Эрец Эйден
Жан-Батист Мишель

# Неизведанная территория

Erez Aiden
Jean-Baptiste Michel

# UNCHARTED:

## BIG DATA AS A LENS ON HUMAN CULTURE

Эрец Эйден
Жан-Батист Мишель

# Неизведанная территория :

Как «большие данные» помогают раскрывать тайны прошлого и предсказывать будущее нашей культуры

Издательство АСТ
Москва

УДК 141.339
ББК 86.42
Э 11

Серия «Наука XXI век»

Erez Aiden and Jean-Baptiste Michel

UNCHARTED:
BIG DATA AS A LENS ON HUMAN CULTURE

*Перевод с английского Павла Миронова*
*Дизайн обложки: студия OpenDesign*

*Печатается с разрешения авторов и литературного агентства Brockman, Inc.*

*Фото Эреца Эйдена © Eliza Grinnel*
*Фото Жана-Батиста Мишеля © Bret Hartman*

**Эйден, Эрец.**

Э 11 Неизведанная территория : Как «большие данные» помогают раскрывать тайны прошлого и предсказывать будущее нашей культуры / Эрец Эйден и Жан-Батист Мишель; пер. с англ. П. Миронова — Москва: Издательство АСТ, 2016. — 351, [1] с. — (Наука XXI век).

ISBN 978-5-17-088935-8

Насколько велики на самом деле «большие данные» – огромные массивы информации, о которых так много говорят в последнее время? Вот наглядный пример: если выписать в линейку все цифры 0 и 1, из которых состоит один терабайт информации (вполне обычная емкость для современного жесткого диска), то цепочка цифр окажется в 50 раз длиннее, чем расстояние от Земли до Сатурна! И тем не менее, на «большие данные» вполне можно взглянуть в человеческом измерении. Эрец Эйден и Жан-Батист Мишель – лингвисты и компьютерные гении, создатели сервиса Google Ngram Viewer и термина «культуромика», показывают, каким образом анализ «больших данных» помогает исследовать трудные проблемы языка, культуры и истории.

УДК 141.339
ББК 86.42

© Erez Lieberman Aiden and Jean-Baptiste Michel, 2013
© Павел Миронов, перевод, 2014
© Издание на русском языке AST Publishers, 2016

*Моему папе, который всегда верил,*
*что я умею считать*
— ЭРЕЦ ЭЙДЕН —

*Моей семье*
— ЖАН-БАТИСТ МИШЕЛЬ —

## Оглавление

## Глава 1
# ЗАЗЕРКАЛЬЕ

Давайте представим, что у нас есть робот, способный прочитать каждую книгу на каждой полке всех крупных библиотек мира. Он может их прочесть невероятно быстро и запомнить каждое прочитанное слово в своей бесперебойно работающей памяти. Чему мы могли бы научиться у такого робота-историка?

Вот вам простой пример, знакомый каждому американцу. В наши дни принято говорить, что южные штаты полны (*are full*, множественное число) южан. Мы также говорим, что северные штаты полны (*are full*) северян или что штаты Новой Англии полны (*are full*) жителями. Однако мы говорим: *the United States is full of citizens* (то есть «США полон жителей», единственное число). Почему мы используем единственное число? Вопрос лежит не только в области грамматики — это, скорее, вопрос нашей национальной идентичности.

После основания Соединенных Штатов Америки основополагающий документ — Статьи Конфедерации — наделил центральное правительство слабыми полномочиями и описывал новое государство не как

национальное объединение, а, скорее, как «дружеский союз» между отдельными государствами, чем-то напоминающий современный Европейский союз. Люди воспринимали себя не гражданами США, а гражданами определенного штата (государства).

И в этом смысле граждане говорили о Соединенных Штатах во множественном числе, что было вполне закономерно для союза различных и в целом независимых государств. Например, в обращении президента Джона Адамса 1799 года говорится о «Соединенных Штатах *и их* договорах с ее Британским Величеством» (курсив наш. — *Э. Э. и Ж.-Б. М.*). В наше время для президента США это совершенно немыслимо.

Когда же слова «Мы, народ…» (Конституция США, принятая в 1787 году) стали обозначать «одну нацию» (Клятва верности флагу, включенная в «Кодекс о флаге США» в 1942 году)?[1]

Если бы мы спросили об этом людей-историков, то, возможно, они бы указали нам на самый знаменитый ответ из финала знаменитой книги Джеймса Макферсона по истории гражданской войны — «Боевой клич свободы»[2]:

*…Некоторые масштабные последствия войны кажутся очевидными. Были побеждены раскол и рабство, чтобы никогда не возникнуть вновь, даже через полтора столетия после Аппоматокса. Этот итог означал серьезную трансформацию американского общества и изменение государственного устройства, уточнившегося, если не сформировавшегося, в результате войны. До 1861 года слова „Соединенные Штаты" чаще всего использовались как существительное во множествен-*

*ном числе: the United States are republic („Соединенные Штаты представляют собой республику"). Война привела к тому, что „Соединенные Штаты" стали в английском языке существительным в единственном числе.*

Макферсон был не первым, кто выдвинул такое предположение; эта тема обсуждается уже не менее сотни лет. Стоит хотя бы вспомнить выдержку из статьи в газете *Washington Post,* опубликованной в 1887 году[3]:

*Какое-то время, буквально несколько лет назад, о Соединенных Штатах говорилось во множественном числе. Было принято говорить: „Соединенные Штаты имеют" или „Соединенные Штаты являлись". Однако война все изменила. Вопрос грамматики был навсегда решен на линии огня от Чесапика до Сэбин-Пасс. Решение приняли не Уэллс, не Грин, не Линдли Мюррей, а сабли Шеридана, мушкеты Шермана и артиллерия Гранта… Поражение мистера Дэвиса и генерала Ли означало переход от множественного числа к единственному.*

Даже через сто лет после того, как была написана эта потрясающая история о языке, артиллерии и приключениях, сложно сдержать волнение. Кто бы мог представить, что люди станут сражаться за грамматику или что «мушкеты Шермана» решат спор о тонкостях словоупотребления?

Но стоит ли этому верить?

Возможно. Джеймс Макферсон — бывший президент Американской исторической ассоциации и настоящая легенда среди историков. Его самая знаме-

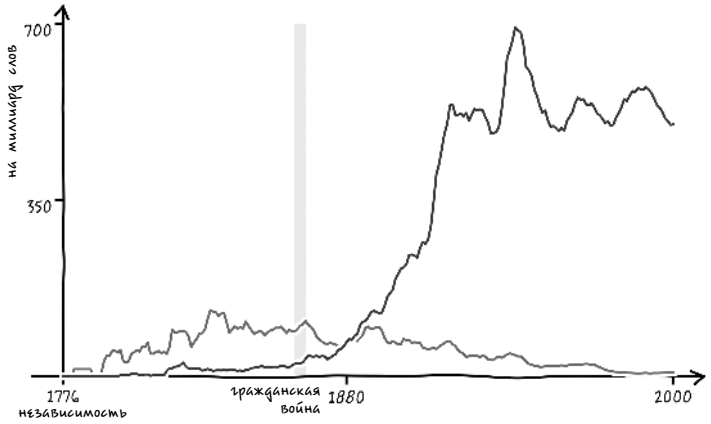
нитая работа «Боевой клич свободы» получила Пулитцеровскую премию. Более того, кто бы ни написал в 1887 году статью в *Washington Post*, Макферсон, вероятнее всего, сам испытал этот синтаксический переворот, и его свидетельству сложно не верить.

Тем не менее Джеймс Макферсон, каким бы великим он ни был, не непогрешим. А свидетели иногда неправильно интерпретируют факты. Можно ли с этим что-то сделать?

Возможно. Давайте предположим, что мы попросили нашего робота — гипотетического робота, прочитавшего все книги из всех библиотек, — поделиться с нами своим механистическим мнением.

Представим, что в ответ на наш вопрос услужливый робот-историк обращается к своей бездонной памяти и рисует график, приведенный на стр. 13[4]. На нем показано, насколько часто использовалось с течением времени понятие «Соединенные Штаты» в единственном или множественном числе в книгах на английском языке, опубликованных в США. Горизонтальная ось — течение времени, год за годом. На вертикальной оси указана частота употребления двух фраз в среднем на каждый миллиард слов текста за год. К примеру, робот прочитал 313 388 047 слов в книгах, опубликованных в 1831 году. Внутри этих слов робот видит фразу *the United States is* (то есть единственное число) 62 759 раз. Иными словами, в этом году данное выражение встречалось 20 раз на миллиард слов, что отражено в высоте синей линии за 1831 год.

Подобный график дает четкое представление о том, когда именно люди стали упоминать Соединенные Штаты в единственном числе.




Есть только одна небольшая проблема: судя по гипотетическому графику гипотетического робота, история, которую мы вам рассказываем, неверна. Во-первых, переход от множественного числа к единственному не был мгновенным. Он был постепенным, начался в 1810-х и продолжался вплоть до 1980-х — то есть более полутора столетий.

Но еще важнее то, что во времена Гражданской войны не происходило никакого резкого перехода. В сущности, период войны не особенно сильно отличался от времени до нее или после. Хотя в послевоенный период и началось некоторое ускорение процесса, однако оно произошло не ранее чем через пять лет после сдачи в плен генерала Ли. Согласно нашему роботу, единственное число не стало общеупотребительным вплоть до 1880 года (спустя пятнадцать лет после окончания войны)[5]. И даже сейчас время от

времени можно увидеть колыхание знамен лингвистической «конфедерации».

Разумеется, все это выглядит довольно умозрительно, поскольку наша история о роботе с навыками скоростного чтения, превосходящего в своей способности к анализу и свидетеля событий, и историка-лауреата, кажется совершенно надуманной.

Однако все это действительно так.

Макферсон, несмотря на всю свою гениальность, ошибался в вопросе о единственном числе. Свидетель помнил события неточно. А робот, о котором мы вам рассказывали, существует на самом деле. График, приведенный чуть выше, был действительно нарисован роботом. И своей очереди еще ждут миллиарды других графиков. В наши дни миллионы людей по всему миру видят историю совершенно по-новому — цифровыми глазами робота.

## *Форма света*

Стоит сказать, что не впервые на наше видение мира влияет появление той или иной новой линзы.

В конце XIII века по всей Италии получило активное распространение новое изобретение — очки. Всего лишь за несколько десятилетий очки прошли путь от никому не известной вещи до экзотического, а затем и вполне привычного аксессуара. Своеобразный предшественник смартфона, очки стали незаменимой вещью для множества итальянцев, совмещая в себе моду и функциональность. Они стали одним из первых триумфальных примеров использования пеленосных технологий.

По мере распространения очков по Европе и всему миру оптометрия превратилась в серьезный бизнес, а технологии изготовления линз стали лучше и дешевле. Разумеется, люди начали экспериментировать и изучать, что будет при совместном использовании нескольких линз. Прошло совсем немного времени, и люди поняли, что при должной инженерной сноровке можно достичь невероятной степени увеличения. Появилась возможность изготовления составных линз, с помощью которых можно было открывать новые миры, невидимые невооруженному человеческому глазу[6].

Например, с помощью таких линз можно было увеличивать изображение самых крошечных вещей. Микроскопы позволили узнать как минимум два факта, связанных с вековой тайной жизни. Во-первых, они показали, что окружающие нас животные и растения состоят из крошечных отдельных частиц. Сделавший это открытие Роберт Гук заметил, что расположение этих частиц напоминает монастырские кельи, и назвал их «клетками»[7]. Во-вторых, микроскопы позволили нам узнать о существовании микробов[8]. Эта совершенно отдельная вселенная организмов, часто состоящих из единственной клетки, населена большей частью обитателей нашего мира. До изобретения микроскопа никто даже не представлял себе, что существование подобных форм жизни возможно.

Составные линзы использовались также и для приближения удаленных объектов. Вооружившись телескопом, дающим 30-кратное увеличение — по нынешним стандартам детская игрушка, — Галилей смог заняться разгадкой тайн космоса[9]. Куда бы он ни посмо-