

# Культурная аналитика

Еще парочка примеров



# What a Deep Neural Network thinks about your #selfie

Oct 25, 2015



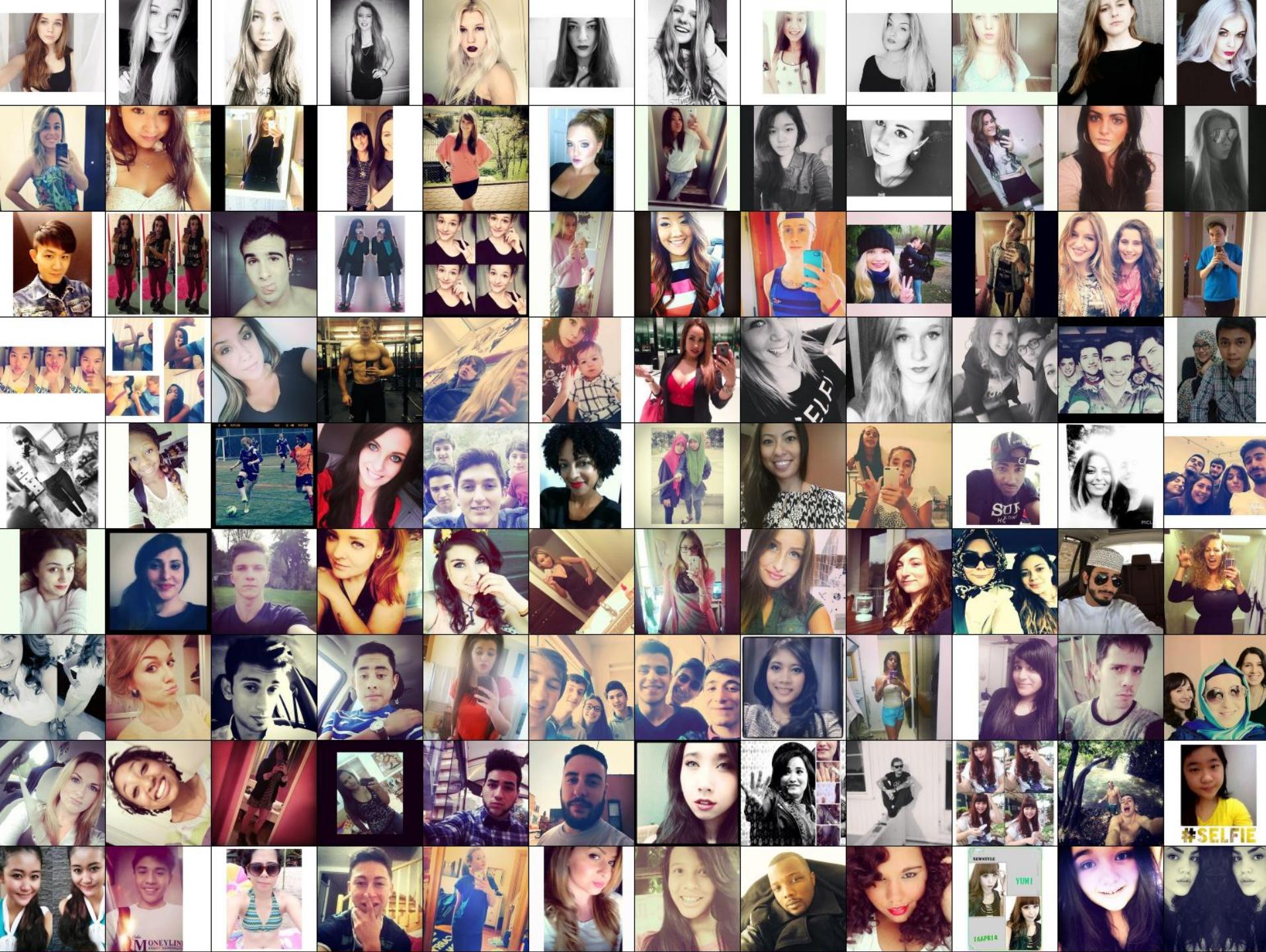
Convolutional Neural Networks are great: they recognize things, places and people in your personal photos, signs, people and lights in self-driving cars, crops, forests and traffic in aerial imagery, various anomalies in medical images and all kinds of other useful things. But once in a while these powerful visual recognition models can also be warped for distraction, fun and amusement. In this fun experiment we're going to do just that: We'll take a powerful, 140-million-parameter state-of-the-art Convolutional Neural Network, feed it 2 million selfies from the internet, and train it to classify good selfies from bad ones. Just because it's easy and because we can. And in the process we might learn how to take better selfies :)

[karpathy.github.io/2015/10/25/selfie/](http://karpathy.github.io/2015/10/25/selfie/)

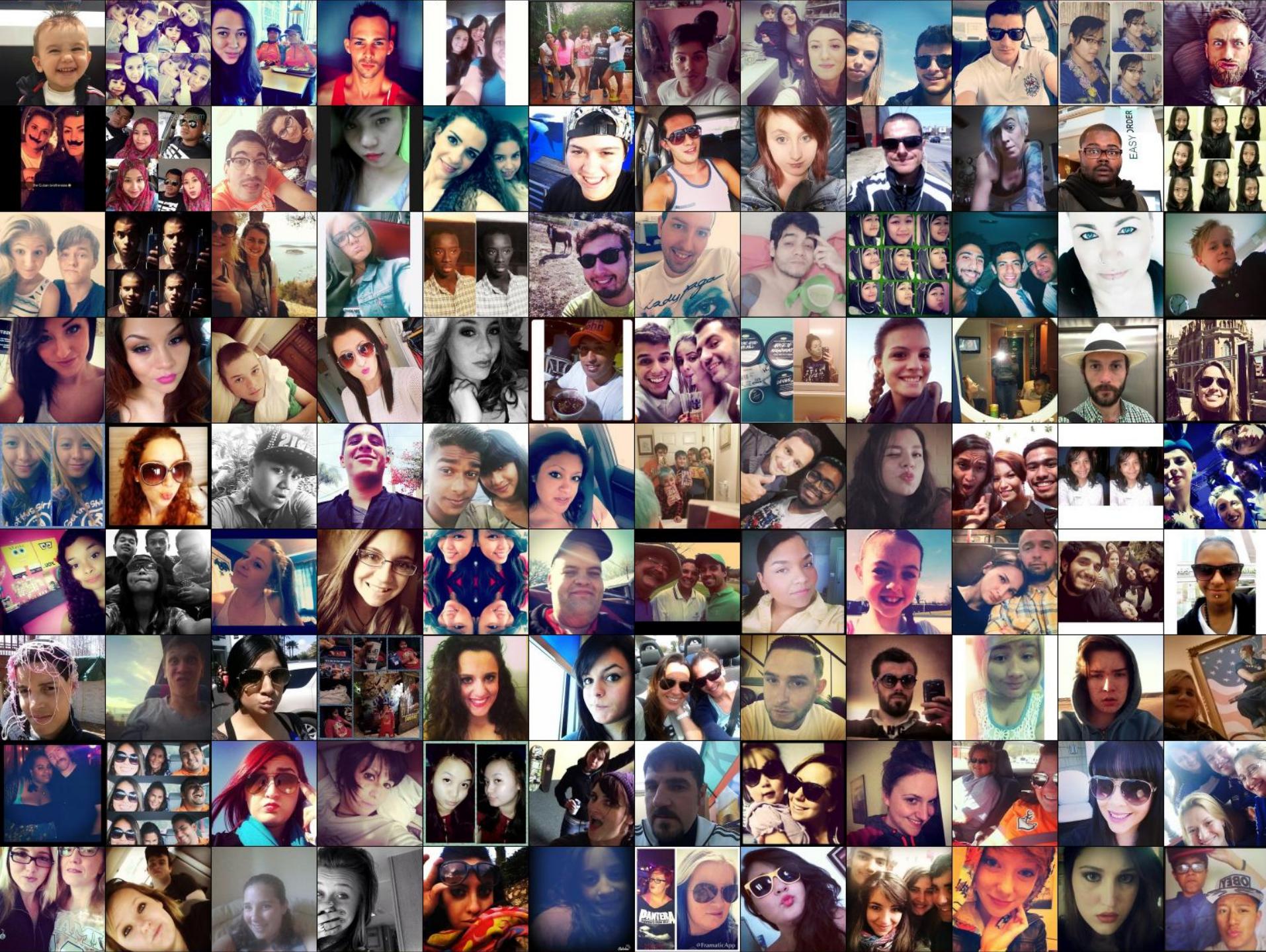
# Что он сделал?

1. Отобрал 5 млн изображений в Instagram по тегу #selfie 
2. С помощью обученной модели компьютерного зрения отобрал 2 млн, которые точно содержат хотя бы одно лицо 
3. Набрал примеры красивых и некрасивых селфи  | 
4. Обучил на этих данных свою модель (сверточную нейросеть) отделять хорошие селфи от плохих
5. Натравил эту модель на 50 тыс. селфи, которые модель не видела

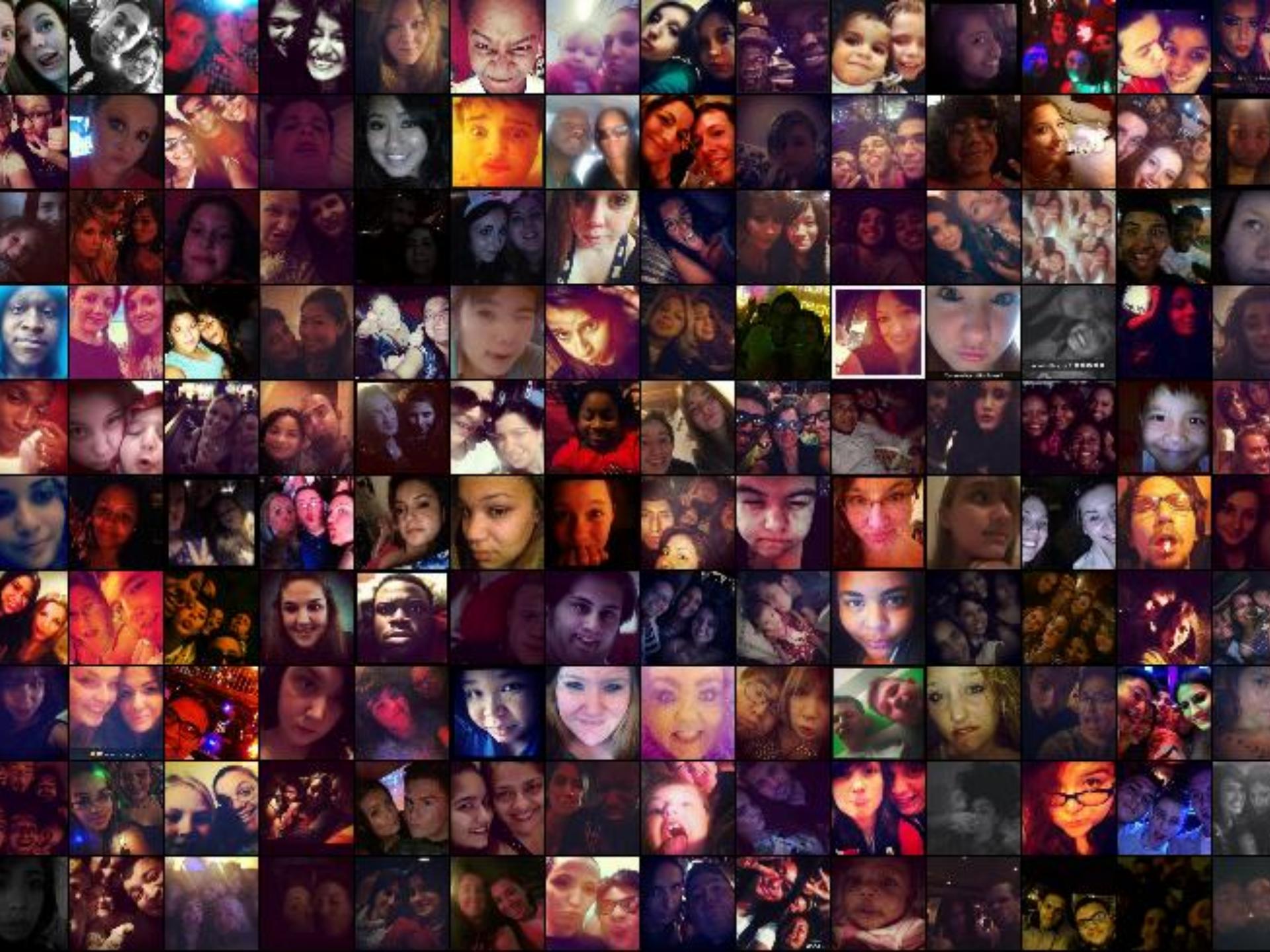
The good...



...the bad...



...and the ugly



# Правила хорошего селфи (v. 2015)

Нужно:

- Быть девушкой;
- Отводить под лицо примерно треть площади фото;
- Демонстрировать длинные волосы;
- Использовать фильтры. Популярные варианты — ч/б, или наоборот перенасыщенные цветом;
- Добавлять белые рамки.

Не нужно:

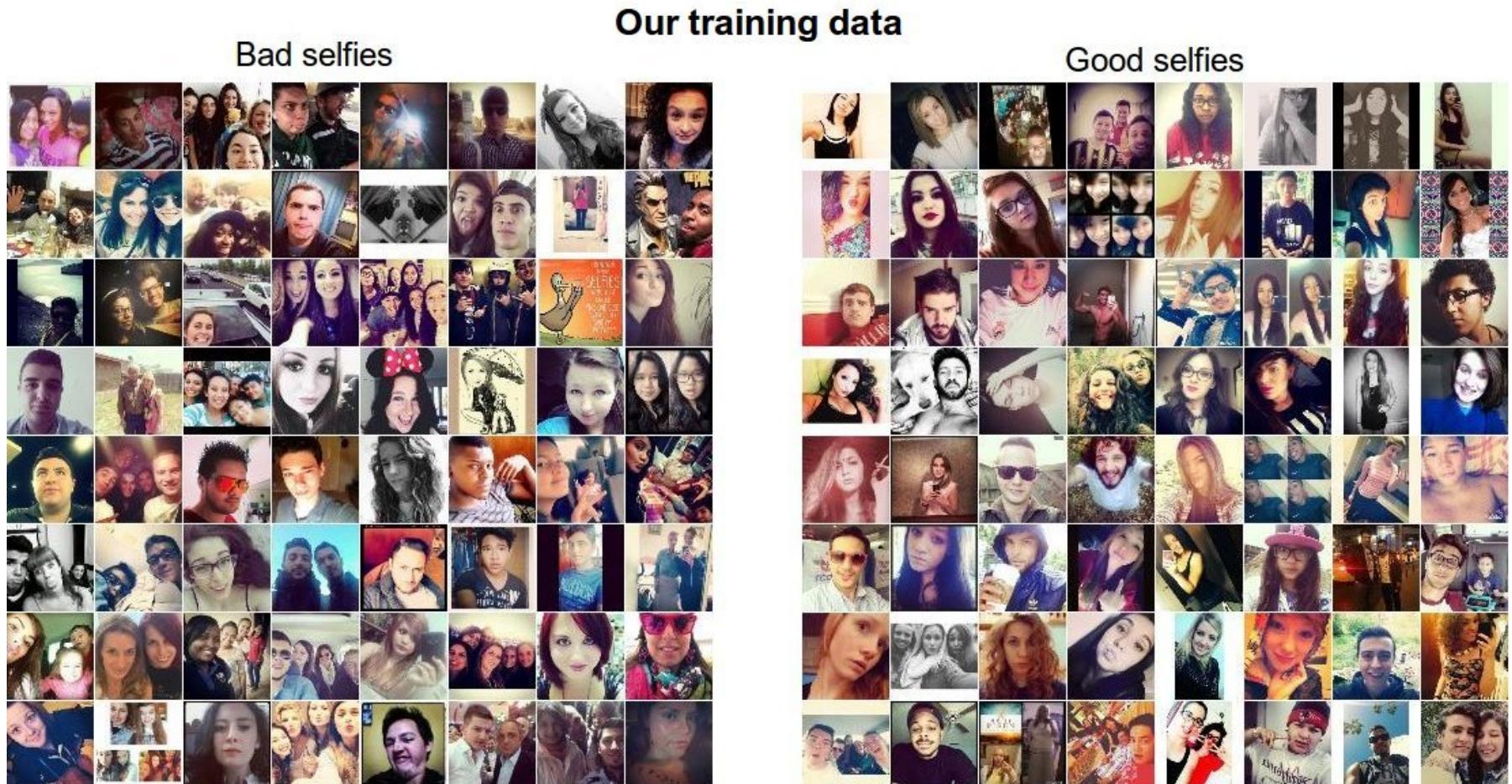
- Снимать в темноте;
- Делать голову/лицо слишком большими;
- Делать групповые селфи.

Стоп-стоп-стоп, а как  
это вообще сделано,  
непонятно

# Как это сделано

- Нейросетевая модель обучалась на основе лайков (да, тогда инстаграм еще не закрыл лайки ❤️)
- Автор поделил авторов селфи на группы с примерно одинаковым числом подписчиков
- В каждой группе отсортировал селфи по числу лайков
- Фотографии, собравшие больше сердечек, были признаны хорошими; меньше — плохими.

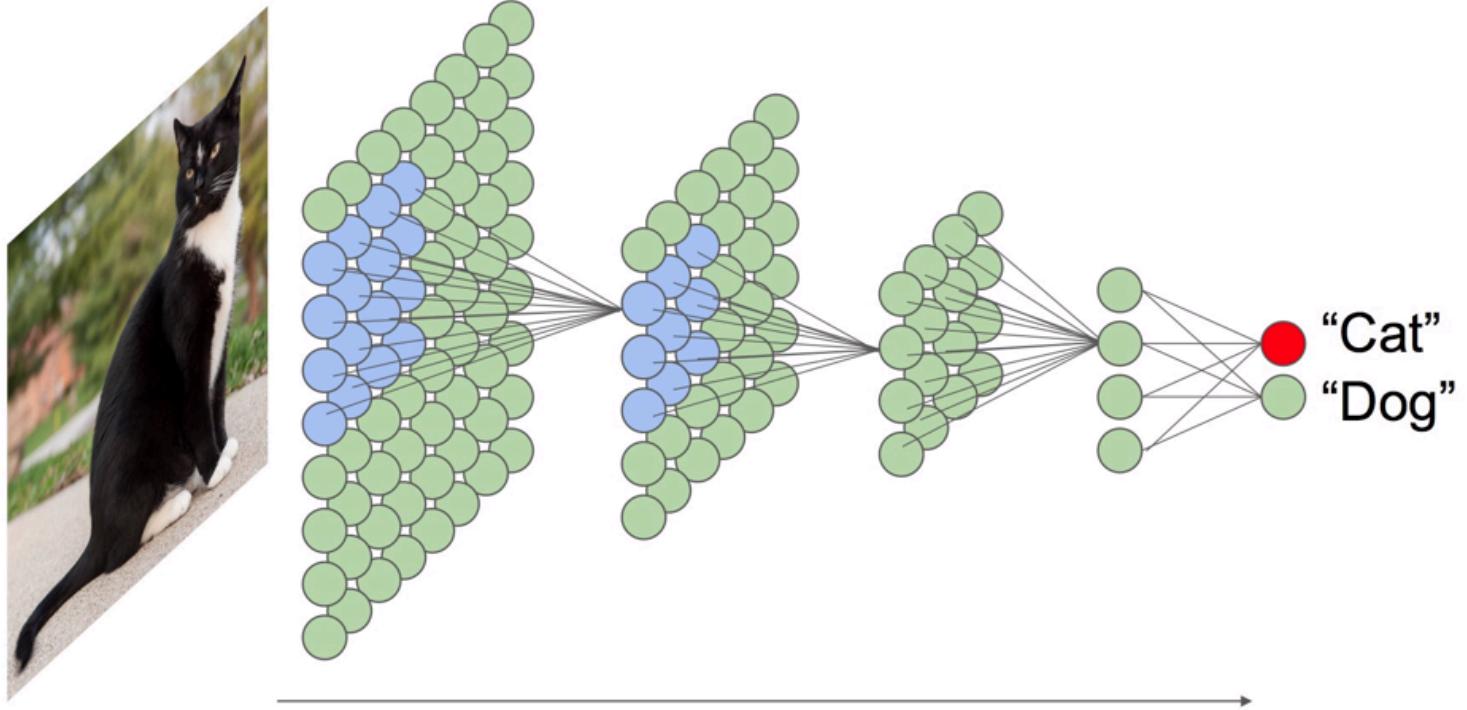
# На чем обучалась нейросеть:



# Так, а что такое нейросеть?

- Компьютерная модель (алгоритм), которая умеет обобщать данные
- Если данных много, модель накапливает достаточно информации, чтобы обобщать не хуже человека
- В результате получается решать некоторые «нешаблонные» задачи (которые нельзя просто «запрограммировать»)

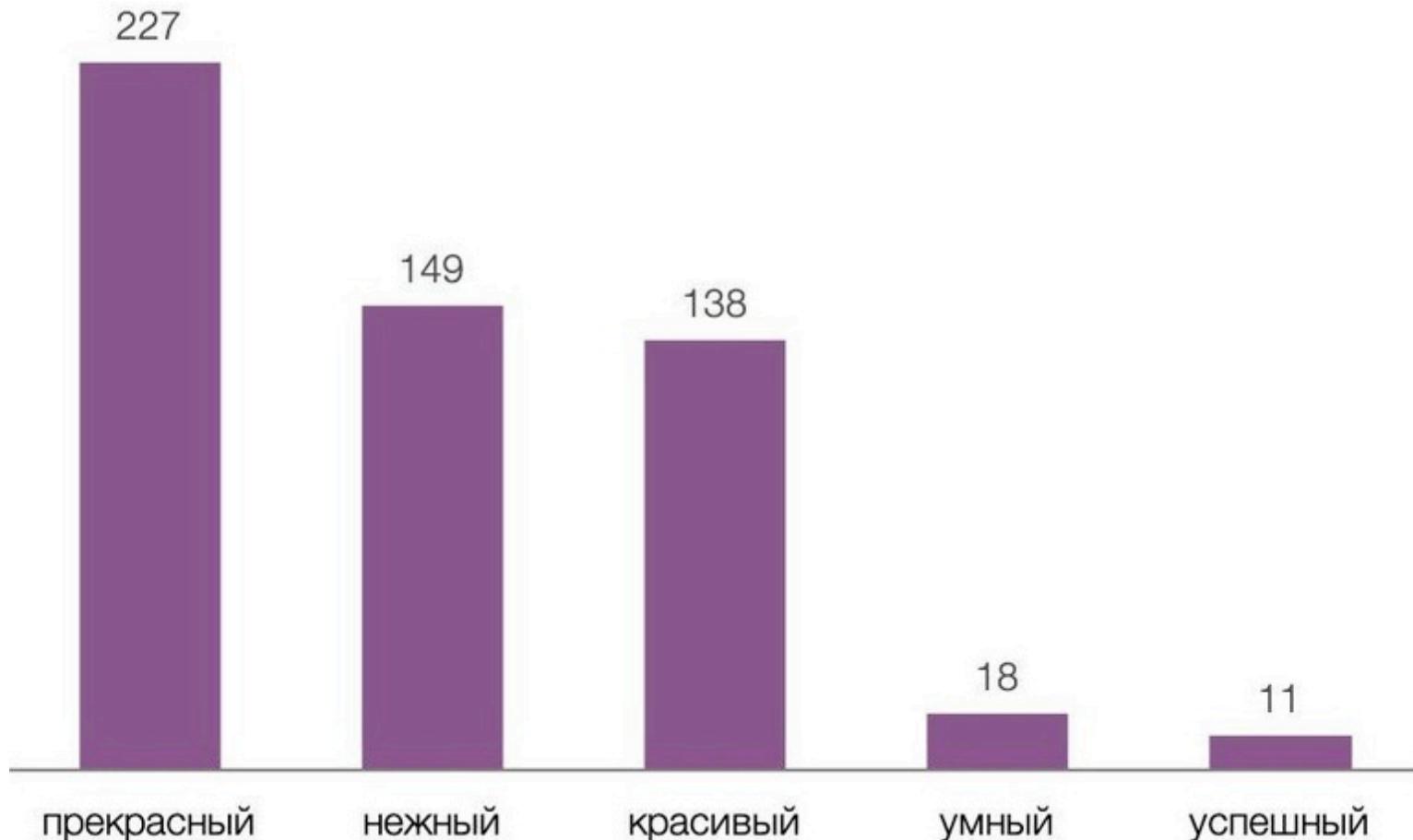
# Типичный пример такой задачи



# В нашем случае

- модель получила 50000 других селфи, которые НЕ использовались при обучении
- И неплохо их отсортировала!
- А человек — проинтерпретировал

# Cultural analytics на коленке: что желают женщинам на 8 марта?

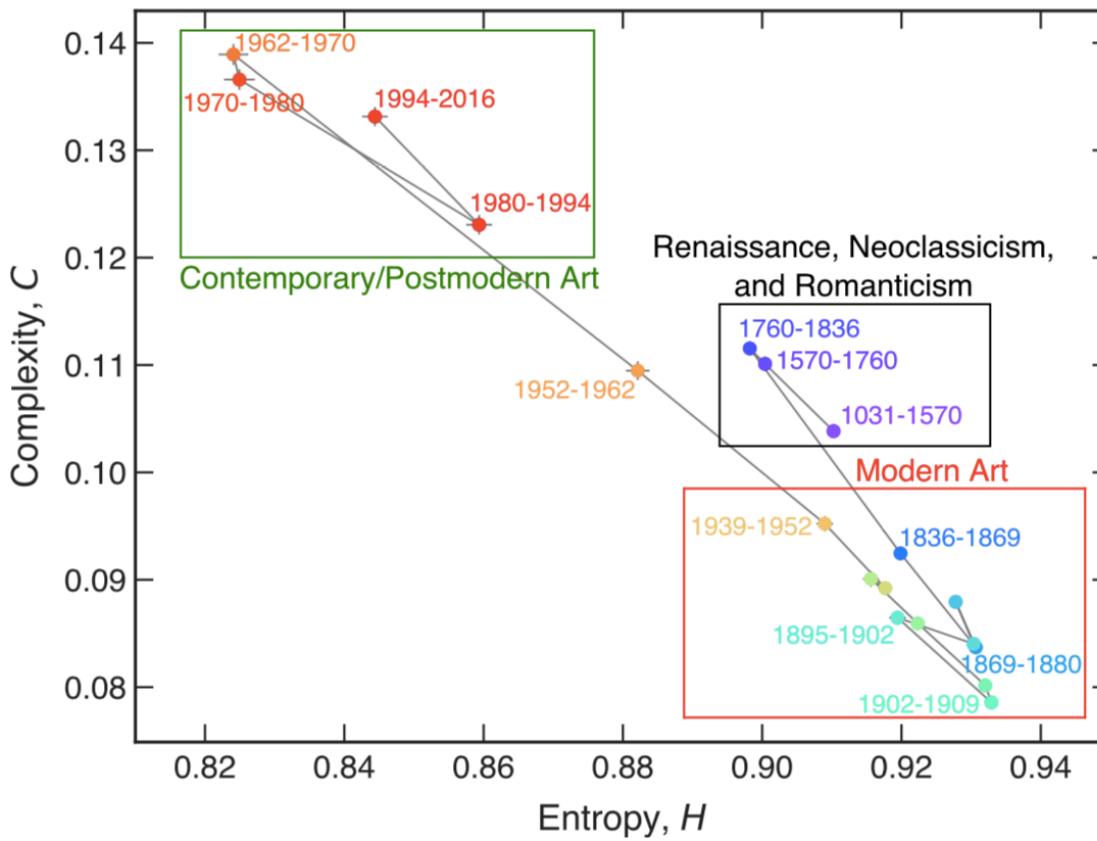


Исследовала Кира Харлашова, DH HSE

[sysblok.ru/society/smejtes-i-detej-rozhajte-iz-chego-sdelany-pozdravlenija-s-8-marta/](http://sysblok.ru/society/smejtes-i-detej-rozhajte-iz-chego-sdelany-pozdravlenija-s-8-marta/)

Более  
«высокодуховный»  
пример

# ЭВОЛЮЦИЯ ИСКУССТВА

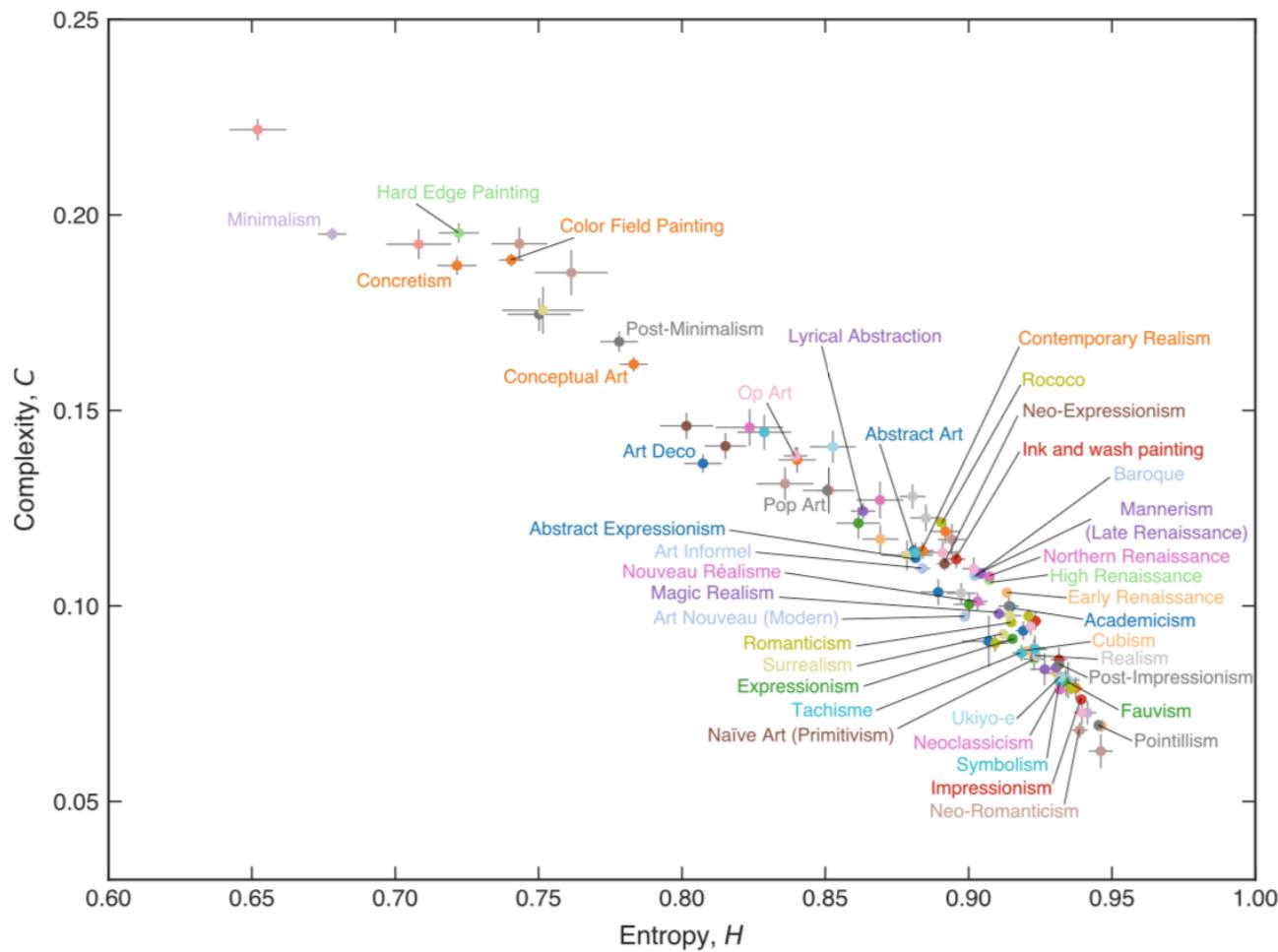


(пример с прошлого раза)

- 137 тыс. картин с WikiArt
- Энтропия (entropy)
- Сложность (complexity)
- Видны переходы:
  - от домодернистского искусства — к модерну
  - От модерна — к постмодерну

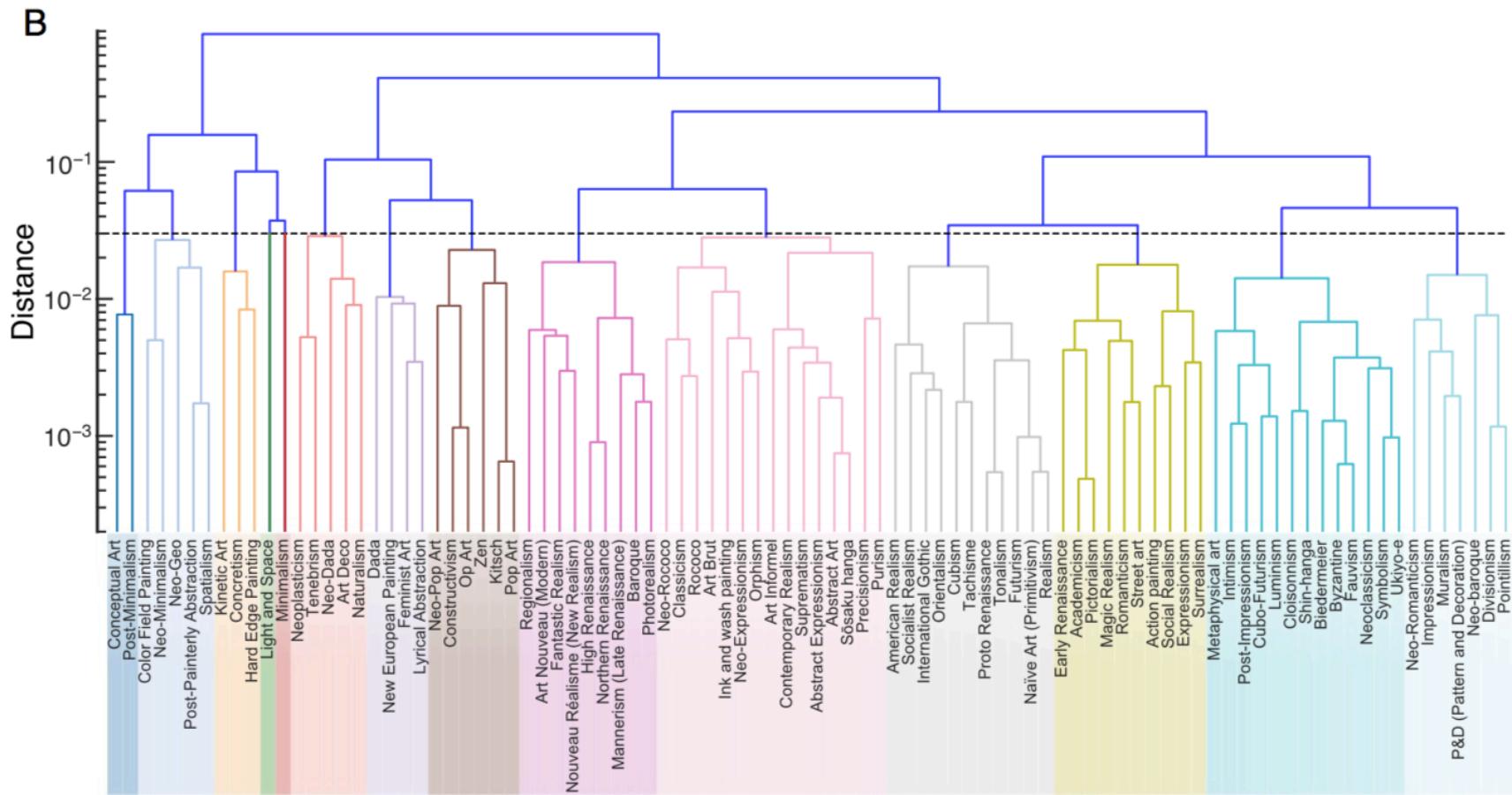
Sigaki H.Y.D., Perc M., Ribeiro H.V. History of art paintings through the lens of entropy and complexity // PNAS. 2018. Vol. 115, № 37. P. E8585–E8594.





Sigaki H.Y.D., Perc M., Ribeiro H.V. History of art paintings through the lens of entropy and complexity // PNAS. 2018. Vol. 115, № 37. P. E8585–E8594.

# Кластеризация жанров



Sigaki H.Y.D., Perc M., Ribeiro H.V. History of art paintings through the lens of entropy and complexity // PNAS. 2018. Vol. 115, № 37. P. E8585–E8594.

# Культуромика

можно ли «делать науку» в Google Ngram Viewer?



# Цель этого занятия

- Познакомиться с идеей культуромики
- Посмотреть на примеры культуромики
- Понять проблемы культуромики
- Подумать, можно ли сделать «культуромунику здорового человека»

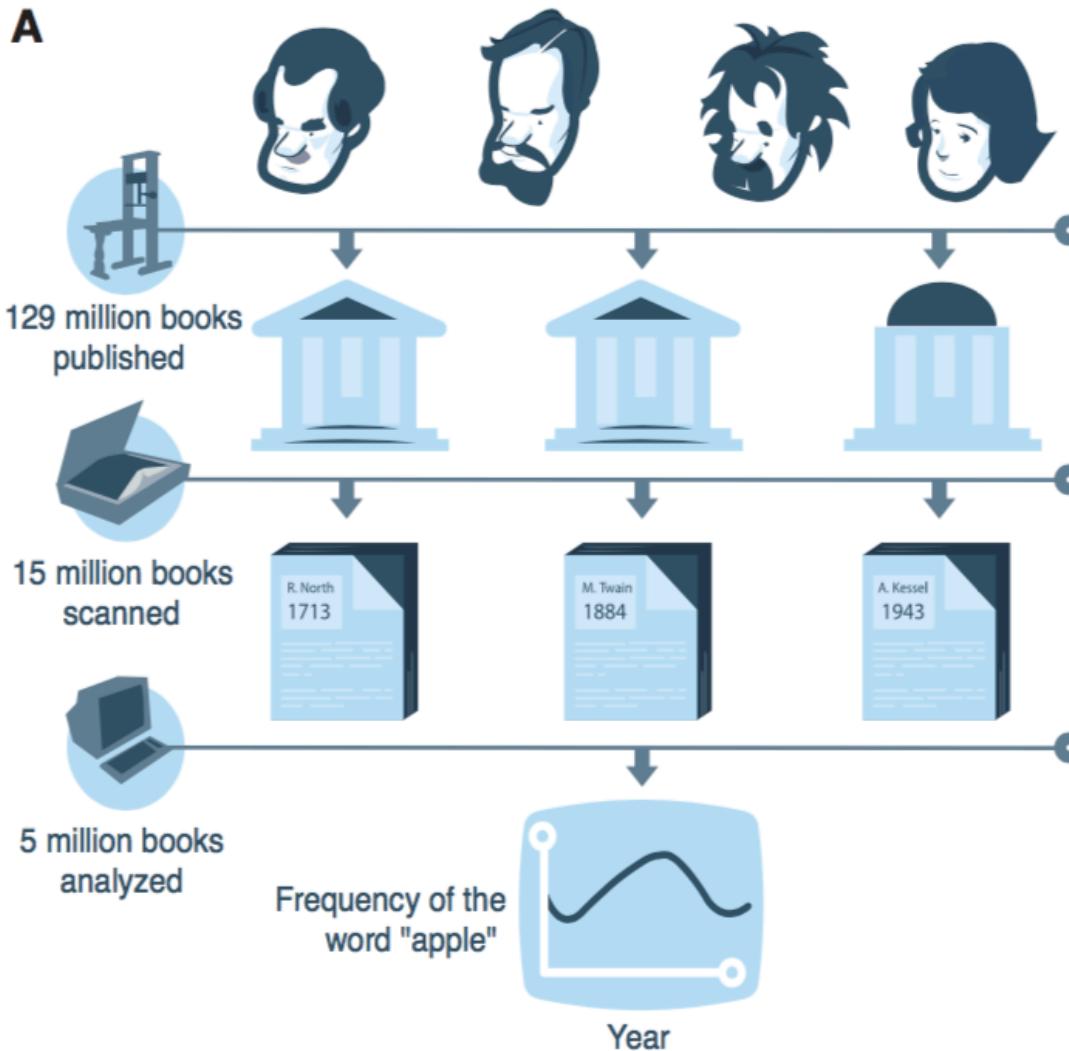
# С чего все началось?

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively.

[*Мы собрали корпус оцифрованных текстов, в котором содержится около 4% всех когда либо напечатанных книг. Анализ этого корпуса позволяет нам делать количественные исследования культурных трендов*]

Michel J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 2011. Vol. 331, № 6014. P. 176–182., перевод мой

# Google Books



# Корпус Google Books действительно большой

- The corpus cannot be read by a human. If you tried to read only the entries from the year 2000 alone, at the reasonable pace of 200 words/minute, without interruptions for food or sleep, it would take eighty years.
- The sequence of letters is one thousand times longer than the human genome: if you wrote it out in a straight line, it would reach to the moon and back 10 times over.
- The resulting corpus contains over 500 billion words, in English (361 billion), French (45B), Spanish (45B), German (37B), Chinese (13B), Russian (35B), and Hebrew (2B)

Michel J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 2011. Vol. 331, № 6014. P. 176–182.

# Но впечатлило всех не это, а масштаб замысла

<...> this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology.

[<...> этот подход может дать новое знание в широком спектре областей: лексикографии, эволюции, исследованиях коллективной памяти, скорости внедрения технологий, феноменов славы и популярности, а также цензуры и исторической эпидемиологии]

В статье был замах на «новую науку»

“Culturomics” extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

[“Культуромика” расширяет границы точных количественных исследований, распространяясь на множество гуманитарных и социальных феноменов.]

Michel J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 2011. Vol. 331, № 6014. P. 176–182., перевод мой

The background of the image is a dense, circular arrangement of numerous small, colorful sticks or straws, creating a textured, radial pattern. The colors are primarily shades of yellow, brown, and green, with some red and blue accents. The sticks are densely packed in the center and taper off towards the edges.

14 January 2011 • 318

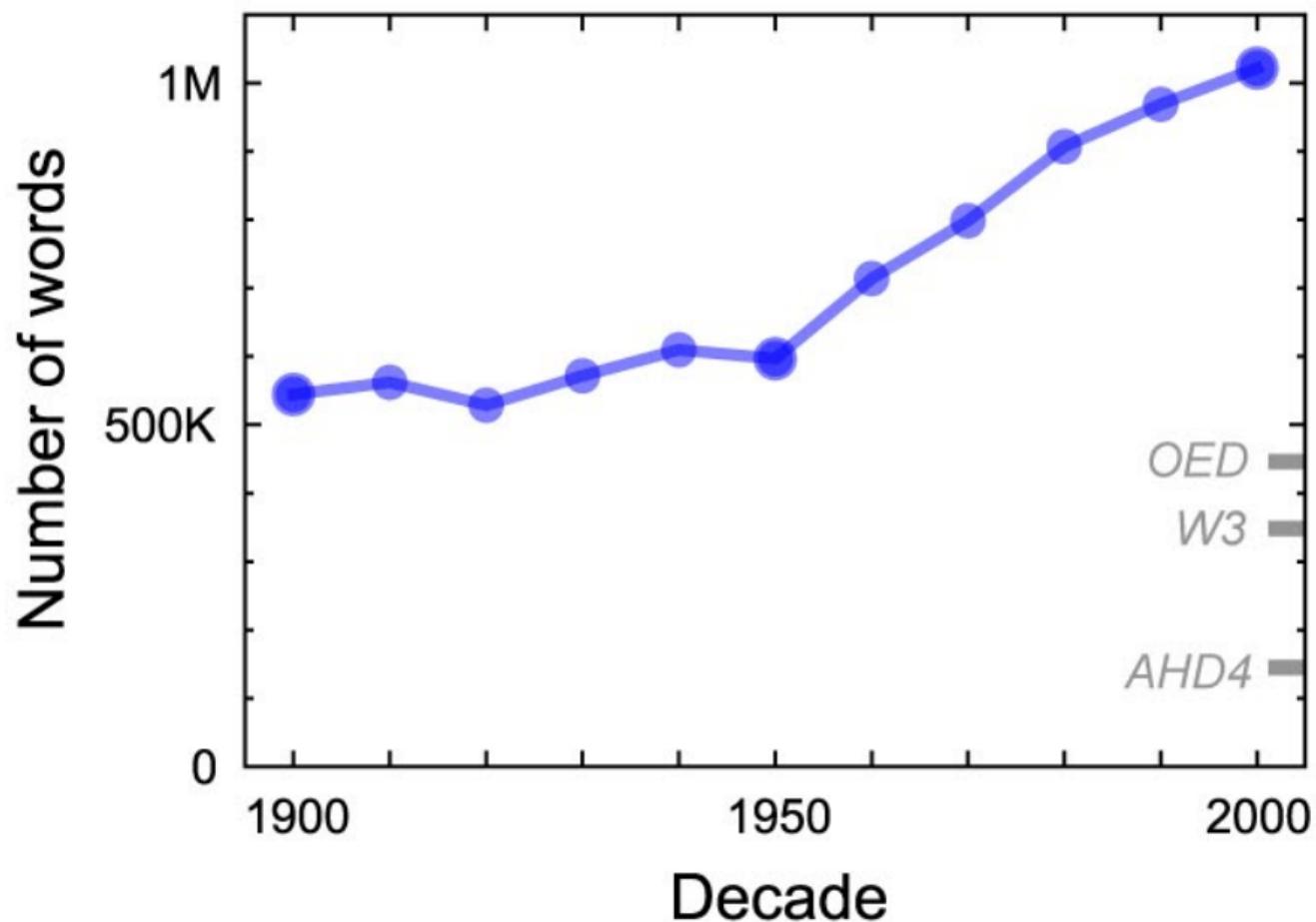
# science

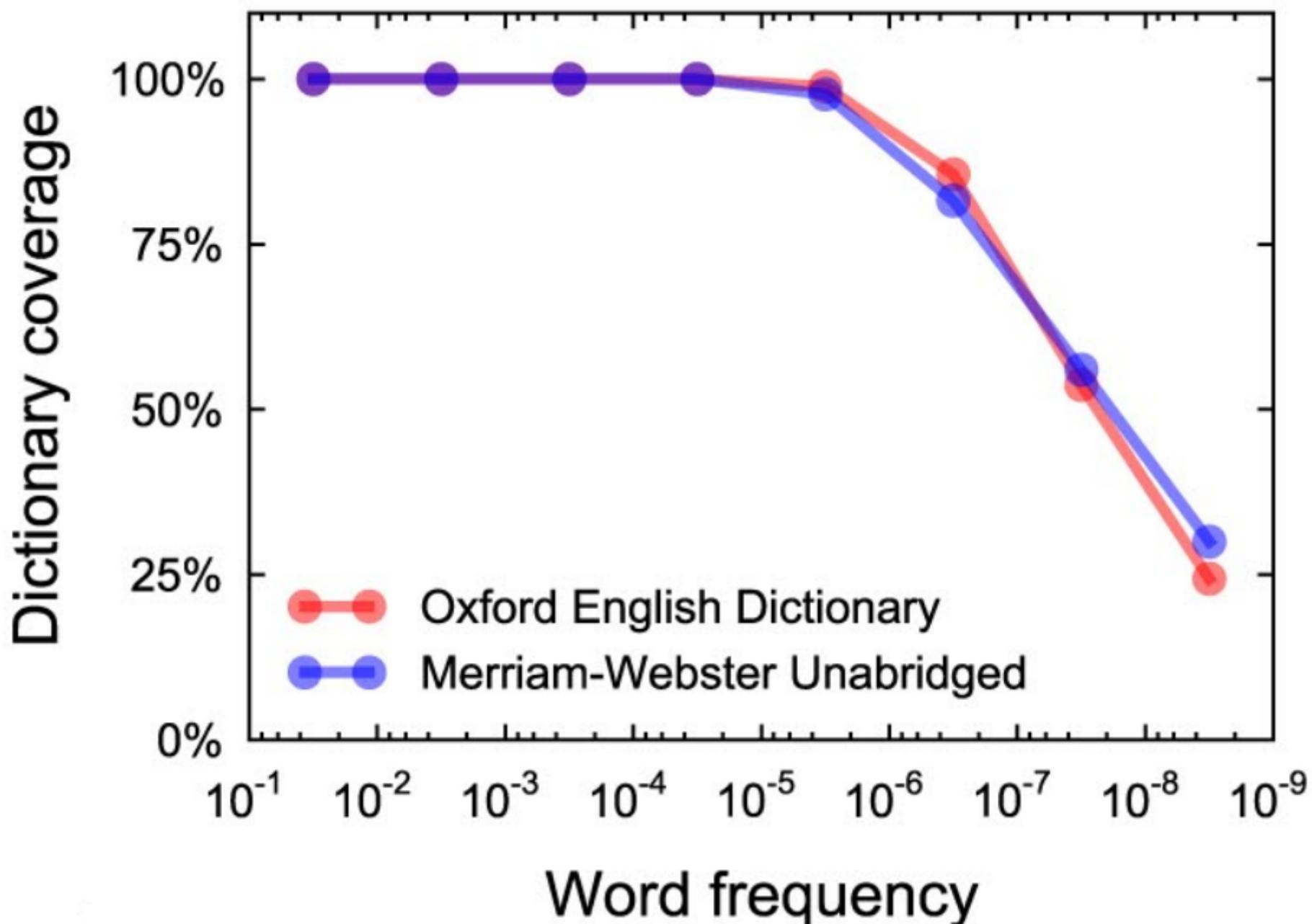
В тот же день открылся Google Ngram Viewer

[books.google.com/ngrams](https://books.google.com/ngrams)

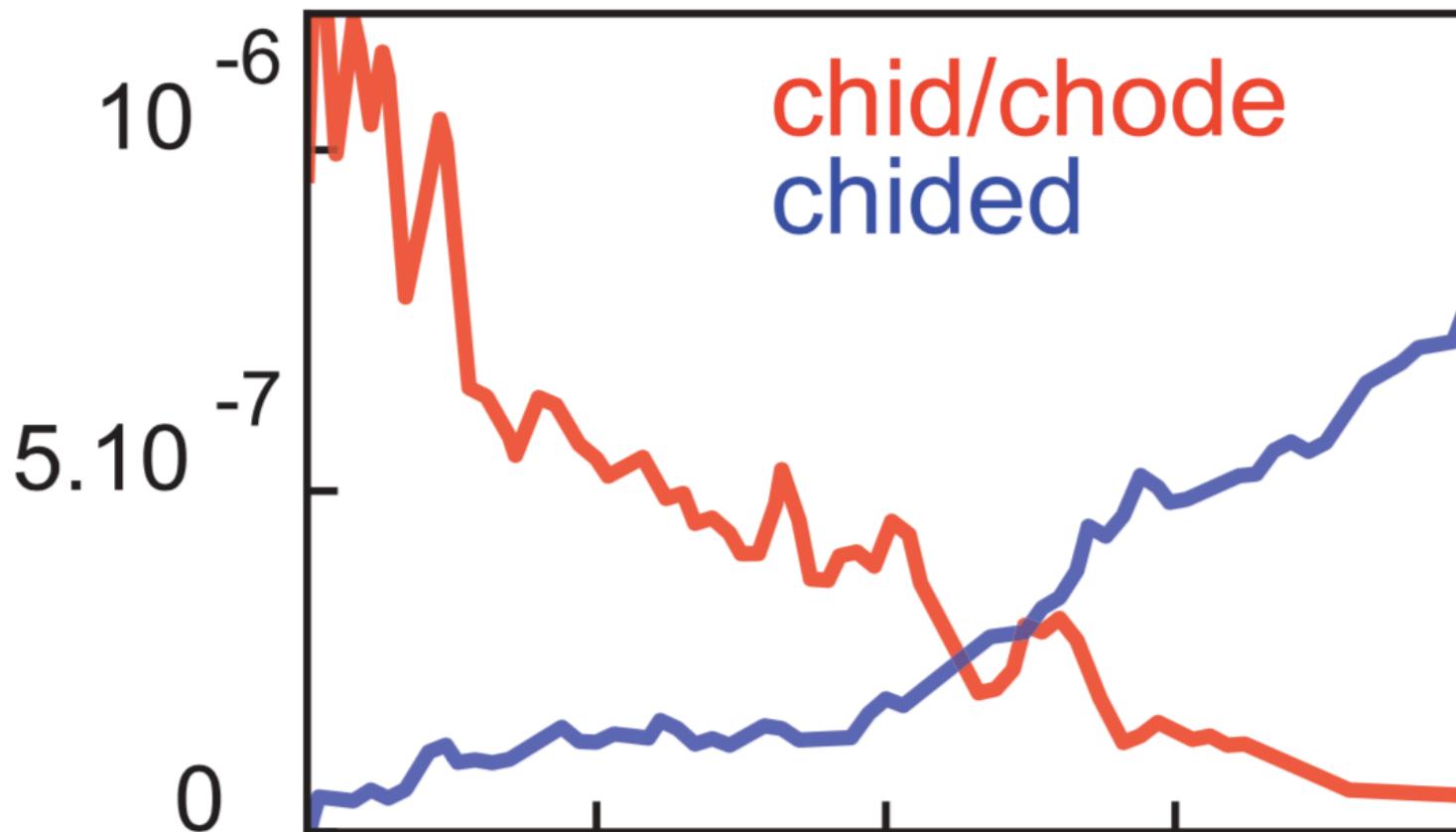
Начали с лингвистики и  
лексикографии

# Размер лексикона

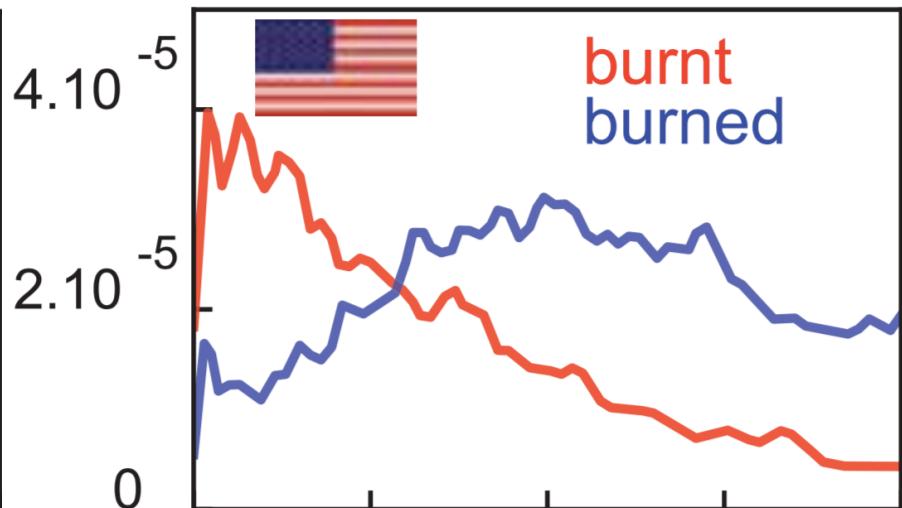
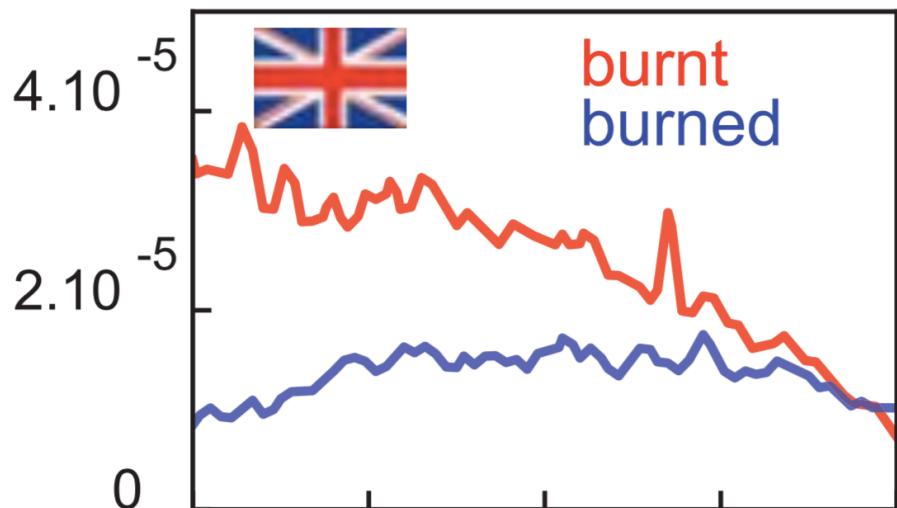




# Регуляризация английских глаголов



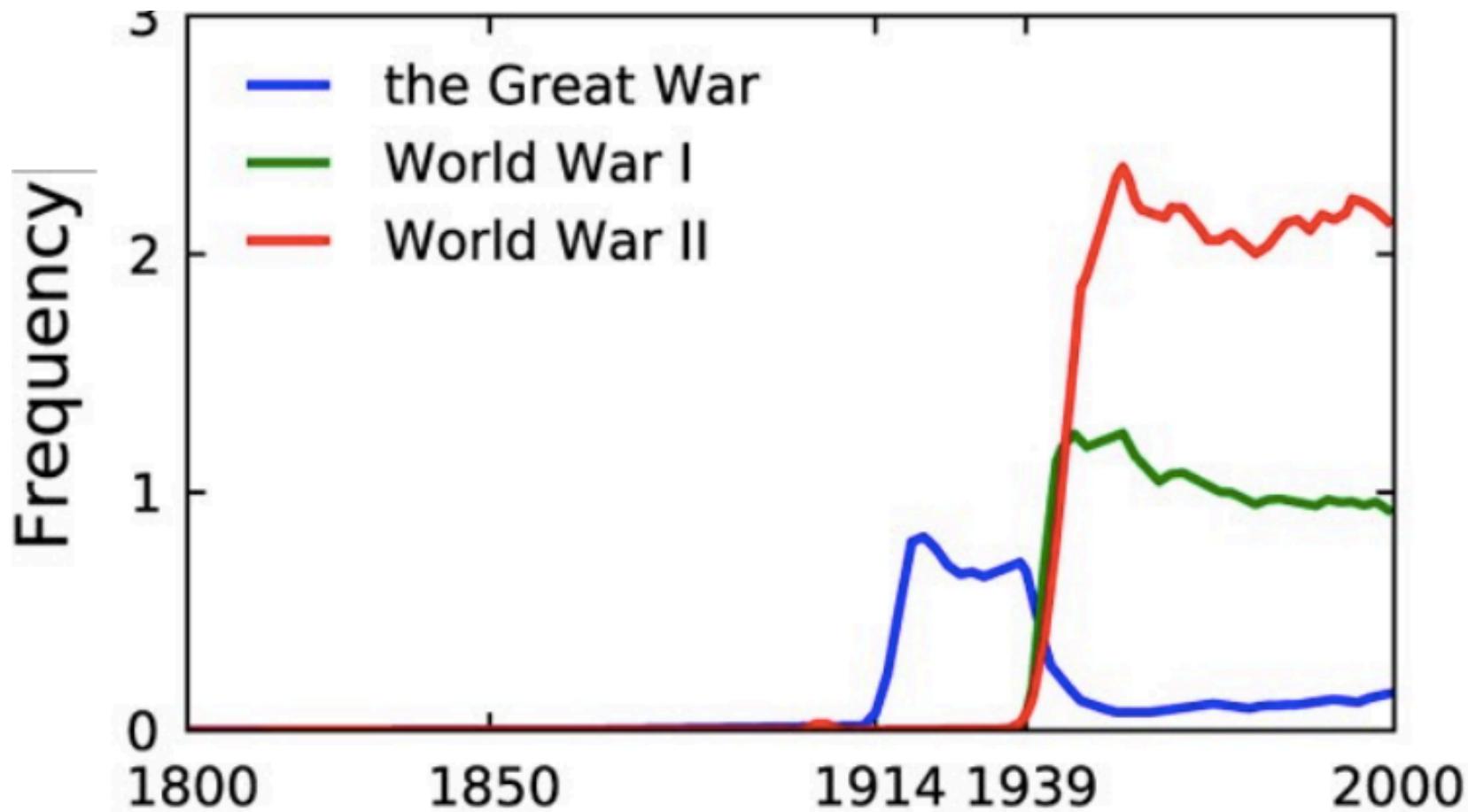
... происходит по-разному на разных сторонах Атлантики



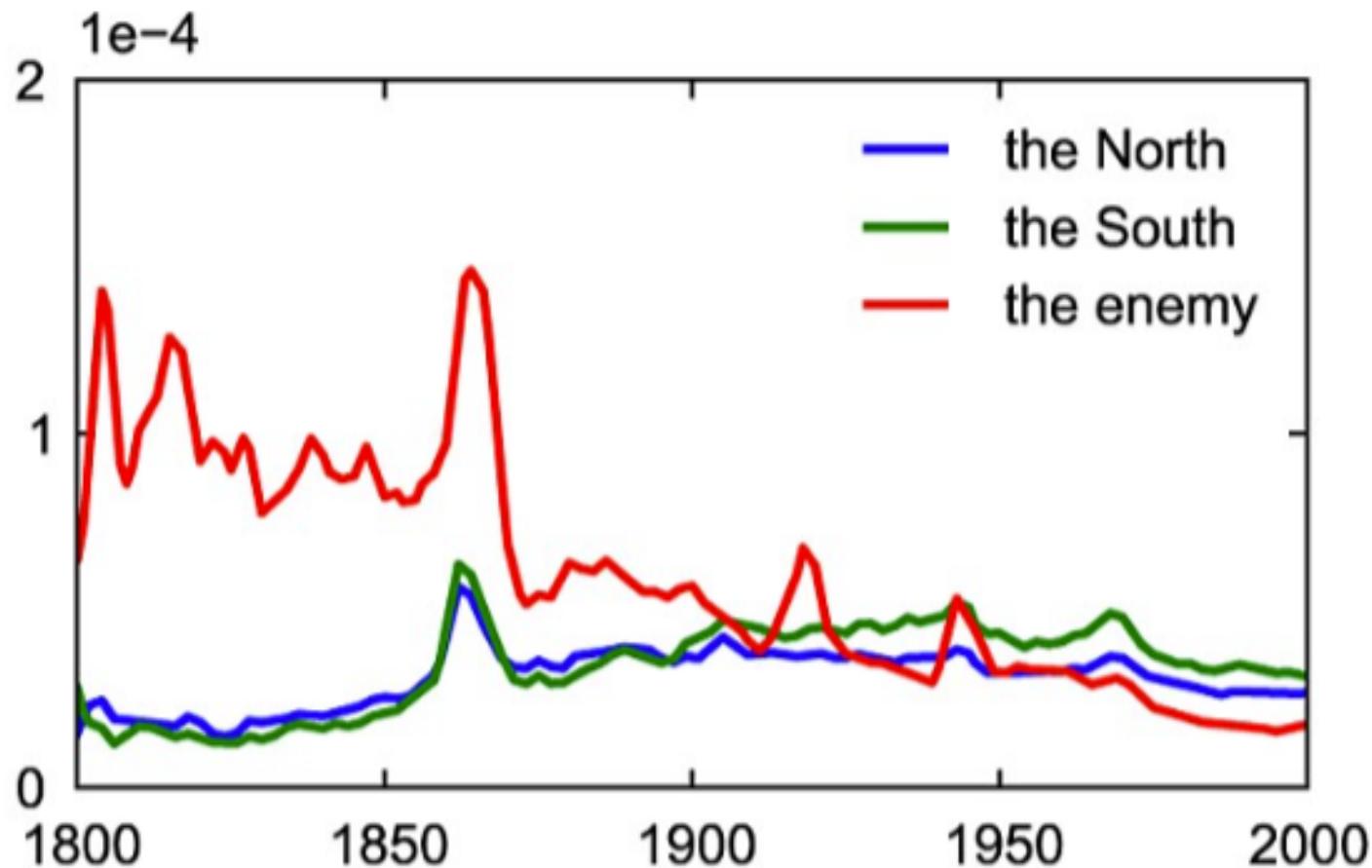
Michel J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 2011. Vol. 331, № 6014. P. 176–182.

# Исторические события

# (уже не) Великая война

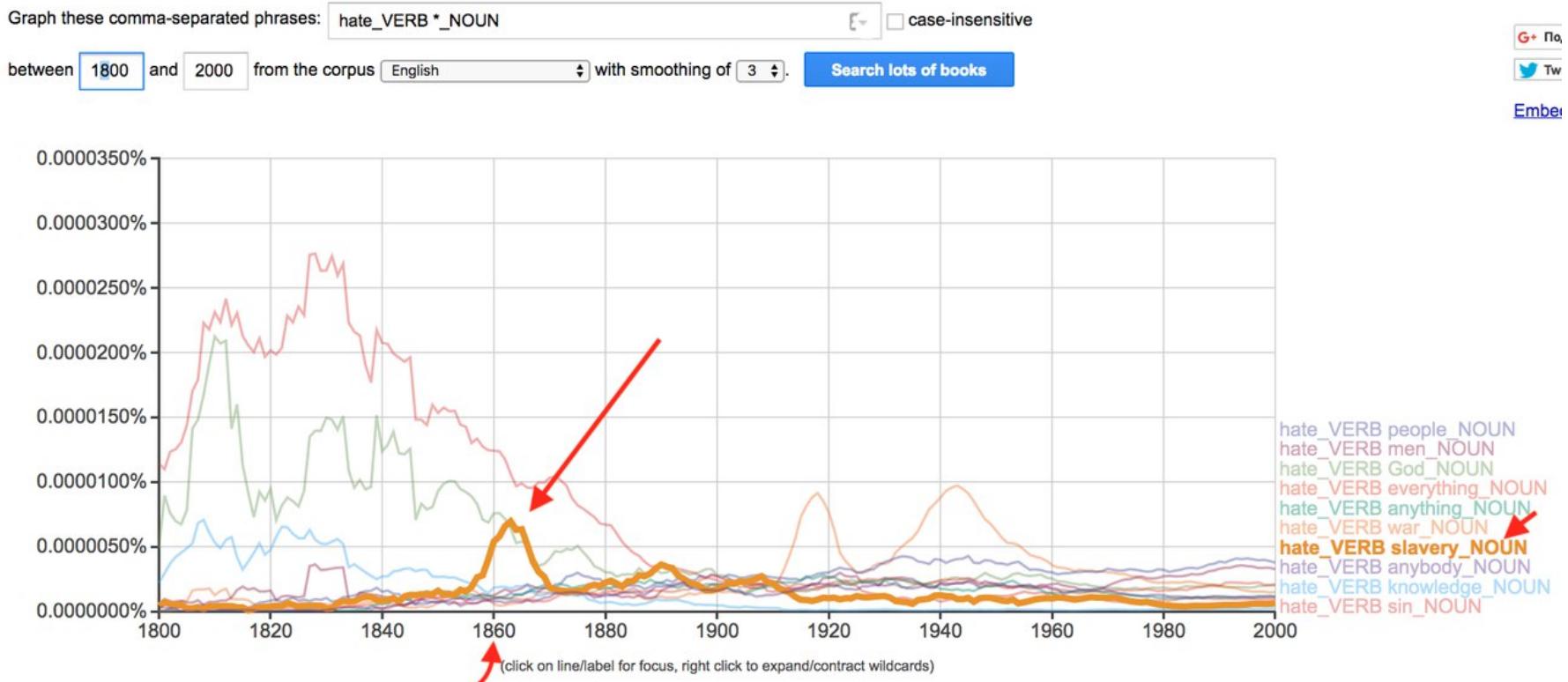


# Еще войны



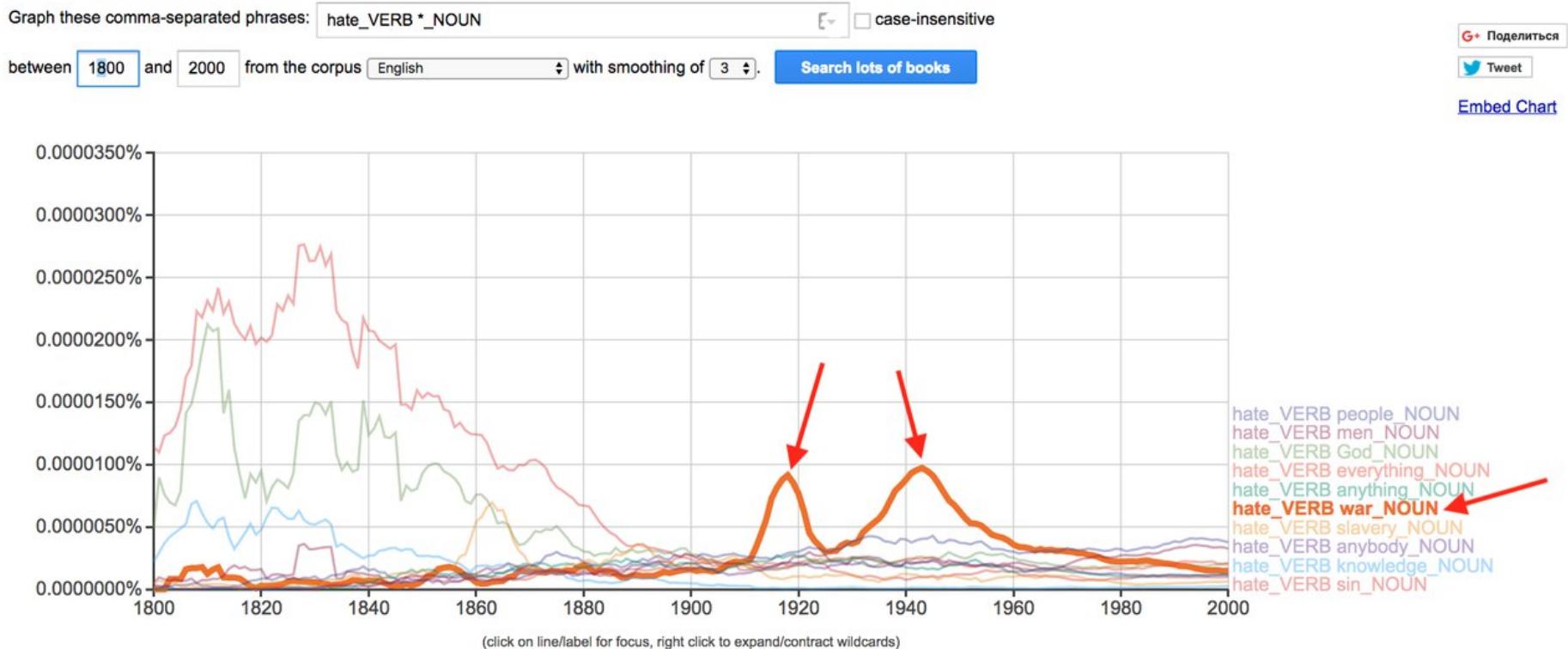
# Мой пример: ненавидеть + сущ

## Google Books Ngram Viewer



# Мой пример: ненавидеть + сущ

## Google Books Ngram Viewer



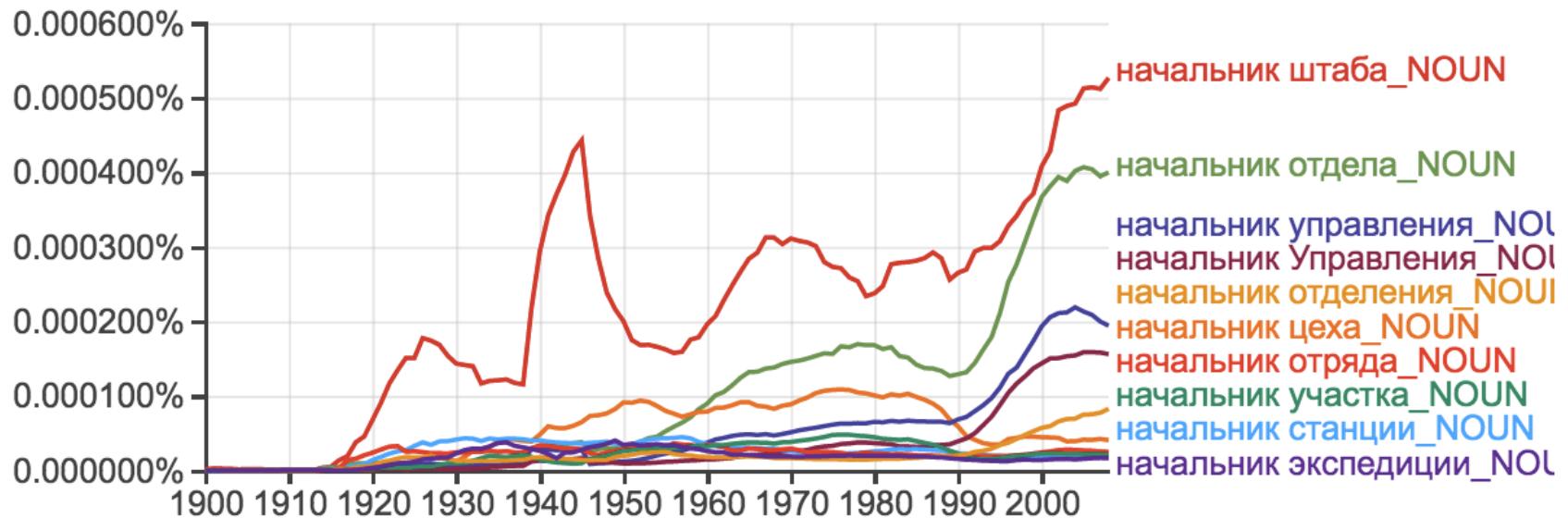
# Возможности и синтаксис поискового запроса:

- Форма слова (начальником)
- Несколько форм ((начальника + начальнику))
- Все формы одной лексемы (наука\_INF)
- С учетом/без учета регистра
- Часть речи (hate\_VERB, hate\_NOUN)
- Wildcards (love \*\_NOUN)
- Синтаксические зависимости (транспорт=>\*\_ADJ)
- Поиск в разных корпусах:  
("internet:eng\_2012,интернет:rus\_2012")

Подробнее см. тут: [books.google.com/ngrams/info](https://books.google.com/ngrams/info)

Более подробно про лингв.разметку:  
[aclweb.org/anthology/P12-3029](https://aclweb.org/anthology/P12-3029)

# Мой пример (русский): начальник штаба



# Уничтожение оппозиции в ВКП(б) и цензура в СССР

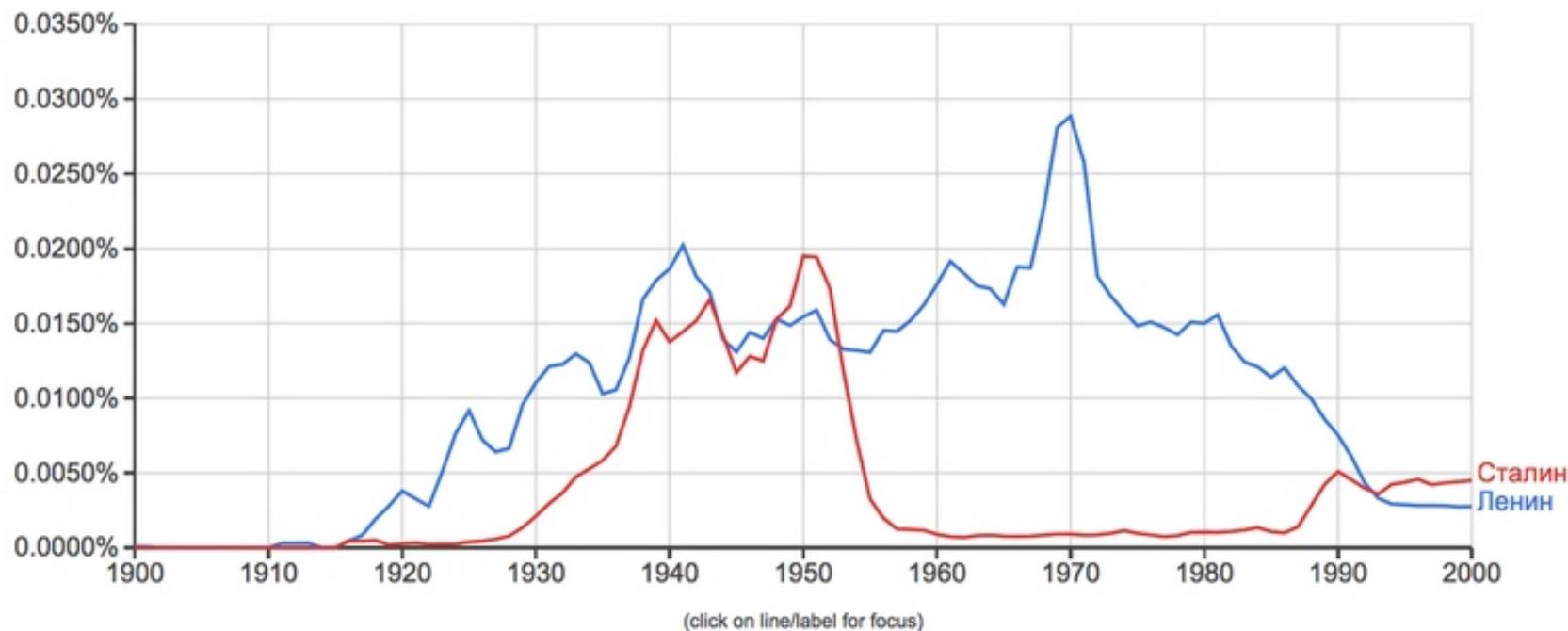


# Мой пример: Ленин/Сталин

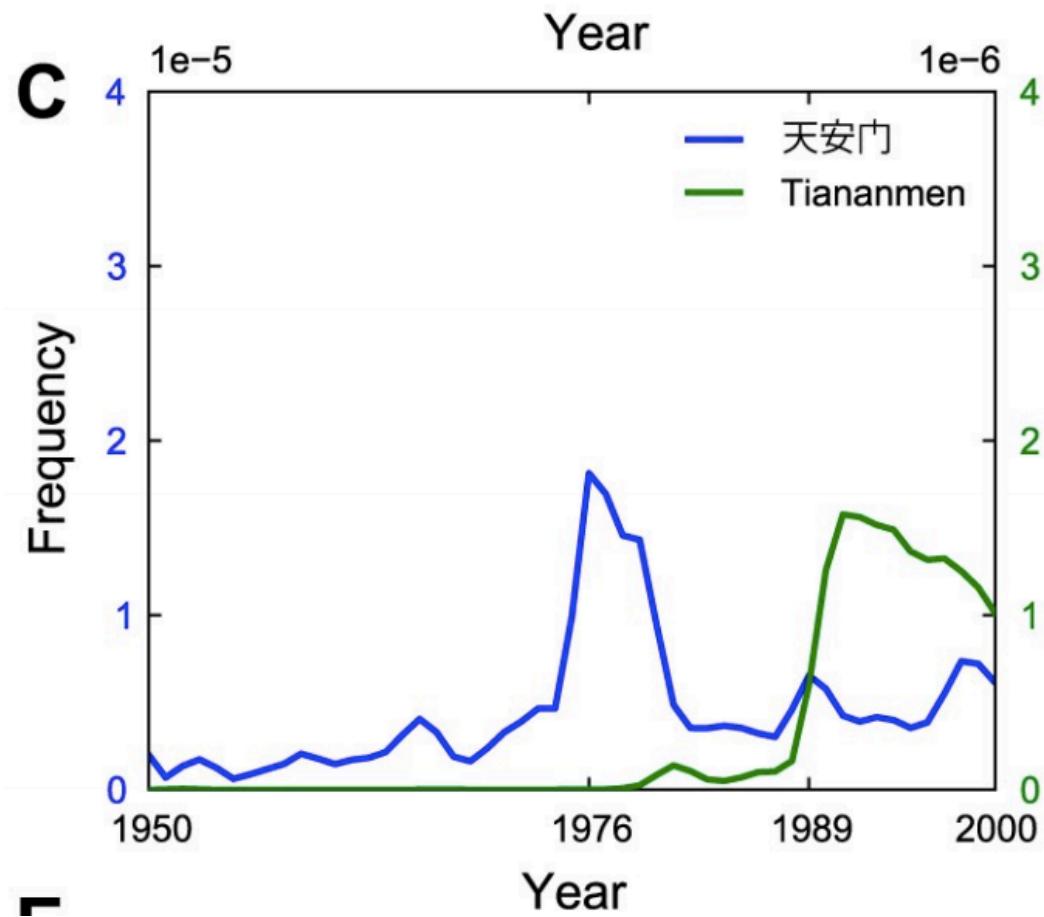
## Google Books Ngram Viewer

Graph these comma-separated phrases:    case-insensitive

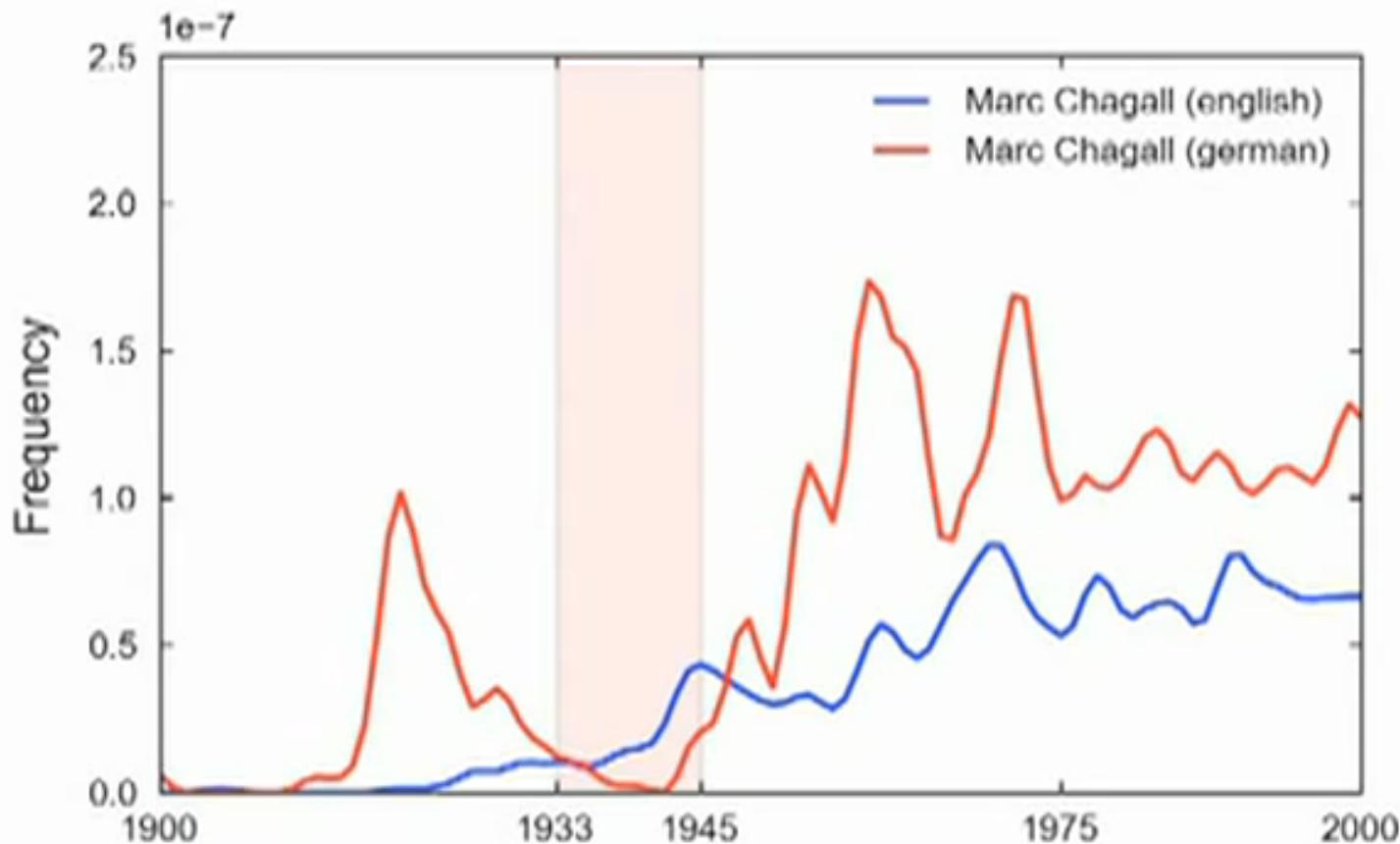
between  and  from the corpus  with smoothing of



# Цензура в Китае

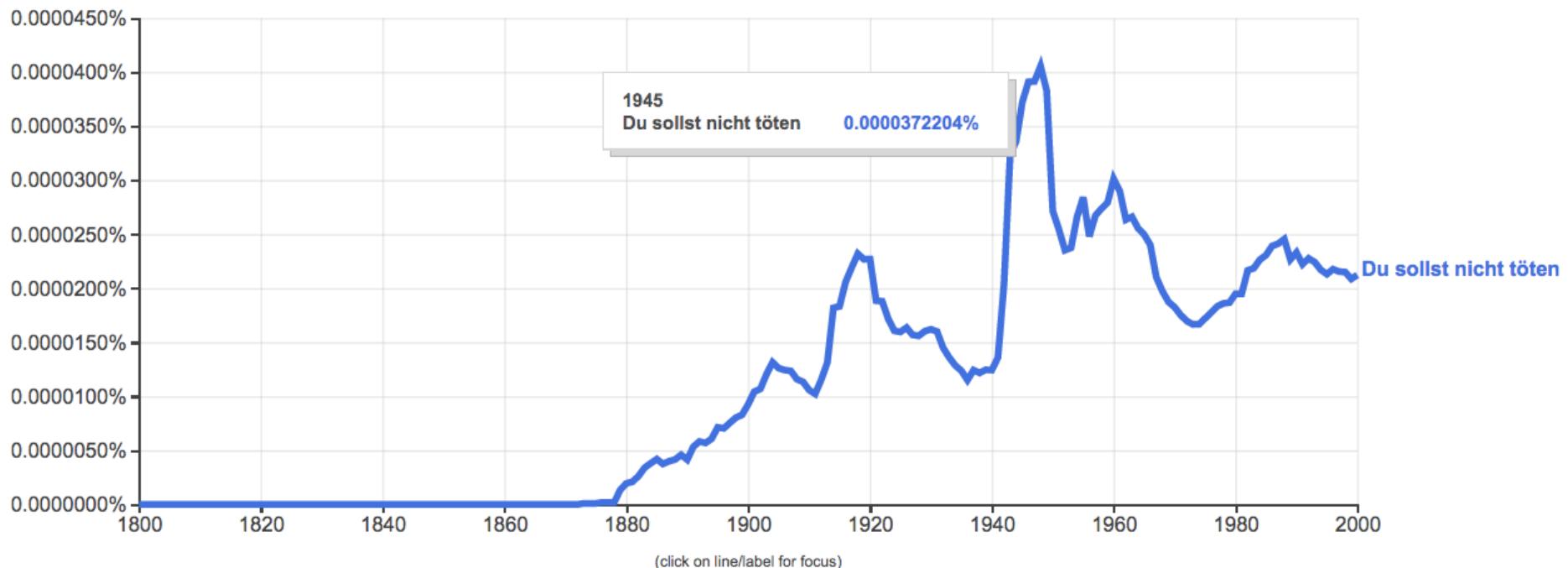


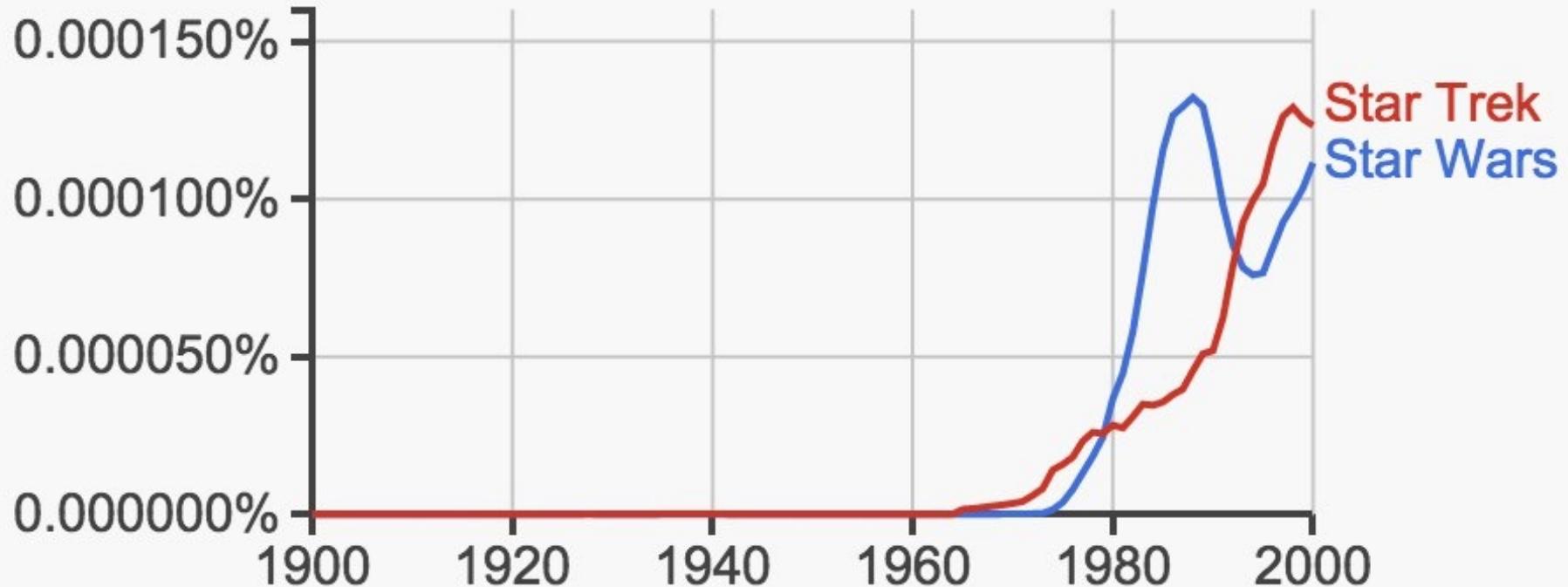
# Цензура в Германии?



# И еще про Германию

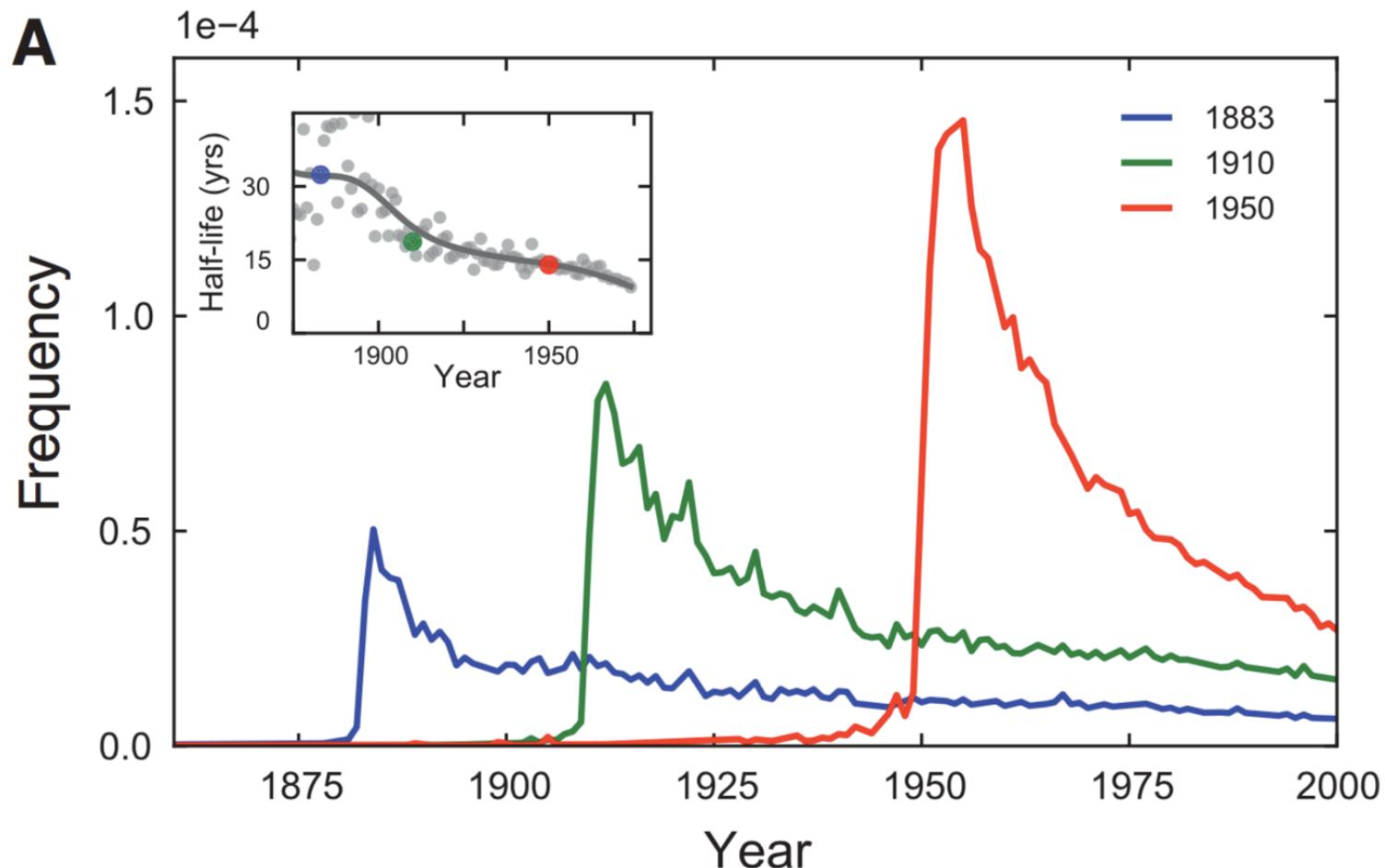
Попробуйте Du sollst nicht töten в немецком корпусе:



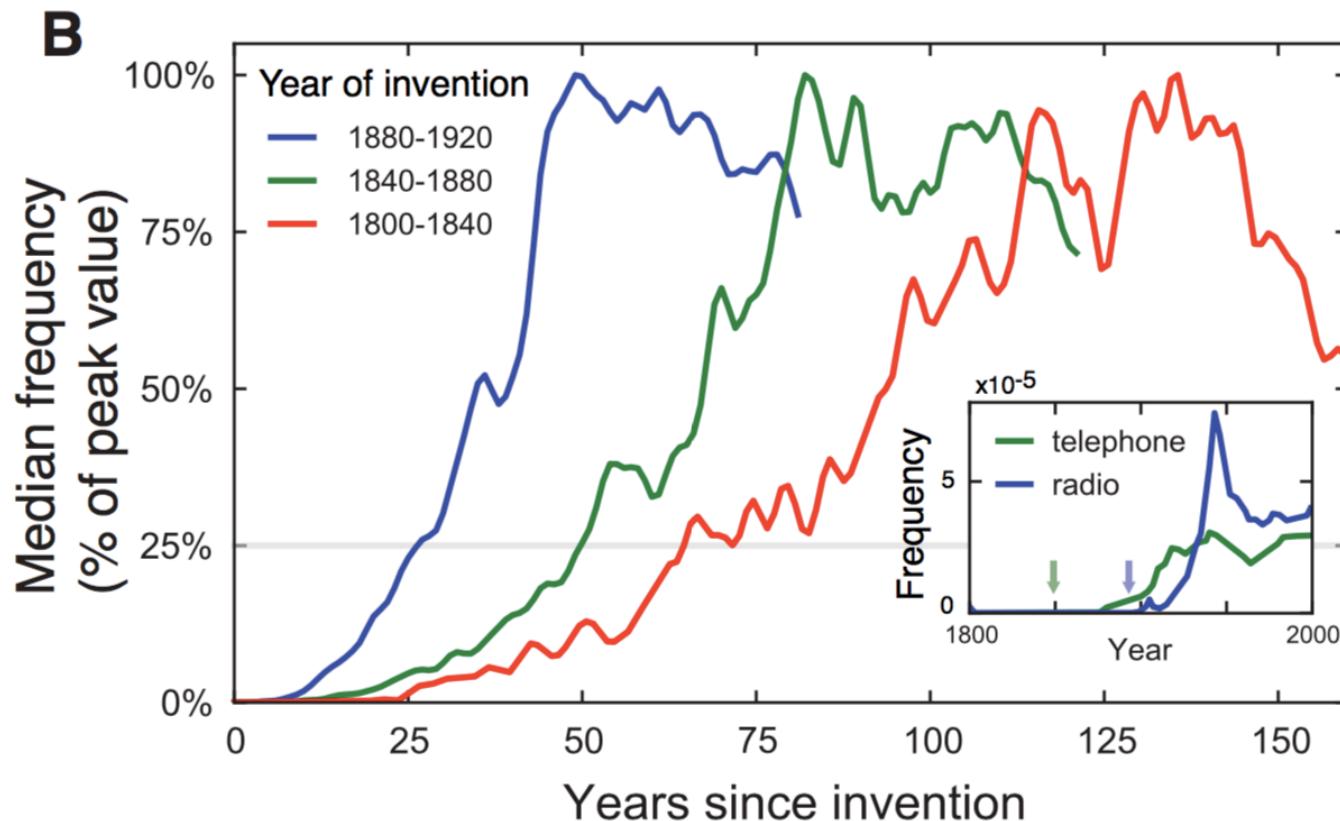


Общественные и  
культурные тренды

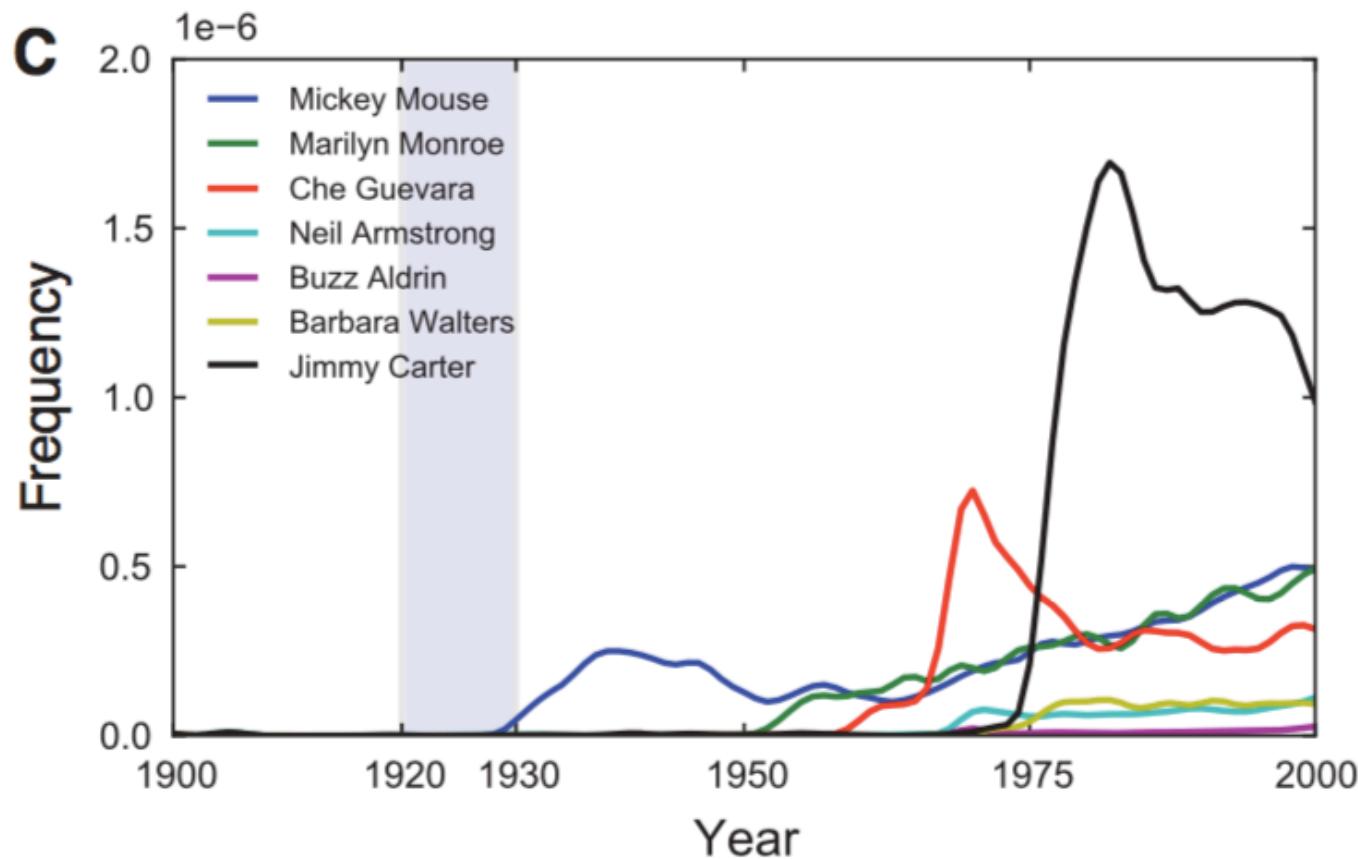
# Ускорение истории?



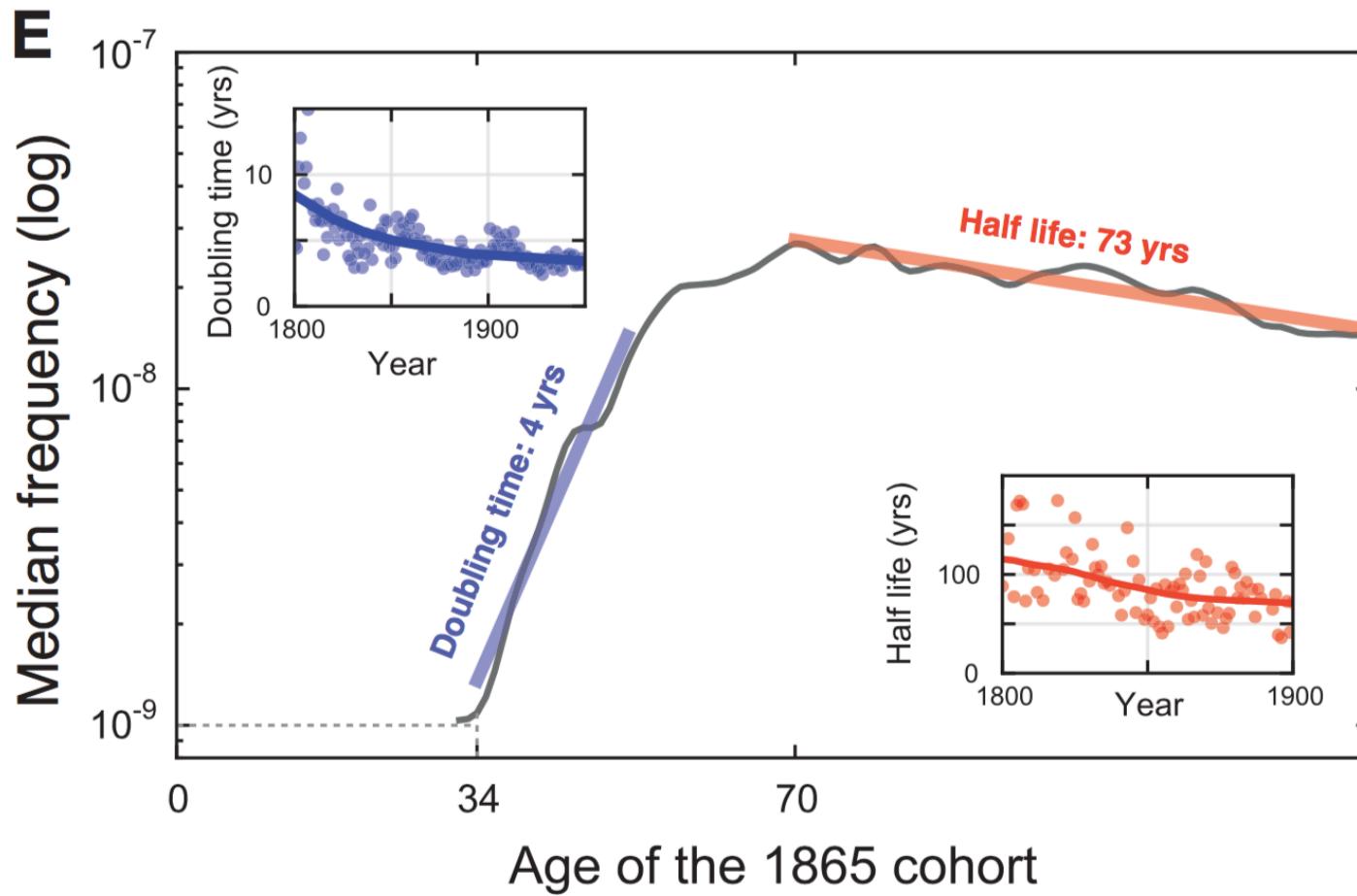
# Ускорение технического прогресса



# Слава и забвение



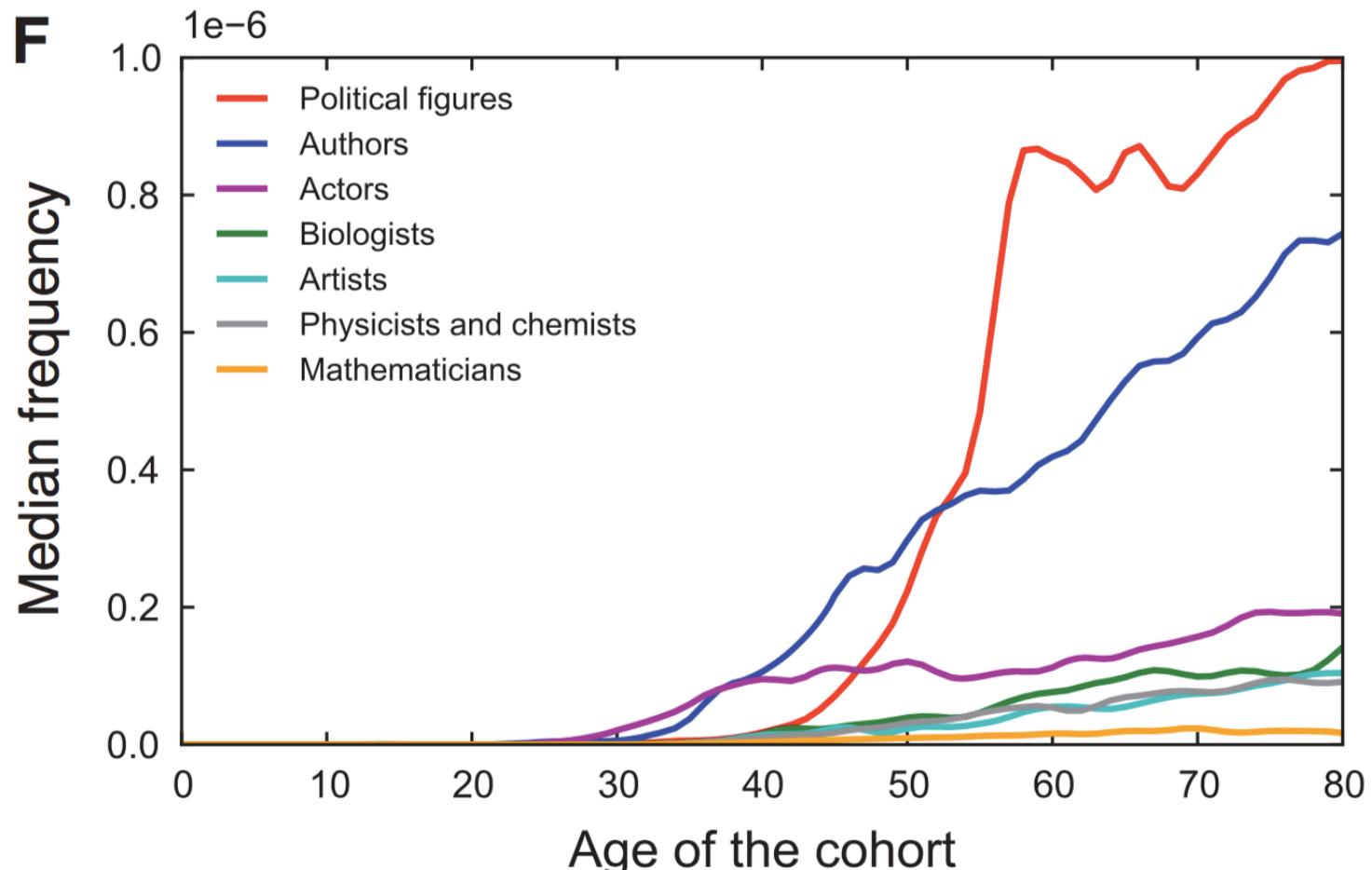
# Слава и забвение



In the future, everyone will  
be famous for 7.5 minutes”

– Whatshisname

# Слава и забвение



# Мой пример:



Чем хороша  
культуромика в Google  
Books?

Обсудим

# Чем хороша культуромика в Google Books? Мои пункты:

- Очевидно, что некоторые явления истории/культуры/языка там действительно отражаются
- Воспроизводимость исследований
- Опора на большие агрегированные данные, тысячи разнородных свидетельств
  - Ни одно свидетельство не весит слишком много

# Последний пункт спорный

- A primary issue is that the corpus is in effect a library, containing one of each book. A single, prolific author is thereby able to noticeably insert new phrases into the Google Books lexicon, whether the author is widely read or not.

Pechenick E. A., Danforth C. M., Dodds P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution

- <...>when content is treated as digital information, a flattening of the structures of knowledge occurs which means that ‘thirteen hundred words of gibberish and the Declaration of Independence are digitally equivalent’ (Brown and Duguid, 2002, p. xiii).

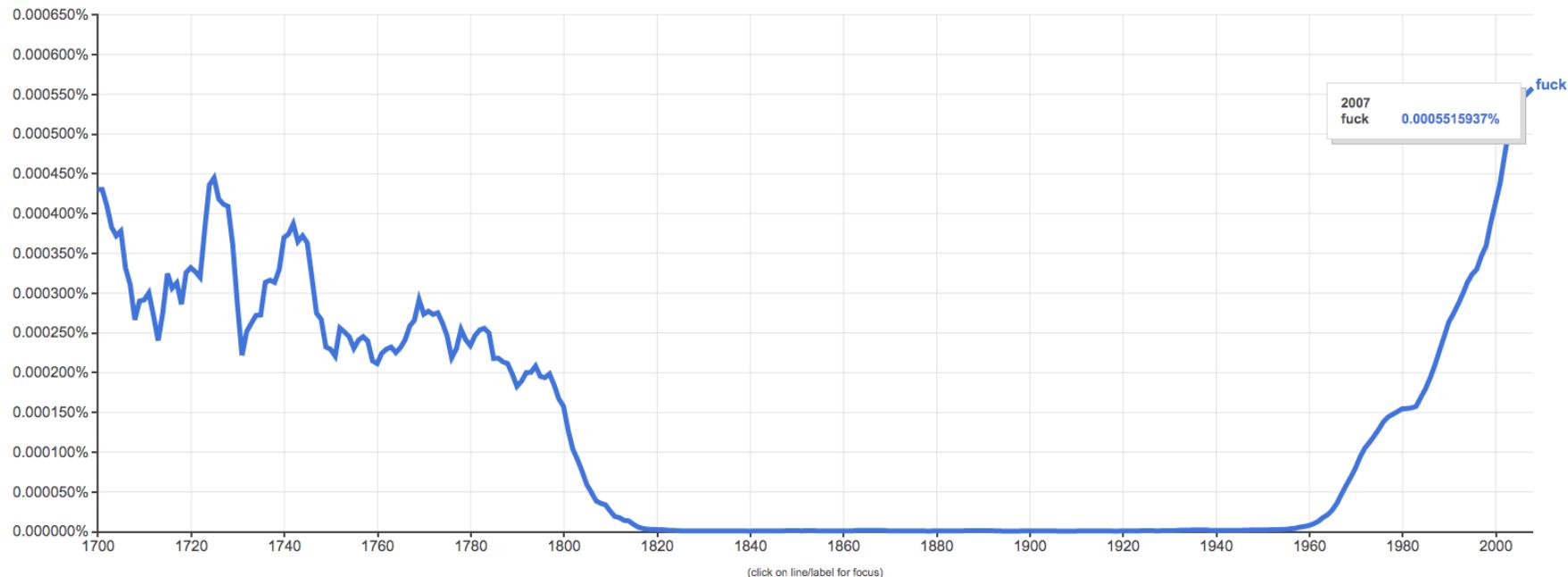
Gooding P. (2012). Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3), 425–431.

Чем плоха  
культуромика в Google  
Books?

# 1. Технические проблемы

# Google Books Ngram Viewer

Graph these comma-separated phrases:   case-insensitive  
between  and  from the corpus  with smoothing of



59. The sucking in of Air through a Quill is done in  
the same manner as Respiration ; for it is the same as if our  
Mouth were as long as the Quil.

60. If we try to **suck** a heavy Liquor through a Quill  
dipped into it, we ought to find so much the greater Dif-  
ficulty as the Quantity of Liquor we make to rise is greater ;  
because this Liquor pressing by its Weight upon the ex-  
ternal Air which endeavours to raise it in the Quill, <sup>1</sup> hin-  
ders it from impelling and assisting the Air which is in the  
Lungs, so much as it usually does ; by which means the  
Air in the Lungs is weakned, and has just so much less  
Force to thrust the Parts of the Body outwards, than the  
Air which is applied to the external Surface of the Body  
has to thrust them inwards, as the Liquor which is caused  
to rise in the Quill is heavier.

61. I shall finish what I have to say concerning these  
Sort of Motions, with explaining that Swelling which  
Surgeons make in the Flesh, by the Application of Cup-

*59. How it is  
that we suck  
in Air.*

*60. Why it is  
more difficult  
to suck a  
heavy Li-  
quor.*

*61. Concern-  
ing the Use  
of Cupping-  
Glasses.*

Был ли вайфай в XVIII  
веке?



Добавить в библиотеку

Написать отзыв

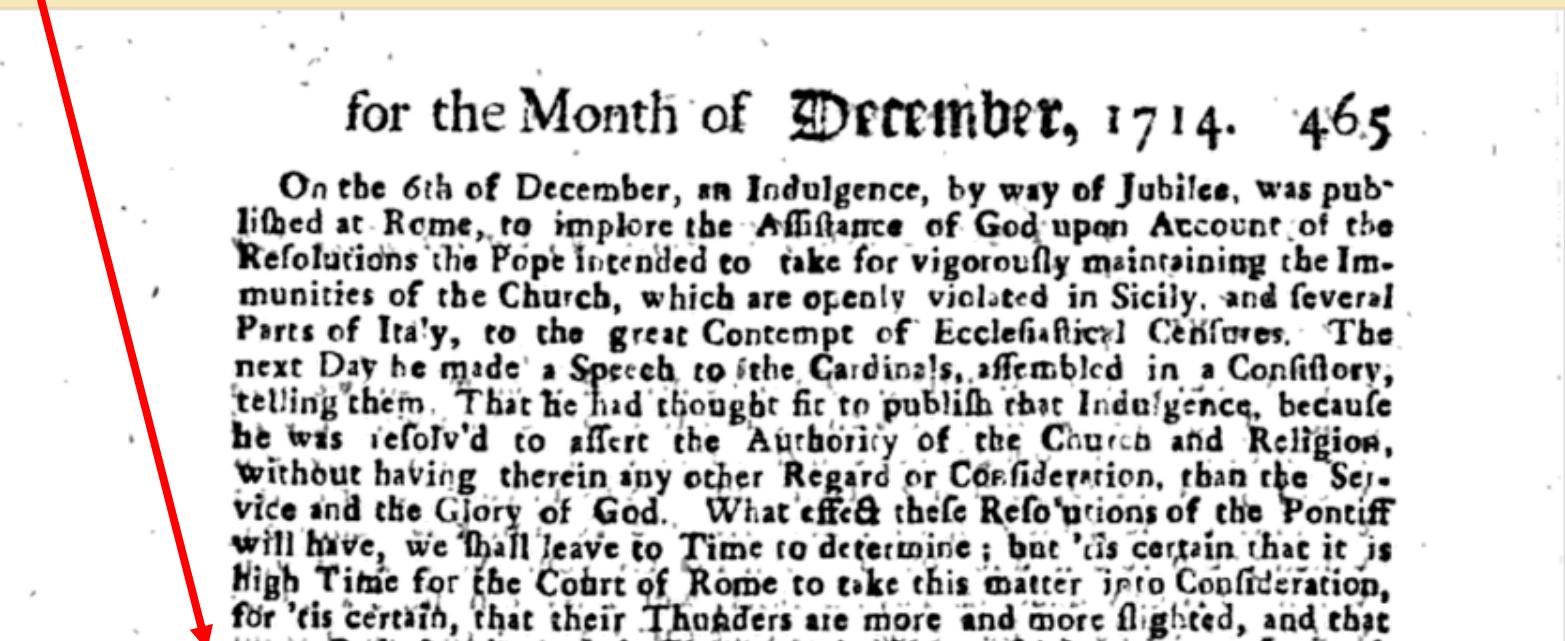
Стр. 465

В этой книге: результат 2 из 2 по запросу ""wifi"" - [Предыдущая](#) [Следующая](#) - [Просмотреть все](#)[Закрыть панель поиска](#)

## for the Month of December, 1714. 465

On the 6th of December, an Indulgence, by way of Jubiles, was published at Rome, to implore the Assistance of God upon Account of the Resolutions the Pope intended to take for vigorously maintaining the Immunities of the Church, which are openly violated in Sicily, and several Parts of Italy, to the great Contempt of Ecclesiastical Censures. The next Day he made a Speech to the Cardinals, assembled in a Consistory, telling them, That he had thought fit to publish that Indulgence, because he was resolv'd to assert the Authority of the Church and Religion, without having therein any other Regard or Consideration, than the Service and the Glory of God. What effect these Resolutions of the Pontiff will have, we shall leave to Time to determine; but 'tis certain that it is High Time for the Court of Rome to take this matter into Consideration, for 'tis certain, that their Thunders are more and more slighted, and that every Body laughs at their Excommunications, which were once so much respected and dreaded, that they were sufficient to shake the Thrones of the Greatest Prince of the Christian World. The Church daily laments the loss of these Days, which she calls the Golden Age of the Church, and **will** use all possible Means to bring back those Times, when an insolent Priest trod under Foot a great Emperor; but as their enormous Impudence was the chief Cause of the loss of their extravagant and scandalous Power, we hope that the violent Means they will use to retrieve the same, will prove the occasion of its entire Destruction.

While they are chiefly intent at Rome, on the Recovery of their pretended ancient Rights, they cannot but be surpriz'd at a Report spread throughout Germany, that the Pope is endeavouring to bring about an



# Google Books Ngram Viewer

Graph these comma-separated phrases: Толстой,Достоевский

case-insensitive

between 1880 and 2000 from the corpus Russian with smoothing of 4

**Search lots of books**



Θ  
фрига

Ђ  
Ер.

Ђ  
Ять

I  
И десятеричное

## 2. Проблемы репрезентативности и состава выборки

# Нерепрезентативный состав Google Books

<...> a fundamental flaw in its methodology:  
its reliance on Google Books for its sample.  
Google Books has focused on digitizing  
academic libraries. I would argue that books  
found in academic libraries are not necessarily  
representative of cultural trends across  
society.

Anita Guerini (2011) Analyzing Culture with Google Books: Is It Social Science?

# Нерепрезентативный состав Google Books

- One of the problems with the Google ngram corpus is that really we have no idea what genres are represented in it, or how their relative proportions may vary over time. So it's possible that an apparent decline in the frequency of words for moral values is actually a decline in the frequency of certain genres — say, conduct books, or hagiographic biographies.
- A decline of that sort would still be telling us something about literary culture; but it might be telling us something different than we initially assume from tracing the decline of a word like “fidelity.”

Ted Underwood (2012). How not to do things with words

# Пример

Books that substantively mention memory before 1800 are most often either volumes dedicated to instructing the reader in the arts of mnemonics or examples of monody, a poem dedicated to the memory of specified individual. The frequency of memory compared with other terms in the corpus declines in the first half of the 20th century, only to increase again after 1960 (see Figure 9). Such usage fits well with a familiar story: the cognitive revolution that reintroduced the psychological study of mental states after decades of neglect. However, a click on the “Search in Google Books” function reveals that the increase is also due to the academic study of the “collective memory”

Michael Pettit (2016) Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer

# Непрепрезентативный состав Google Books

<...> the evolution of the corpus throughout the 1900s is increasingly dominated by scientific publications rather than popular works. We have shown that even the first data set specifically labeled as fiction appears to be saturated with medical literature.

Pechenick E. A., Danforth C. M., Dodds P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution

<...> the corpus includes very few examples of the cheap but largely ephemeral dime novels and pulp fiction that were widely popular with American working-class readers (Egnal, 2013).

# Ограничение определенным типом культурных объектов

It <...> excludes the visual culture, oral traditions, and symbolic rituals that have been at the heart historical and anthropological interpretations of culture for decades.

Michael Pettit (2016) Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer

# Смешение разных культур без возможности разделить

It assumes that cultures are undifferentiated and totalizing wholes whose self-understanding is driven by the thought leaders who make it into print. The tool provides no ways of disaggregating distinct subcultures. In short, it risks replicating a version of the past that 50 years of social history and cultural studies have debunked (Hall, 1981). Instead, we need to recognize that the Google Books corpus does not represent culture as an undifferentiated whole, but foregrounds the culture of particular subgroups.

Michael Pettit (2016) Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer

# Диахронические диспропорции Google Books

- The oldest works were published in the 1500s.
- The early decades are represented by only a few books per year, comprising several hundred thousand words.
- By 1800, the corpus grows to 60 million words per year;
- by 1900, 1.4 billion;
- and by 2000, 8 billion.

Поэтому авторы статьи в Science в основном приводили примеры про XIX – XX вв.

We survey the vast terrain of “culturomics”, focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000.

[Мы исследуем просторы культуромики, фокусируясь на лингвистических и культурных феноменах, отраженных в английском языке между 1800 и 2000 гг.]

Michel J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 2011. Vol. 331, № 6014. P. 176–182., перевод мой

3. Частотности слов —  
слишком грубая и  
примитивная метрика

# СЛОВА МЕНЯЮТ ЗНАЧЕНИЯ

If we crowdsource “leadership” using twenty-first-century reactions on Mechanical Turk, for instance, we’ll probably get words like “visionary” and “professional.” “Loud-voiced” probably won’t be on the list — because that’s just rude. But to Homer, there’s nothing especially noble about working for hire (“professionally”), whereas “the loud-voiced Achilles” is cut out to be a leader of men, since he can be heard over the din of spears beating on shields

<...>

The laws of perspective apply to history as well. We don’t have an objective overview; we have a position in time that produces its own kind of distortion and foreshortening.

Ted Underwood (2012). How not to do things with words

# Слова меняют значения

Searching for individual words is risky. A word's meaning is as likely to change over time as its frequency. For example, I am not convinced that "God" meant precisely the same thing in 1800 versus 1900 versus 1950

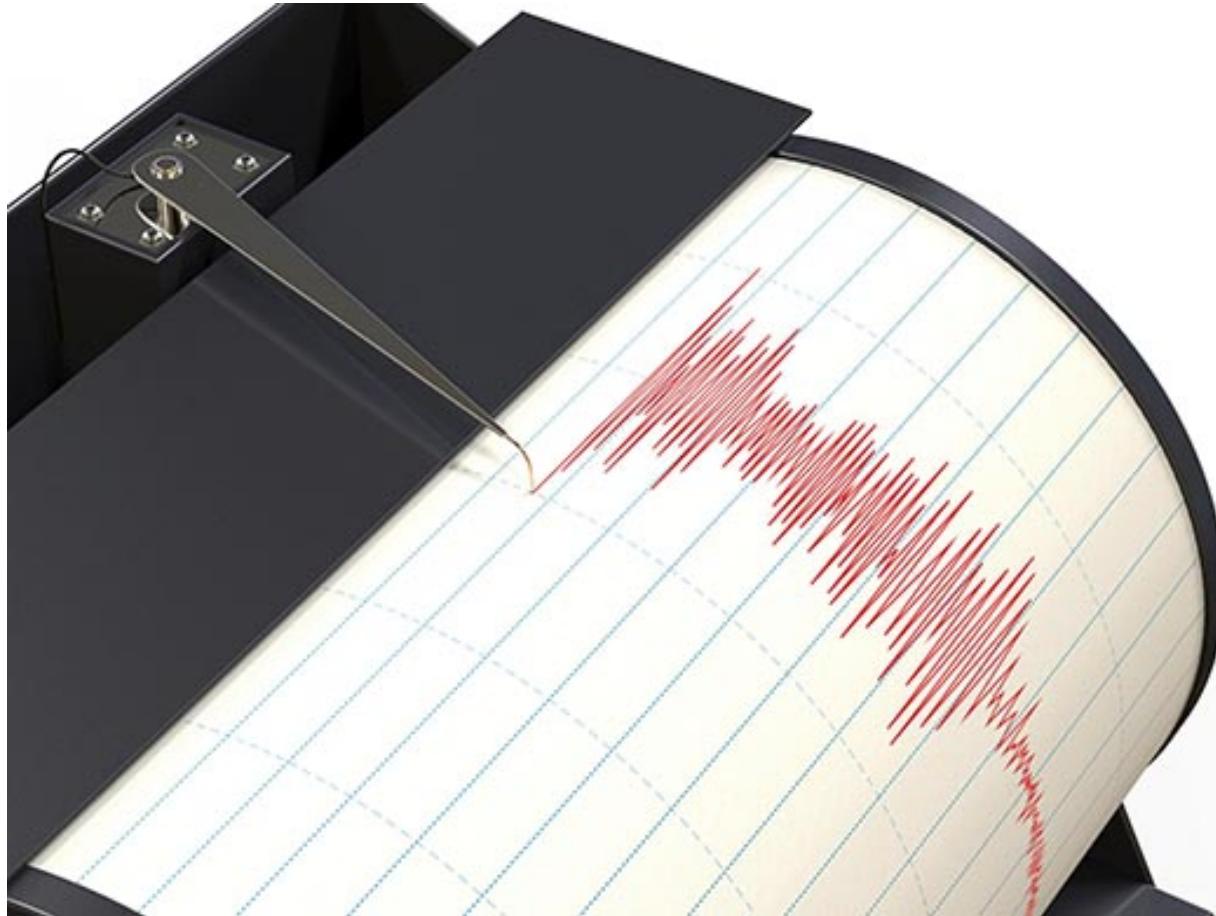
Michael Pettit (2016) Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer

# Корпусный поиск — это как анализ данных сейсмографа

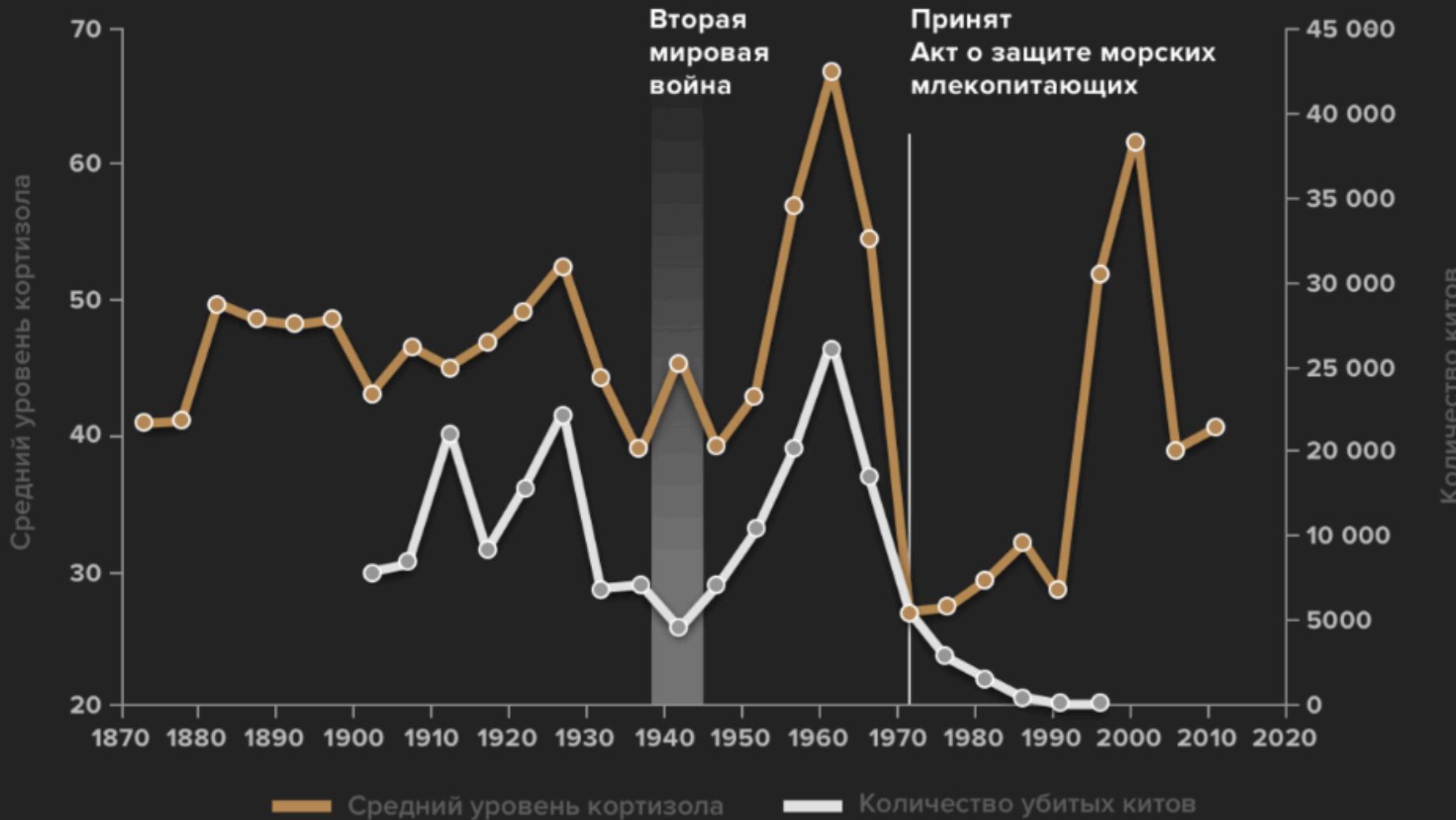
At its heart, the use of word frequency with a reasonably sized (if problematic) data set simply provides one more form of evidence to be added to all the rest. Knowing that the term 'electricity' peaks between 1870 and 1900 is useful evidence, but does not provide either an explanation for why, or a description of how it is being used.

Tim Hitchcock (2011) [Culturomics, Big Data, Code Breakers and the Casaubon Delusion](#)

Корпусный поиск — это как  
анализ данных сейсмографа



# Уровень гормона стресса у китов в XX веке



Коктейль из  
очевидного и  
неинтерпретируемого

# Инструмент не анализирует

These large-scale visualisations of language may be the raw material of history, the basis for an argument, the foundation for a narrative, the evidence put in the appendix in support of a subtle point, but they do not serve as a work of history.

Tim Hitchcock (2011) [Culturomics, Big Data, Code Breakers and the Casaubon Delusion](#)

# Инструмент не анализирует

<...> the humanities will always have to rely on human analysis to some degree. <...> When texts are deconstructed to the extreme of granularity, and interactions become mediated by automated tools, there is no book. Instead, there is merely a massive corpus of words which carry no great epistemological significance.

Gooding P. (2012). Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3), 425–431

# Практика

Поиските сами интересные примеры в Google Ngram  
Viewer

# Задания

(по умолчанию диапазон 1900 - 2000)

1. Сравните частотности слов *science* и *religion* в англоязычных текстах
2. А теперь *наука* и *церковь* — в русскоязычных
3. Разберитесь, как построить график, который показывал бы разницу между частотностью упоминания Ленина и Сталина (т.е. чтобы линия была **одна**; выше нуля, когда Ленина больше, ниже — когда больше Сталина)
4. Постройте **один** график с **четырьмя** линиями, которые отражали бы частотность слов *color* и *colour* в двух корпусах — британского и американского английского
5. Постройте график, который бы отражал **отношение** частотности словосочетания *в Украине к сумме частотности* словосочетаний *в Украине* и *на Украине*
6. Постройте график частотности слов *телефон, телевизор, компьютер, пейджер, телеграф* но частотность пейджа пусть будет умножена на 20