высказывание приписывают Бернару Шартрскому,  
французскому философи XI-XII

# Научных публикаций очень много: желаемое

## Google Академия



Стоя на плечах гигантов

- “...Мы подобны карликам, усевшимся на плечах великанов; мы видим больше и дальше, чем они, не потому, что обладаем лучшим зрением, и не потому, что выше их, но потому, что они нас подняли и увеличили наш рост собственным величием” высказывание приписывают Бернару Шартрскому, французскому философи XI-XII
- “Today we are privileged to sit side-by-side with the giants on whose shoulders we stand.” Gerald Holton, “On the recent past of physics,” American Journal of Physics, 29 (December, 1961), 805.

# Научных публикаций очень много: желаемое

Details



Science  
Volume 134, Issue 3473  
Jul 1961

ARTICLE

Impact of Large-Scale Science on the United States

Big science is here to stay, but we have yet to make the hard financial and educational choices it imposes.

[View article page](#)

Alvin M. Weinberg

1961 by the American Association for the Advancement of Science



Throughout history, societies have expressed their aspirations in large-scale, monumental enterprises which, though not necessary for the survival of the societies, have taxed them to their physical and intellectual limits. History often views these monuments as symbolizing the societies. The Pyramids, the Sphinx, and the great temple at Karnak symbolize Egypt; the magnificent cathedrals symbolize the church culture of the Middle Ages; Versailles symbolizes the France of Louis XIV; and so on. The societies were goaded into these extraordinary exertions by their rulers—the pharaoh, the church, the king—who invoked the cultural mystique when this was sufficient, but who also used force when necessary. Sometimes, as with the cathedrals, local

# Научных публикаций очень много: желаемое

## Details



Science  
Volume 134, Issue 3473  
Jul 1961

### ARTICLE

#### Impact of Large-Scale Science on the United States

Big science is here to stay, but we have yet to make the hard financial and educational choices it imposes.

[View article page](#)

Alvin M. Weinberg

1961 by the American Association for the Advancement of Science



who also used force when necessary. Sometimes, as with the cathedrals, local pride and a sense of competition with other cities helped launch the project. In many cases the distortion of the economy caused by construction of the big monuments contributed to the civilization's decline.

When history looks at the 20th century, she will see science and technology as its theme; she will find in the monuments of Big Science—the huge rockets, the high-energy accelerators, the high-flux research reactors—symbols of our time just as surely as she finds in Notre Dame a symbol of the Middle Ages. She might even see analogies between our motivations for building these tools of giant science

## Научных публикаций очень много: реальность

- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания

## Научных публикаций очень много: реальность

- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
  - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект

## Научных публикаций очень много: реальность

- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
  - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект
  - ... люди могут хакнуть и обессмыслить любую метрику

## Научных публикаций очень много: реальность

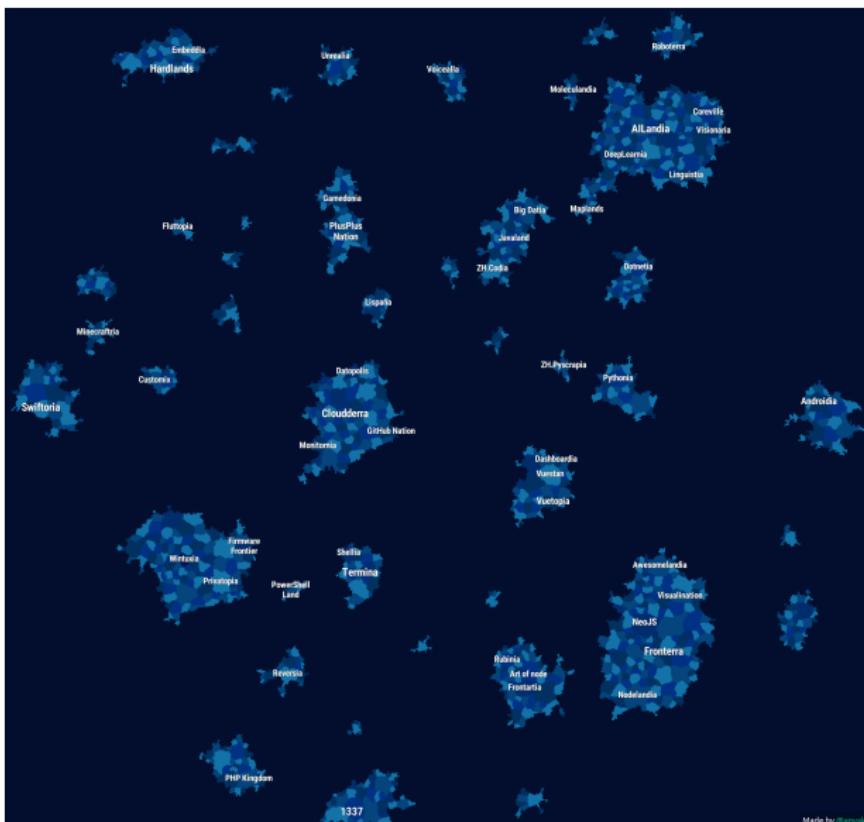
- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
  - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект
  - ... люди могут хакнуть и обессмыслиТЬ любую метрику
- Исследователи больше любят новые исследования: на материале 726 медицинских статей, содержащих 17 895 научных ссылок, авторы приходят к выводу, что вне зависимости от журнала более 70% цитируемых работ опубликованы не более 10 лет до публикации работы. [Chow et al., 2023]

## Научных публикаций очень много: реальность

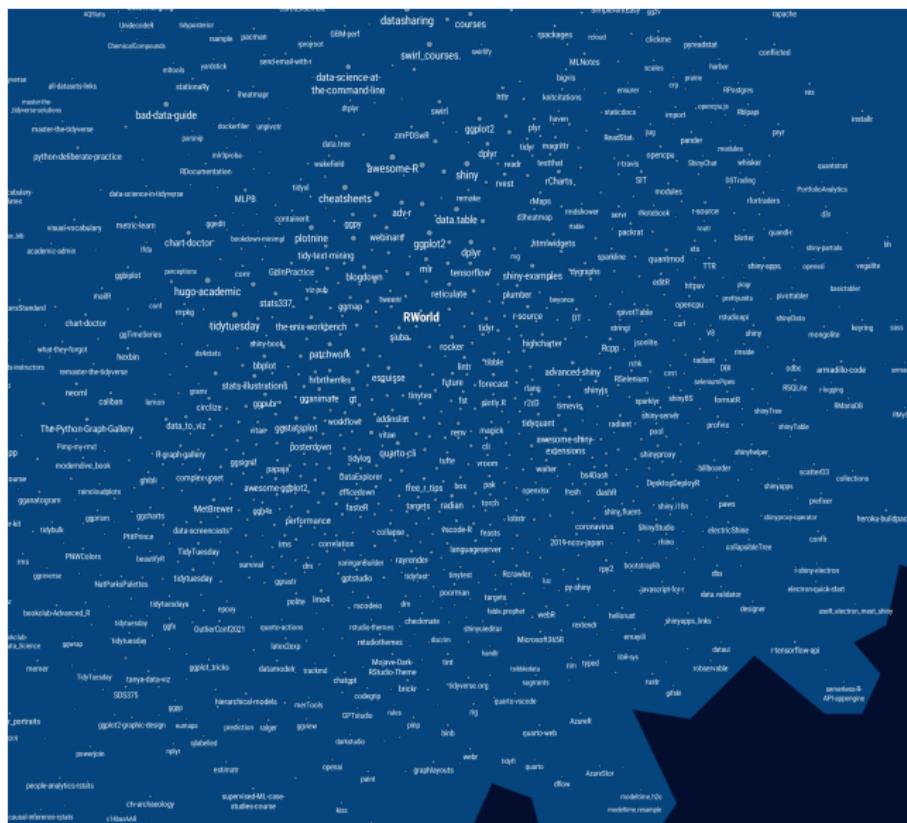
- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
  - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект
  - ... люди могут хакнуть и обессмыслиТЬ любую метрику
- Исследователи больше любят новые исследования: на материале 726 медицинских статей, содержащих 17 895 научных ссылок, авторы приходят к выводу, что вне зависимости от журнала более 70% цитируемых работ опубликованы не более 10 лет до публикации работы. [Chow et al., 2023]
- Даже цифра может подгнить: авторы обнаружили значительную долю “мертвых” URL статей, которые упоминаются при цитировании в публикациях в медицине. [Klein et al., 2014]

# Ландшафты

# Карта репозиториев гитхаба (Андрей Кашча)



# Карта репозиториев гитхаба (Андрей Кашча)



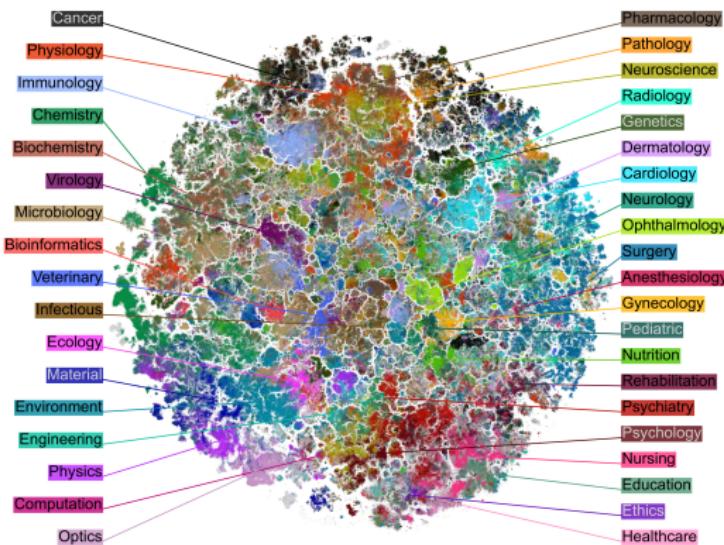
## [Gonzalez-Marquez et al., 2023]

*The number of publications in biomedicine and life sciences has rapidly grown over the last decades, with over 1.5 million papers now published every year. This makes it difficult to keep track of new scientific works and to have an overview of the evolution of the field as a whole. Here we present a 2D atlas of the entire corpus of biomedical literature, and argue that it provides a unique and useful overview of the life sciences research. <...>*

<https://static.nomic.ai/pubmed.html> (интерактивная версия)

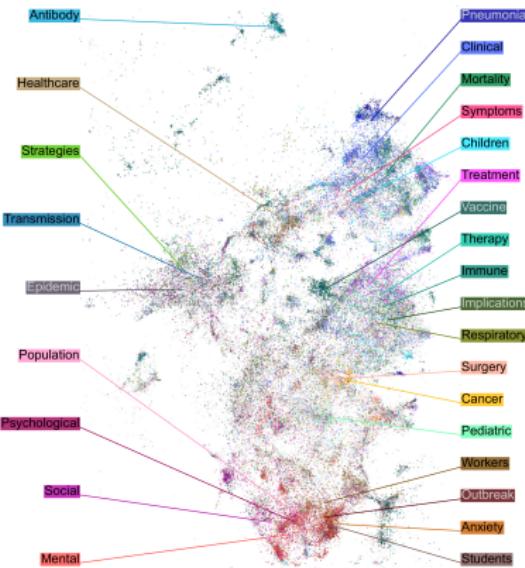
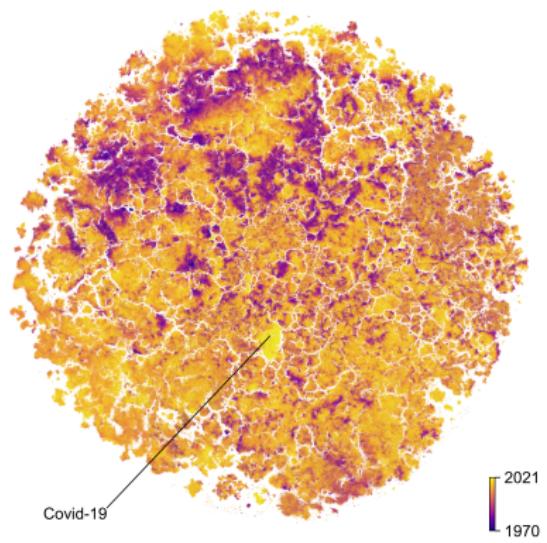
Это препринт!

## [Gonzalez-Marquez et al., 2023]



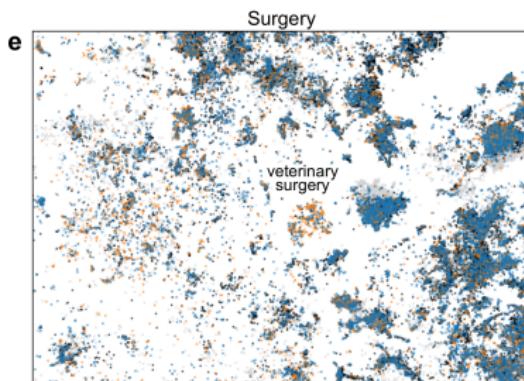
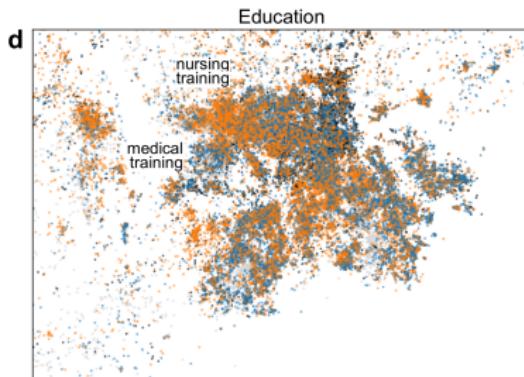
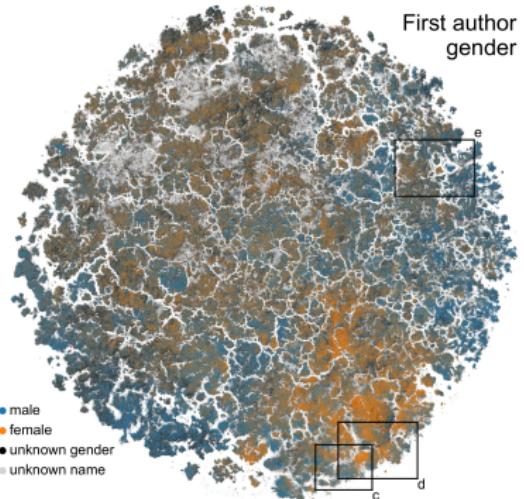
2D эмбеддинги на основе 21 миллиона аннотаций, которые были трансформированы в 768-мерное векторное пространство при помощи PubMedBERT [Gu et al., 2021], а дальше сплюснутая в 2D при помощи t-SNE [Van der Maaten and Hinton, 2008]. Цвета основаны на названиях журналов.

## [Gonzalez-Marquez et al., 2023]



Регион карты, посвященный Covid-19. Цвета приписаны на основе названий работ. Кроме того здесь есть около 15% работ не посвященных короновирусу.

## [Gonzalez-Marquez et al., 2023]



Статьи раскрашены по полу первого автора.

## Другие проекты Nomic

- map of Wikipedia
- map of Twitter
- другие <https://atlas.nomic.ai/discover>

# Похожее

## Библиометрические исследования?

Библиометрия — дисциплина, возникшая в конце XIX века, в рамках которой можно встретить разные применения математических методов к исследованию научных работ. Наиболее известные применения:

- графы соавторства
- библиографические ссылки
- ключевые слова
- измерение качества журналов
- и др.

## Distant Reading?

Дальнее чтение [Moretti, 2013] — это не какой-то один метод, а целое семейство методов анализа литературных текстов и их структуры, а также подразумевающий некоторый осмыслиенный с точки зрения литературоведения исследовательский вопрос.

"Информационный гриб"  
○○○○○

Ландшафты  
○○○○○○○

Похожее  
○○○

Техническое  
●○○○○○

Исследование лингвистики  
○

Ограничения метода  
○○

References

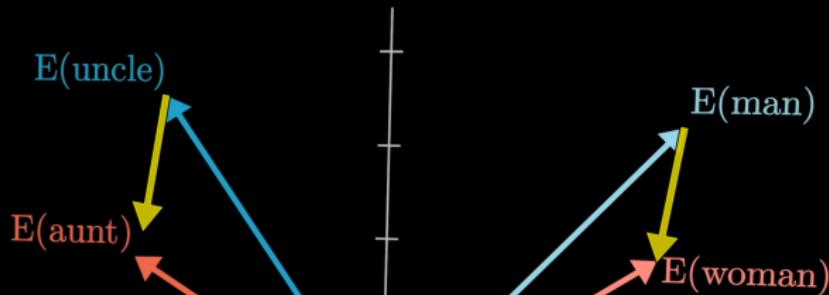
# Техническое

## Эмбеддинги

- Архитектурам машинного обучения любой сложности при работе с языковыми данными нужно уметь преобразовывать слова (на самом деле некоторые кусочки письменных слов) в наборы чисел, которые обычно называют **вектором**.
- Числа для вектора каждого конкретного слова обычно получают на основе контекстов, в которых оно появляется в обучающем корпусе.
- Слова с похожим значением будут направлены в одну сторону. Сравнивать их следует по углу между векторами.
- В работах [Mikolov et al., 2013a,b] от Google была представлена модель word2vec, архитектура нейросети для создания векторных моделей.
- Совсем недавно вышли видео 3Blue1Brown, в которых это обсуждается подробнее:
  - But what is a GPT? Visual intro to transformers
  - Attention in transformers, visually explained

## Эмбеддинги

$$E(aunt) - E(uncle) \approx E(woman) - E(man)$$



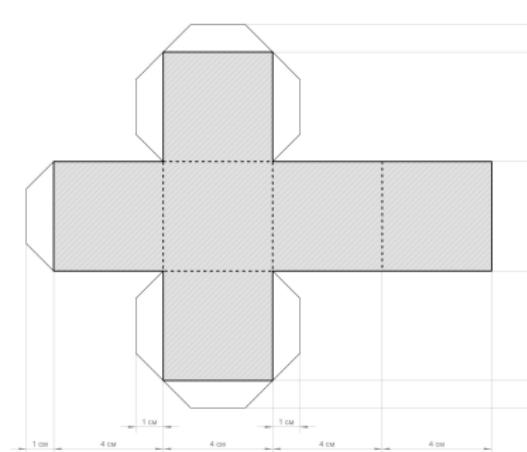
Взято из видео 3Blue1Brown.

## doc2vec

- Чтобы анализировать тексты [Le and Mikolov, 2014] предложили разбивать их на абзацы и конкатенировать векторы, которые входят в абзац, а потом использовать их для кластеризации текстов.
- Если применять эту логику к предложениям, то это позволяет не терять информацию о месте слова.

## Уменьшение размерности

- Эмбеддинги — многомерные вектора чисел, например, в GPT-3 50 тысяч токенов закодировано при помощи векторов длиной 12 тысяч. Смотреть на это пространство глазами нельзя, но можно попробовать уменьшить размерность.



## Уменьшение размерности

- Эмбеддинги — многомерные вектора чисел, например, в GPT-3 50 тысяч токенов закодировано при помощи векторов длиной 12 тысяч. Смотреть на это пространство глазами нельзя, но можно попробовать уменьшить размерность.
- Популярные алгоритмы:
  - Singular value decomposition (SVD)
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - Uniform Manifold Approximation and Projection (UMAP)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)

# Исследование лингвистики

# Ограничения метода

*Мне известен дикий край, где библиотекари отказались от суеверной  
и напрасной привычки искать в книгах смысл, считая, что это все  
равно что искать его в снах или в беспорядочных линиях руки...*

Х. Л. Борхес, “Вавилонская библиотека”

## Список литературы I

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222, 2015.

Natalie LY Chow, Natalie Tateishi, Alexa Goldhar, Rabia Zaheer, Donald A Redelman, Amy H Cheung, Ayal Schaffer, and Mark Sinyor. Does knowledge have a half-life? an observational study analyzing the use of older citations in medical and scientific publications. *BMJ open*, 13(5):e072374, 2023.

Rita Gonzalez-Marquez, Luca Schmidt, Benjamin M Schmidt, Philipp Berens, and Dmitry Kobak. The landscape of biomedical research. *bioRxiv*, 2023. doi: <https://doi.org/10.1101/2023.04.10.536208>.

## Список литературы II

- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. Scholarly context not found: one in five articles suffers from reference rot. *PLoS one*, 9(12):e115253, 2014.
- Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

## Список литературы III

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Franco Moretti. *Distant reading*. Verso Books, 2013.

Derek J. de Solla Price. *Little science, big science*. Columbia University Press, 1963.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.