

---

## Digitizing Texts

- or -

*Google Ngram Viewer Turns Snippets into Insight*

---

Computers have the ability to store enormous amounts of information.

But while it may seem good to have as big a pile of information as possible, as the pile gets bigger, it becomes increasingly difficult to find any particular item.

In the old days, helping people find information was the job of phone books, indexes in books, catalogs, card catalogs, and especially librarians.

## Goals for this Lecture:

1. Investigate the history of digitizing text;
2. Understand the goal of Google Books;
3. Describe logistical, legal, and software problems of Google Books;
4. Introduce the type of computer software that converts scanned text to searchable documents;
5. To see the compromises that Google had to make in order to realize its goal of digitizing all books.

---

## Google Books

---

This is the story of Google Books. It starts out promising a world wide library of all the books ever printed, accessible to everyone. Gradually it changes into something that is still useful, but not quite the miracle that was advertised. This story shows that even the most powerful computer company in the world can't always get what it wants, because computers and computer programmers must work in a complicated world.

For thousands of years, books have been understood to be the symbol and carrier of culture and wisdom. Even though a particular tablet, scroll, or book can eventually fall apart, the information it carries, in the form of writing, can be copied fresh, and in this way, we are still able to read Julius Caesar's war memoirs, the story of Genghis Khan in "The Secret History of the Mongols", and the Indian epic of Prince Rama.

---

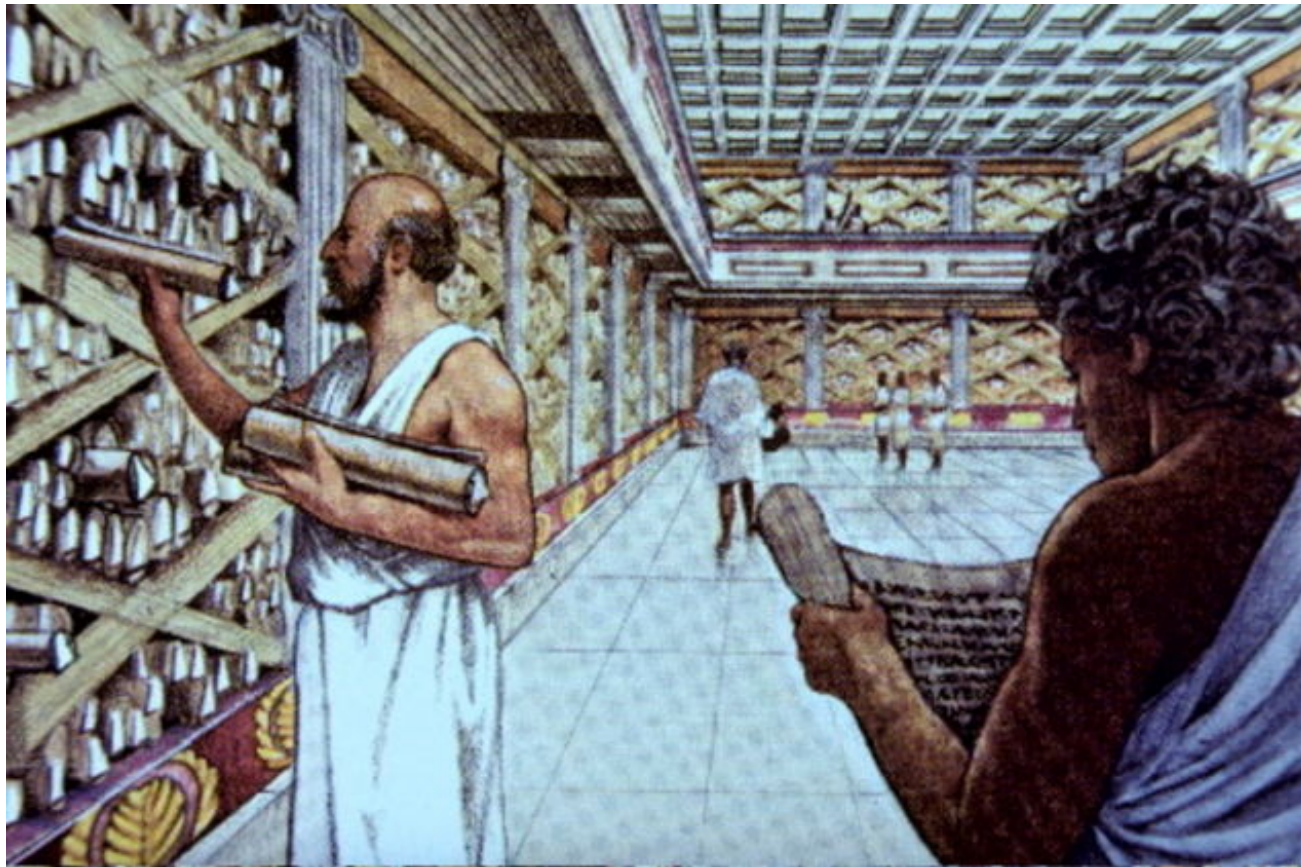
## The Library of Alexandria

---

The reverence for books, and the passion for collecting them, has a most famous illustration in the library of Alexandria, an ancient city founded by Alexander the Great, where the Nile empties into the Mediterranean Ocean. A steady stream of trading ships entered its harbor daily.

By decree, inspectors boarded each ship and requested the crew to temporarily turn over any scrolls they carried, in any language, which were copied and added to the collection. Over time, Alexandria became known as the storehouse of all wisdom and knowledge of the ancient world, and students would come to study with the wise teachers who guarded the library treasures. It was started in the third century BC by Alexander the Great's successor.

In later years, the library suffered catastrophes of fire, war, political violence, religious wars, and the relentless crumbling of papyrus, until nothing was left of the building or its books. But the memory of the library of Alexandria has since become a symbol of an ideal collection of all the writings of all time and places, available to anyone.



Now that we are learning how computers are changing our future, one question we might ask is:

Could computers construct a modern library of Alexandria?

That is:

- Can we make all the world's books available to everyone?
- Is this a good idea that everyone will approve of?
- How could we do it?
- How much would it cost?
- Who would be willing to do it?
- How would the resulting library actually work?

This is not a simple task!

Once we try to answer these questions, we realize that we don't even know how to begin to estimate the difficulty of such a project!

Can you tell me:

- How many (unique) books are there in the world?
- Where are these books now?
- How much computer storage space does a typical book require?
- How can a physical book be entered into a computer?
- Do we have to get permission to put a book into the computer?
- Do we have to get permission to let someone access a computerized book?

Our computer solution to reconstructing the library of Alexandria probably starts with the simple idea of, “Well, just put all the books onto the computer in one place, and tell people where that is!”

That’s because we know, from experience with Internet browsers, that:

- the Internet somehow has stored a lot of information already;
- that information is easy to access;
- that information can be accessed quickly;
- nobody seems to charge any money for access (to most things, anyway!);

But there are significant differences between personal web pages and books.



At one time, few people could read, and books had to be copied by hand.

The Greeks and Romans used professional scribes to make copies of scrolls by hand, a process that could take many weeks.

In medieval Europe, monks would make copies of books as part of their religious duties, and these books generally were only used within the monastery.



---

## Publishers

---

When Gutenberg perfected the printing press about the middle of the 1400's, it was possible to make hundreds of copies of a book, market them, and make a living doing so.

The publishing business was born, relying on converting handwritten manuscripts into printed texts, and printed texts into francs, marks, pounds and dollars.

Books and information became a kind of **valuable property**.



---

## Authors

---

In ancient times, no one made money by writing books. Instead, they did it for the love of learning, or at the request of friends, or for prestige.

Once publishing became a business, however, publishers gradually realized that, at least if an author was still alive, it was useful to pay a royalty in return for exclusive rights to publish the author's work. Now the publisher was essentially **licensing** the property of the author.



---

## Copyright Laws

---

At first, agreements between publishers and authors were irregular or informal. Sometimes a publisher wouldn't make an arrangement with an author, and sometimes a "pirate publisher" would issue a separate, cheaper edition of a book without getting rights from the author or legal publisher.

A system of copyright laws developed, to provide **legal protection** for the property of authors and publishers, with fines and punishments for violations.



---

## Copyright Lawyers

---

The growth of copyright laws created a new class of copyright lawyers, who monitored publications, looking for violations, and threatening legal action to protect what began to be called **intellectual property**.

Copyright law was extended to music, song lyrics, stage plays, and art work. and had to adjust to problems arising from new technology, including copy machines, audiotape, VCR's.

Even the makers of player pianos were sued, on the argument that the paper tape represented an illegal copy of a song.



---

## Our Simple Plan May Run Into Trouble

---

Our idealistic plan of recreating the library of Alexandria may start to seem a little bit like a bad joke that begins:

A computer scientist and an author and a publisher and a lawyer and a librarian walk into a bar...

In fact, it might not be merely a joke, but a legal disaster. If every book is a valuable piece of property, then someone who comes along and plans to vacuum up all this property into a new enterprise could be facing thousands of angry owners.

So now we have a wonderful idea (books for everyone) which has become computerized (free online books for everyone), but which would have to be realized in a world in which books are property, authors and publishers are owners, lawyers are enforcers, and librarians are well-meaning but somewhat helpless observers.

And into this world walks a company named Google which says,

*“It seems like such a good idea, how hard could it be, let’s do it!”*





---

## Google is one of the leaders in digitizing text

---

The following is from a 2007 New Yorker article “*Google’s Moon Shot*”:

The story of how Sergey Brin and Google’s other co-founder, Larry Page, met as graduate students in computer science at Stanford in the mid-nineties, and devised a series of elegant software algorithms that allowed Web searchers to find relevant information quickly and efficiently, has become part of Silicon Valley lore.

Less well known is that, at the time, Brin and Page were also working on Stanford’s Digital Library Technologies Project, an attempt, funded by the federal government, to organize different kinds of stored information, including books, articles, and journals, in digital form.

*“There was an attitude in computer science that putting things on dead trees was obsolete, and getting it all into a searchable, digital format was a quest that had to be accomplished someday,”* Terry Winograd, a Stanford professor who was a mentor to Page and Brin, said.



---

## Google announces Google Books

---

In 2004, Google announced a plan to systematically scan a copy of **every book** ever published. They could only estimate the total number of books as around 130 million.

The result would be a huge database called [Google Books](#).

As the project got under way, every week a truck would pull up to the Cecil H. Green Library at Stanford and take at least a thousand books to an undisclosed location to be scanned. This process was soon repeated at other University libraries such as Oxford and Harvard, and gradually the search spread further to fill in missing entries in this universal library.

---

## Other book digitization efforts

---

Google is not the only book scanning venture.

- **Amazon** has digitized hundreds of thousands of the (new) books it sells.
- **Carnegie Mellon University** has a million books digitized in the *Universal Digital Library*.
- **Project Gutenberg** has digitized 50,000 (old) items in the public domain, no longer subject to copyright protection.

Still, only Google has embarked on a project of a scale commensurate with its corporate philosophy: “to organize the world’s information and make it universally accessible and useful.”

As of October 2015, Google had scanned 25 million books, new, old, in and out of copyright, still far short of their goal of 130 million.

---

## The Google Books project is welcomed by some

---

Libraries cooperated with Google because they recognized the need to move their holdings into the modern times, but were dealing with many constraints:

- limited budget;
- limited hours of operation;
- lack of space for storing books;
- difficulty of accessing rarely used books;
- the need to carefully index and catalog books;
- the need to restore books to the shelf upon return;
- danger of accidental or deliberate damage to books;
- book theft;
- slow operation of interlibrary loan;
- growing demands for computers and computer areas;



---

## **Publishers oppose Google Books**

---

Although Google had negotiated with the libraries, it did not notify publishers and authors that it was scanning copyrighted works.

Publishers already felt that sales to libraries cost them business. A single copy might be read by 20 or 30 library patrons, and all those potential sales were lost.

Now, Google could take a single copy of a book and make it available to everyone in the world for free. Wouldn't this mean the end of publishing as a business?

---

## Publishing is a big business

---



---

## **Publishers proposed a licensing organization**

---

Google would be making all books available to everyone.

Publishers felt that Google was hiding behind non-profit libraries, and that Google was certain to earn money by selling ads that would appear alongside the scanned books.

Publishers didn't think that was wrong - but they wanted their share of the revenue.

More importantly, they did not want Google to be in control of the rights associated with books.

Publishers proposed a book licensing organization, which would control the creation, distribution and use of all electronic books charging a fee per use, as ASCAP does for music.

---

## **Authors oppose Google and the publishers**

---

When authors heard about Google Books, they raised new objections.

And they did not simply support the publishers.

Authors felt that electronic or digitized books were new uses of their work, which were not covered by their contracts with publishers.

Therefore, the authors should be presumed to control these rights, and only an individual author should be able to allow the creation of electronic versions of a work.

In announcing Google Books, Google had described it as a tremendous gift to the world. Now publishers and authors were calling it a tremendous theft of their property.

Now the dispute involved thousands of people, all with different motives and needs.



---

## Lawsuits!

---

The Authors Guild sued in September 2005, followed by the Association of American Publishers. Arguments included:

- We tolerate libraries sharing books, but Google makes a profit;
- Google is worse than a pirate - pirates distribute illegal copies, Google makes illegal copies first, and then distributes them;
- Google is a monopoly - Google will control how the information is used;
- Censorship: A for-profit company can be pressured by governments;
- Orphans: books whose copyright owners can't be determined are called orphans. Google doesn't pay anything to scan and display such books. Why should Google get this windfall?
- Privacy: Google will be able to tell what books people read; collecting, using, and even selling such data may violate privacy.

---

## Defense and Compromise

---

Google's case in court depended on the idea that its use of books was legal because it was **transformative**. By only providing snippets of copyrighted book texts, no one was actually able to read a complete book; and instead, Google was creating an entirely new, and noncompetitive, service.

With strong arguments on both sides, it wasn't clear who had the best case.

In such situations, both parties often prefer to reach a settlement, getting a result they can live with rather than risking a total loss. With so much at stake, the Google Books litigants searched for an agreeable compromise.

---

## 2008-2011: A settlement is attempted

---

After much negotiation, the settlement proposed to the judge stipulated:

- Google would pay \$60 to the author of every book that was scanned;
- Google would pay \$125 million to copyright owners, and fund a Book Rights Registry to distribute royalties to owners;
- Google would set up free portals in 4,000 colleges and universities;
- For any copyrighted work, Google would only allow users to see small portions or “snippets”, not the whole thing;
- Google would allow any publisher to withdraw all their books;
- Google would allow author to withdraw their books;
- Google would include a “buy this book” link along with a search result.

March 2011: Settlement was **rejected**; the Authors Guild resumed the suit.

November 2013: US Circuit Court dismissed the suit.

October 2015: the Second US Circuit Court of Appeals rejected the appeal.

April 2016: the Supreme Court rejected the appeal; **the case is over.**



It looks like Google Books is here to stay!

---

## Can Google afford to do a good job?

---

The Library of Congress made an independent estimate of the costs of accurately preserving a 300 page book, including

- \$65, to make a Xerox copy;
- \$185, to make a microfilm image of each page;
- \$1,600, to do a “low level” digital format;
- \$2,500, to do an “enhanced” digital format.

No one, not even Google, can afford to spend \$2,500 per book to digitize 130 million books. A suggested budget of \$800 million leaves less than \$7 per book!

Google realized it had to make many compromises, and discover some efficient methods, just to approximate its goal.

---

## How does Google get searchable text?

---

Turning a book into a searchable text begins with scanning, that is, making a photograph of each page.

To cut down the scanning cost, an automatic scanner was developed; which can turn to the next page after each photograph is made.



Automatic scanning is cheaper than having a human copy each page of the text but it is less reliable.

The quality of the resulting image depends on the strength of the scanner light, the physical state of the book, the printing style, the scanner resolution.

Occasionally, pages are skipped, torn, or folded over!

The resulting image may be too faint to read, or the text on the next page may show through, or the image may be blurred because the book moved slightly.



(503)  
 Query. *How are Spirits to  
 be Discern'd.*

If this Query refers to the Persons themselves, who are, or suppose themselves led by a Good Spirit, it is Answered, That the Manner of Discernment is by Sensation, which is not to be described. This Sensation like the New Song, the New Name, and the White Stone, they only who do experiment, and while they do experiment, are able to distinguish. But then it may be Queried, How shall they who suppose themselves to be led by a True Spirit, and are not, be enabled to discern aright in this matter, since they are without this Spiritual Sensation, by which the others are assur'd concerning the Truth of their Visions and Revelations? *Ans.* They are to seek for it of God, by a perfect Mortification of their own Wills, and especially as to these very supposed Favours; and by most instant and violent Prayer. Thus some who upon the Commission of an evil thing, have been depriv'd of the Good Spirit of God, which they did enjoy; and have been deliver'd up to a Spirit of Delusion for some time,

Query. *or are jSpirits to '  
 .-be' Difern'd.*

If this Query refers to the \*Pe'rsons themselves' - who are or suppose the 'm- selves led by a PP Good Spirit, it is Answered, That the Manner of Discernment is by Sensation, which is not to be described. This Sensation like the New Song, the New Name, and the White Stone, they only who do experiment,

and while they do experiment, are 'able to distinguish. But then it may be Queried, How shall they who suppose the 'n (elves- to be, led by a True Spirit, and are *not, be* enabled to discern aright, in this

matter, since they *are* without this Spiritual Sensation, by which the others are assur'd concerning the Truth of their Visions and Revelations? *Ans.* They are to seek for it of God, by a perfect Mortification of their own Wills, and, especially as to these very supposed Favours; and by most instant and violent Prayer. Thus some who upon the Commission of an evil thing, have been depriv'd of the Good Spirit of God, which they did enjoy; and have been deliver'd up to a Spirit of Delusion for some time,



---

## Photographs aren't text files

---

A picture of a page may contain the information we are interested in, but it's stored as a photograph, not as a text document.

Once a book has been turned into a series of photographs, we need to turn the photographs into a single text, containing the words.

It's much too expensive to hire typists to look at the photographs and type up the information, so a cheaper, automatic solution is necessary.

This step is only possible because of amazing software that can teach a computer how to “see” a photograph, “recognize” the alphabetical characters in the photo, and “read” the resulting words.

---

## The Problem Now: Turn Pictures of Text into Text

---

By working with libraries, Google got free access to millions of books.

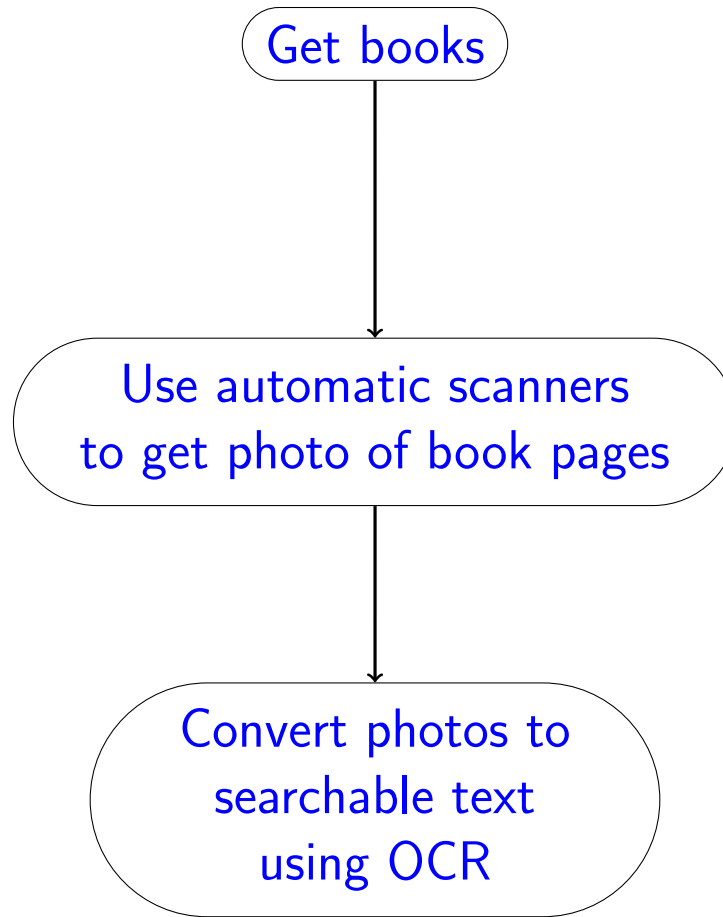
By developing automatic scanners, Google turned millions of books into hundreds of millions of photographs of book pages.

Now what Google needed was to transform the many photographed pages from one book into a searchable text file, as though someone had typed the book text in by hand.

Google could not afford to hire humans to examine the photographs and type the information into files.

Instead, Google turned to **Optical Character Recognition (OCR)**.

OCR is automatic and hence cheaper, but can be less reliable than a human reader. But this was the only way for Google to reach its goal.



---

## Optical Character Recognition (OCR)

---

Youtube video “How does optical character recognition (OCR) work?” by Techquickie



---

## Optical Character Recognition (OCR)

---

Google Books relies on Optical Character Recognition, **OCR**.

As you read these slides, your eyes and brain carry out a type of OCR! Your eyes record the patterns of light and dark that make up characters and your brain recognizes characters, groups them into words, retrieves the meaning of those words, and interprets the resulting sentences.

We don't actually need the computer to understand what it sees, just to recognize the character and word patterns.

The information we give to the computer will be a photograph of a book page generated with a scanner; – it's just a pattern of pixels, not words!

An OCR program examines pixel patterns and tries to match them against a known alphabet. This is tough, since the letters can be at any size, at an angle, in different colors or fonts. However, a good OCR program, given a good quality photographic image, can do an amazing job.

---

## OCR creates a text file

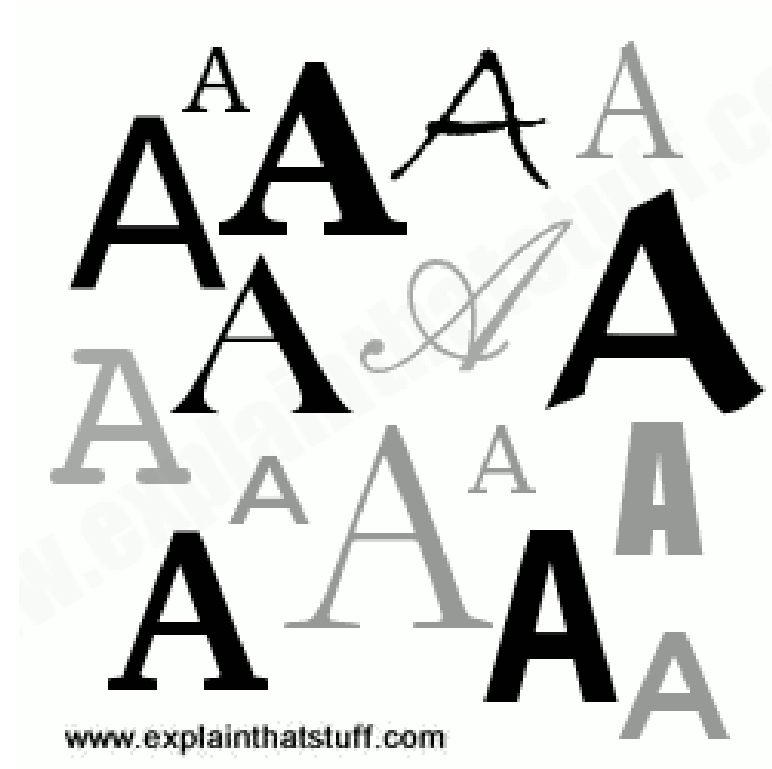
---

The OCR program creates a text file, just like a document created, for instance, by Microsoft Word.

Unlike the original photographic image, you can do lots of things with a text file created by OCR:

- you can search for a keyword;
- edit it;
- incorporate it into a web page;
- compress it;

OCR would be hard even if we were only looking for examples of the letter “A”! the letter could be bold or italic or underlined, in various fonts, at any size, or any angle, even handwritten.



---

## A special font to make OCR easier

---

In the early days, OCR software was not very good. To improve the results, a special font was developed that OCR could recognize well.

This monospaced (fixed-width) font was called **OCR-A**, and was designed with simple, thick strokes that are ugly but easily recognized.

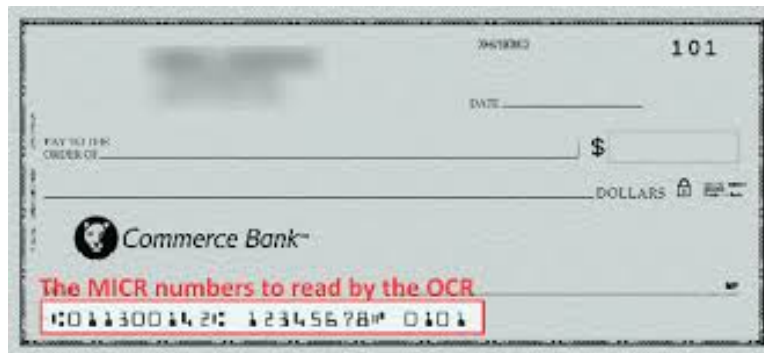
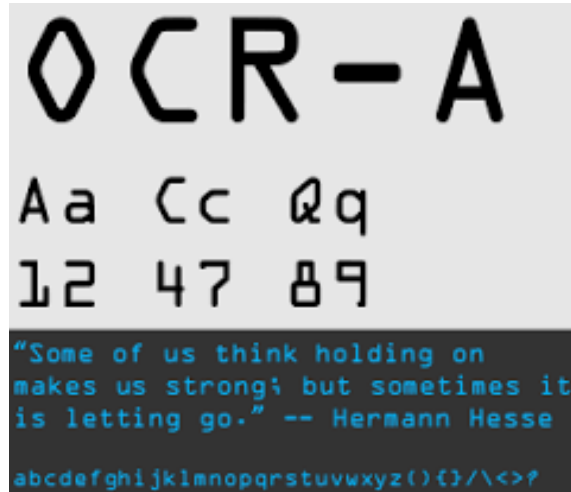
The OCR-A font is still in use by banks, and you can see it if you look at a check or the numbers on your credit card. However, it is not so easy for humans to read so it would never be used as a font in books.



---

## Examples of the OCR-A font

---



---

## OCR by pattern recognition

---

In order to improve OCR software, one technique that was tried was **pattern recognition**. In this approach, an OCR computer program was designed which could try to recognize characters, but which also had a limited ability to remember its wrong guesses.

To improve the software, the programmer would “train” the software by showing it many examples, and telling it **yes** or **no** depending on whether its answer was correct.

This kind of approach is an example of what is now called **machine learning**; in some ways, it is similar to how a child learns how to speak, and to correctly identify shapes and colors. We will talk about algorithms for pattern recognition when we do the machine learning unit.

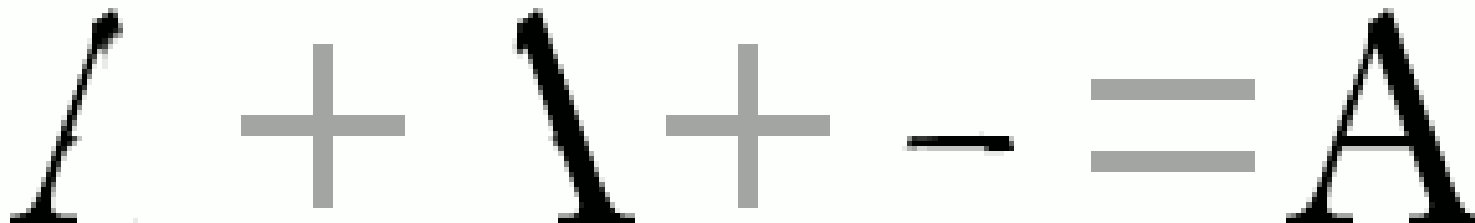
---

## OCR by feature recognition

---

Feature recognition OCR tells the computer how to recognize a letter by searching for a particular pattern of lines or strokes which make up the letter.

For “A” one could try to recognize two angled lines which meet in a point at the top and a horizontal line between them approximately halfway down.



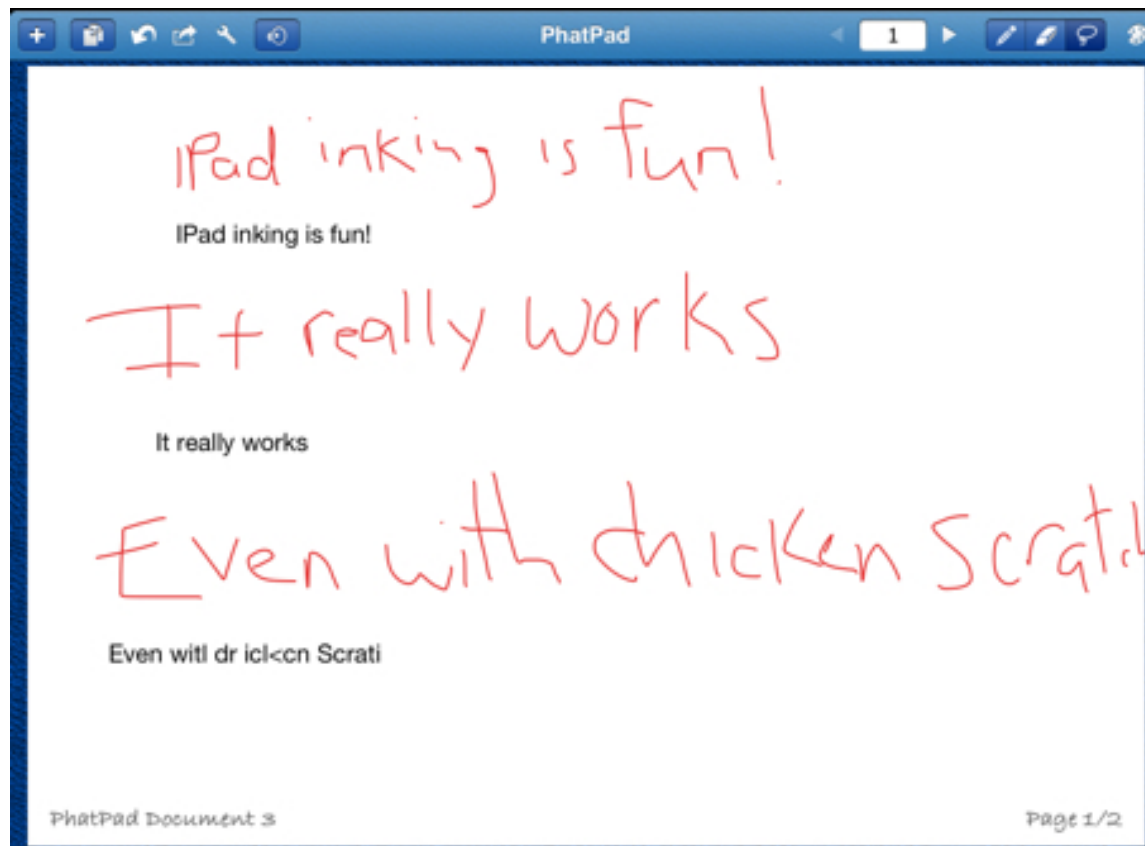
[www.explainthatstuff.com](http://www.explainthatstuff.com)

---

## OCR is now good enough to be widely used

---

So much has been learned about the character recognition problem that, these days, OCR can be used for handwriting recognition and mail sorting.



---

## OCR can still be fooled by mangled text

---

An interesting “inverse” application of OCR is [reCAPTCHA](#), developed at Carnegie Mellon University. OCR software finds it hard to recognize badly written words but humans can usually read them, CAPTCHA puzzles are used to defeat automatic spam programs.



The image shows a reCAPTCHA interface. At the top, two words are displayed in a heavily distorted, cursive font: "overlooks" on the left and "inquiry" on the right. Below these words is a yellow rectangular box containing the text "Type the two words:". Underneath this text is a white rectangular input field with a blue border. To the right of the input field are three small red buttons with white icons: a circular arrow (refresh), a double left arrow (previous), and a question mark (help). Further to the right is the reCAPTCHA logo, which consists of a large white 'C' shape on a red background, followed by the text "reCAPTCHA™" and the tagline "stop spam. read books." in white.

---

## By using OCR, Google must accept a certain error rate

---

To process 130 million books, Google has automated the scanning process, and then used OCR to automate the translation of scanned images into text.

Errors can come from a damaged or badly printed book, a scanner that is misaligned or jostled, a page that is only partially turned, light or dark printing, or unusual fonts.

When a garbled OCR text is created, it could only be fixed by human intervention...someone would have to find the book, turn to the proper page, read the lines, and type them into the text file.

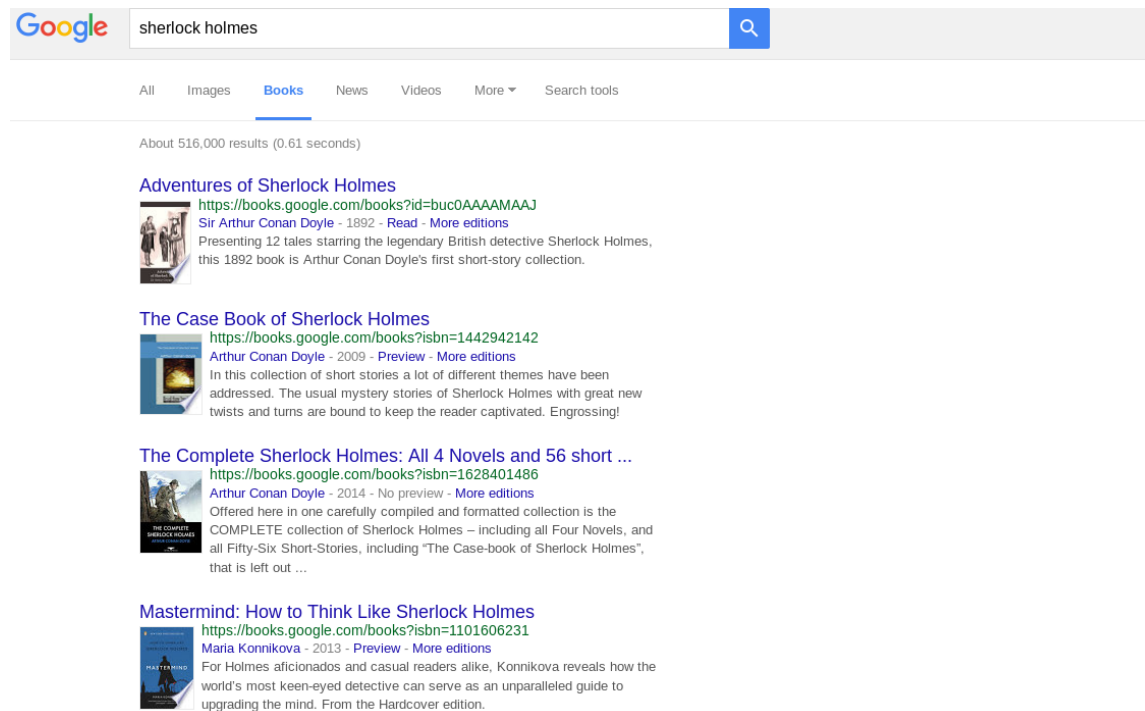
Such a remedy is not affordable. The Google Books team can improve their algorithms, and, might conceivably rescan the book, but they will not hand-correct an error.

---

## Browse the Google Books library

---

If you go to the Google Books website and type in “Sherlock Holmes” (with quotes), a list of matches appears, with the best matches first.



One of the first items on the list is “Adventures of Sherlock Holmes” , with the address **<https://books.google.com/books?id=RxAJAAAAIAAJ>**;

You can select this book, and you will be able to “page through it” , with the additional feature that all the pages on which the words “Sherlock Holmes” appears will have a bookmark allowing you to quickly jump there.

It is important to realize that what you are seeing right now is the **photo-graphic image of the book**, before OCR is applied.




Here is the photographic image, from a scanner, of the first page of the Sherlock Holmes adventure titled [A Scandal in Bohemia](#).

## Adventure II

### A SCANDAL IN BOHEMIA

#### I

O **Sherlock Holmes** she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise, but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen ; but, as a lover, he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer—excellent for drawing the veil from men's motives

You probably don't realize that this page, which looks very clean and readable, presents challenges to an OCR translator.

The heading **Adventure 1** is printed in an unusual font called "Blackletter".

The "T" in *To Sherlock Holmes she is always the woman* is displayed in a fancy style suggestive of an old handwritten manuscript.

There is a hyphenated word split between text lines 3 and 4. Will the OCR know what to do with this?

There is a pencil mark to the right of the word *most* in text line 7.

The phrase *observer–excellent* almost looks like one word.

On this single page we see many potential problems.

We can apply OCR software to this photographed page:

1. Go to the first page of the adventure [Scandal in Bohemia](#)
2. In the menu you will see an image of [scissors](#) which are used to clip a section of text. Click on this and a “+” will appear.
3. Use this to draw a box around the text you want to have translated, say the title and the first paragraph.
4. A popup window will appear; Click [Translate](#) and the result of applying OCR will appear.

## The OCR version

floventure I A SCANDAL IN BOHEMIA I have seldom heard him mention her under any other name in his eyes she eclipses and predominates the whole of her sex it was not that he felt any emotion akin to love for Irene Adler All emotions and that one particularly were abhorrent to his cold precise but admirably balanced mind He was I take it the most perfect reasoning and observing machine that the world has seen but as a lover he would have placed himself in a false position He never spoke of the softer passions save with a gibe and a sneer They were admirable things for the observer excellent for drawing the veil from men's motives

---

## OCR is good, but not perfect!

---

The OCR version is actually a very good copy of the text...but it is by no means perfect. There are some real problems:

- “Adventure” in the title has been interpreted as floventure
- The two line heading has merged into a single line;
- The fancy T caused problems and the first sentence has disappeared;
- The hyphenated word “predominates” has caused problems;
- All the punctuation has disappeared;

The omissions and word changes are particularly disturbing, since they change the meaning, and are harder to notice.

However, this is the result from the current state of OCR software.

---

## Socrative Quiz Ngrams\_Quiz1

CTISC1057

---

1. Everyone is in favor of creating Google Books.
2. Google Books is the only effort to date at digitizing books.
3. To create Google Books, automatic scanners are used to first make a photograph of each page of the book.
4. Using an automatic scanner to photograph pages of a book is 100% accurate.
5. The goal of optical character recognition (OCR) software is to turn an image of text into a searchable file.
6. An OCR translation is always 100% accurate.
7. OCR-A is a special font where all the characters have the same width and are easily distinguishable from each other.

8. OCR-A is a popular font used in printed novels.
9. reCAPTCHA is an application which relies on the fact that computers have trouble reading badly written text.
10. Google's final goal in creating Google Books is to create a photograph of each page of every book.

---

## Goals for this Lecture

---

1. To see how Google creates a searchable data base even for copyrighted books;
2. What an n-gram is;
3. How the Google Ngram Viewer works.
4. How Ngram Viewer can answer interesting questions, especially for word use.



---

## Creating a searchable data base even for copyrighted books

---

Remember that Google agreed not to make the entire text of copyrighted books available online.

However, “**snippets**”, that is, short strings of text, were legally allowed.

A snippet is called an **n-gram**.

For example, any two **consecutive words** is called a 2-gram (or bigram).

Any single word is called a 1-gram (or unigram).

Any three **consecutive words** is called a 3-gram (or trigram).

For example, any four **consecutive words** is called a 4-gram.

For example, any five **consecutive words** is called a 5-gram.

In other words, each book was essentially put through a grinder to make

pieces consisting of strings of words of length 1, 2, 3, 4, 5.

**Example** Consider the sentence

The cow jumps over the moon.

Determine how many 1-grams, 2-grams, 3-grams, and 4-grams we can form from this sentence.

1-grams	2-grams	3-grams	4-grams
The	The cow	The cow jumps	The cow jumps over
cow	cow jumps	cow jumps over	cow jumps over the
jumps	jumps over	jumps over the	jumps over the moon
over	over the	over the moon	-
the	the moon	-	-
moon	-	-	-

So in the sentence **The cow jumps over the moon** there are:

- 6 unigrams
- 5 bigrams (2-grams)
- 4 trigrams (3-grams)
- 3 quadgrams (4-grams)

Notice that there are 6 words in the sentence, so that is the total number of unigrams. The total number of bigrams is one less than the number of words, so 5, etc.

Sometimes we are interested in the number of **unique** ngrams. In our example, there are only 5 unique unigrams because “the” is repeated. All the 2-grams are unique as are all the 3- and 4-grams.

**Example.** Use the sentence

The current head coach of the Seminoles is Jimbo Fisher.

to answer the following questions.

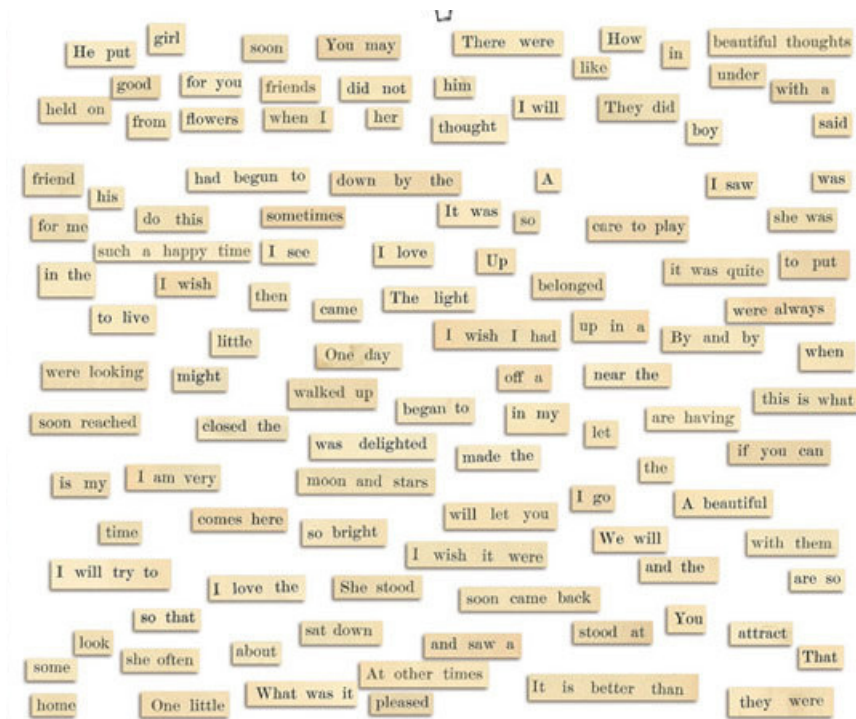
1. How many total 1-grams does this sentence contain?
2. How many unique 1-grams does this sentence contain?
3. How many 2-grams does this sentence contain?
4. How many 3-grams does this sentence contain?

---

## A collection of n-grams

---

You can think of this as cutting a book up into thousands of scraps of paper. Each scrap contains a few words, and is labeled by the title of the book, language, and date of publication.



---

## What's the use of an N-gram?

---

By cutting books up into n-grams, we've thrown away a lot of information!  
Can we do anything interesting with what is left?

The n-grams don't have much meaning, but they do represent words and phrases, and each n-gram is labeled by the author, book, date and other information.

So we can examine the history and popularity of any short phrase, at least in so far as it appears in print.

*Xmas* is a common abbreviation for *Christmas*. Is this a recent invention?

How old are the phrases *bite the bullet* and *the whole nine yards*?

We can also use it to complete a phrase; for example, what do you think the most popular word is after **Mickey**?

---

## Ngram Viewer

---

In order to make it easy for users to work with the database of n-grams created from Google Books, a team at Google built the [Ngram Viewer](http://ngrams.googlelabs.com), which displays the frequency of occurrence of a word or phrase over time.

The Ngram viewer is available at <http://ngrams.googlelabs.com>

---

## What Ngram Viewer can show you

---

When you start Ngram Viewer you see the results of a sample query, namely [Albert Einstein](#), [Sherlock Holmes](#), [Frankenstein](#). So the input is two bigrams (“Albert Einstein” and “Sherlock Holmes”) and one unigram (“Frankenstein”).

Each separate item results in a curve, shown on a common plot.

The  $x$ -axis is the range of publication dates of books you are searching.

The  $y$ -axis gives a percent of the books published in a particular year containing your search phrase.

By moving the cursor along the Albert Einstein curve, you can see it has the value 0.0000231603 for the year 1940.

This means that, of all the two word phrases in all the English books published in 1940, 0.0000231603 percent were “Albert Einstein”.



Is that a big value? A small value? Well, we can see that Einstein's name appears with greater and greater frequency over time, which seems reasonable. We are not so interested in the number but rather its size compared with other search ngrams.

Note that all punctuation (including hyphens) is ignored.

**Example.** Compare the occurrences of Albert Einstein, Frankenstein, Sherlock Holmes

## Google Books Ngram Viewer

Graph these comma-separated phrases:  ☐ case-insensitive

between  and  from the corpus  with smoothing of  [Search lots of books](#)



## Comparing the Ngram Viewer plot to Einstein's life

To show up in Ngram Viewer for a particular year requires at least 40 hits.

Einstein doesn't make it until 1915, which makes sense because he was born in 1879, received his Ph.D in 1905, became a professor in Prague in 1911, and in Berlin in 1914.

His calculations concerning general relativity (1911) were confirmed by Sir Arthur Eddington during the solar eclipse of 1919 and these observations were then reported in the international media so we see the number of occurrences steadily increasing.

However, only between about 1950 and 1970 was he more popular in text than Frankenstein!

## Examining the Ngram Viewer plot for Frankenstein

*Frankenstein* is a novel written by Mary Shelley, and published in 1818.

Amazingly, since about 1970 the word Frankenstein has become much more common in print than either Albert Einstein or Sherlock Holmes.

Although Frankenstein is actually the mad scientist, the name became associated with the monster instead, and now the word has become very widely used.

The growth in the use of the word might be explained by the fact that it now is used generically for any monster.

If you were interested in pursuing this question, it is possible to pick a year, get a list of all the Google Books containing the word Frankenstein, and see a short snippet of text from the book, which could help you determine how the word is being used.

Notice that the plot shows the 1-gram “Frankenstein” occurring before 1818.

Of course there could have been real Baron Frankensteins during that period.

A more likely reason is human error. Each ngram is tagged with the book’s title, author, and publication date, but this information comes from a library’s card catalog and might have been entered incorrectly.

To investigate a question like this, tighten the year range, and ask for the list of Google Books containing “Frankenstein”. Choose such a book, and jump to the location in the book, with some surrounding text.

The context may show it’s about some other (non-monster) person.

Or it may be obvious that this is a modern book that has been incorrectly dated.

Otherwise you can check the book’s catalog information.

## Examining the Ngram Viewer plot for Sherlock Holmes

“Sherlock Holmes” is the name of a fictional private detective created by Sir Arthur Conan Doyle and the name first appeared in print in 1887 and came in widespread popularity with the printing of “A Scandal in Bohemia” in 1891.

New stories and novel continued to appear until the author’s death in 1930.

The phrase gradually decreases in popularity until about 1965.

Around that time, there began a series of new stories, movies, and television shows that continues to this day, which might explain the second rise in popularity.

The phrase “Sherlock Holmes” has also become a common synonym for “detective” and so might be used with meaning a direct reference to the character.

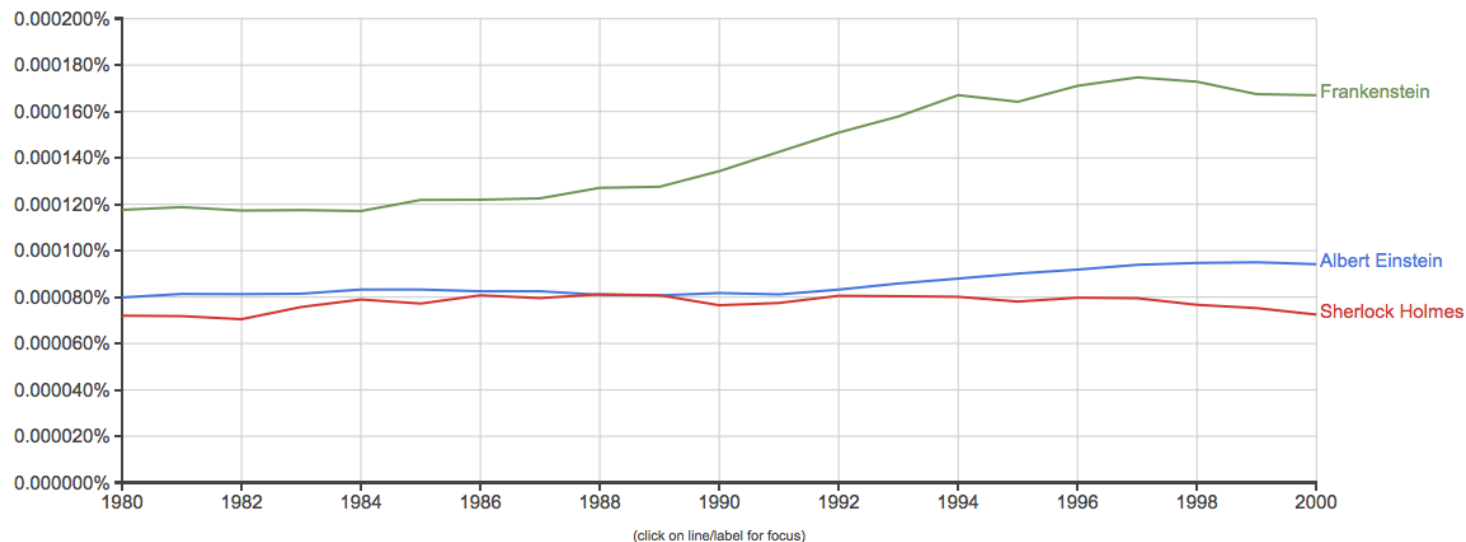
## Narrowing the time span

The plots created by Ngram Viewer can show much more detail if we reduce the range of years to a tighter interval. We do this simply by typing in new values in the date boxes in the upper left.

### Google Books Ngram Viewer

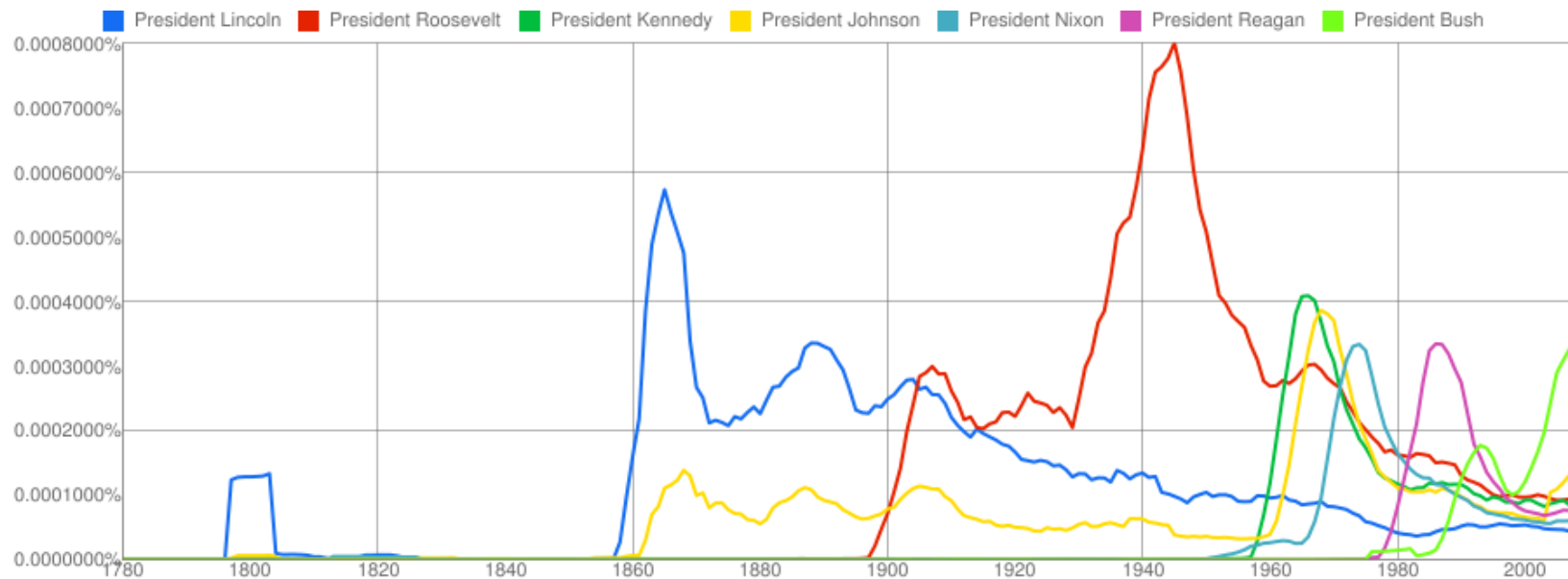
Graph these comma-separated phrases:  ☐ case-insensitive

between  and  from the corpus  with smoothing of  [Search lots of books](#)



## Example. Searching for presidents

To search for US presidents, we include the word “president”; otherwise we’d find references to every person with the same last name!



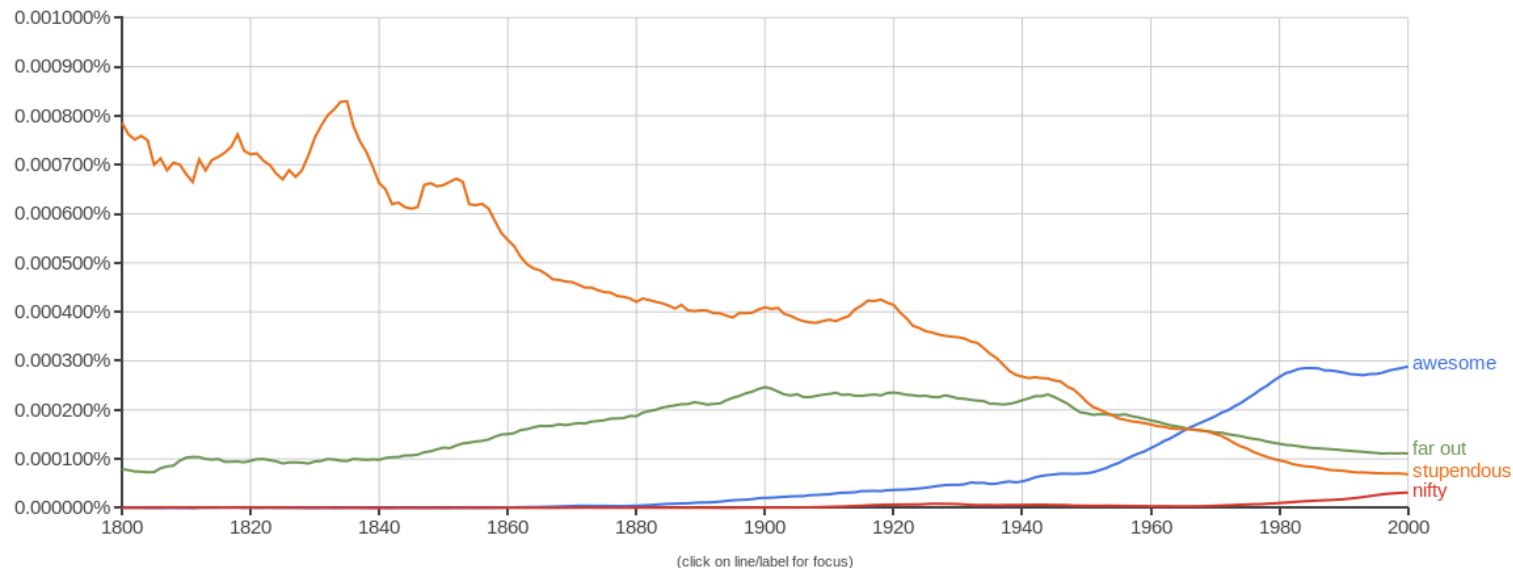
Notice double humps for Roosevelt (red curve) and Bush (light green).

Since Lincoln wasn't even born until 1808, the blip for him at 1800 must be bad catalog information (incorrect publishing date).



**Example.** Suppose we want to know when people stopped saying “far out” and started saying “awesome” or “nifty” or “stupendous”.

The plot suggests that “nifty” and “awesome” are relative newcomers; that “stupendous” is rapidly dropping, while “nifty” and “awesome” are rising and “far out” is holding more or less steady.



## Did they really say Far Out in 1840?

However, we should not be too sure about the results for “far out”.

Remember, what is happening is that a search is made for each occurrence of the bigram “far out” and after all, this phrase has another meaning, simply “not nearby”.

The Ngram Viewer allows us, for any of the search words, to examine the books in which the search word occurred. Here is part of what we see if we ask to look at books from 1911-1924 using only “far out”. Simply click on the years where you want to see the books.

Search in Google Books:

<a href="#">1800 - 1929</a>	<a href="#">1930 - 1987</a>	<a href="#">1988 - 1992</a>	<a href="#">1993 - 1997</a>	<a href="#">1998 - 2000</a>	<a href="#">awesome</a>	English
<a href="#">1800 - 1914</a>	<a href="#">1915 - 1990</a>	<a href="#">1991 - 1994</a>	<a href="#">1995 - 1997</a>	<a href="#">1998 - 2000</a>	<a href="#">nifty</a>	English
<a href="#">1800 - 1834</a>	<a href="#">1835 - 1910</a>	<a href="#">1911 - 1924</a>	<a href="#">1925 - 1975</a>	<a href="#">1976 - 2000</a>	<a href="#">far out</a>	English
<a href="#">1800 - 1810</a>	<a href="#">1811 - 1821</a>	<a href="#">1822 - 1858</a>	<a href="#">1859 - 1939</a>	<a href="#">1940 - 2000</a>	<a href="#">stupendous</a>	English

Let's check some of the books, published between 1911 and 1924, in which the phrase **far out** appears, and try to determine if they were using modern slang back then!

## The New Nature Library - Volume 2 - Page 363



<https://books.google.com/books?id=Go0sAQAAMAAJ>

1914 - [Read](#) - [More editions](#)

**FAR OUT AT SEA** "**Far out** at sea—the sun was high, While veered the wind and flapped the sail; We saw a snow-white butterfly Dancing before the fitful gale **Far out** at sea. The little wanderer, who had lost His way, of danger nothing knew; ...

## The Far Triumph - Page 11



<https://books.google.com/books?id=dEcgAAAAMAAJ>

Elizabeth Dejeans - 1911 - [Read](#) - [More editions](#)

It jutted **far out** from the hillside, its front as clean cut as a slab of marble. At its summit were several ledges and irregular projections; it was on one of these ledges that he must have seen the bit of red. It was either gone now, or hidden from ...

## Handbook of Nature-study for Teachers and Parents: Based ...



<https://books.google.com/books?id=pnVNAAAAYAAJ>

1922 - [Read](#) - [More editions](#)

How does it act for the first two or three hours? How does the empty chrysalis skin look? A BUTTERFLY AT SEA **Far out** at sea — the sun was high. While veered the wind and flapped the sail; We saw a snow-white butterfly Dancing before the ...

## Handbook of Nature-study for Teachers and Parents, Based ...



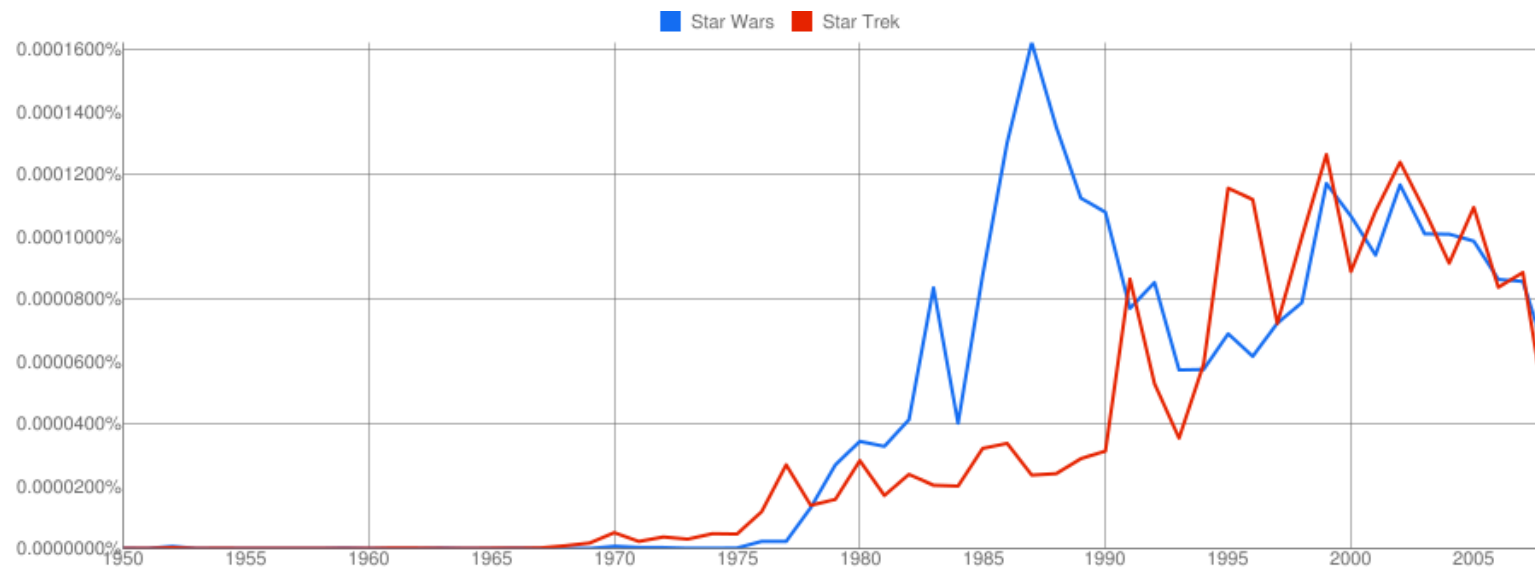
<https://books.google.com/books?id=OHoeAAAAMAAJ>

Anna Botsford Comstock - 1911 - [Read](#) - [More editions](#)

## Example. Star Wars versus Star Trek

Now suppose we want to know the history of the usage of the terms “Star Wars” and “Star Trek”. We simply enter these two bigrams separated by a comma.

We expect to see both terms set to zero until quite recently.



It's interesting to look at who's "winning" from year to year.

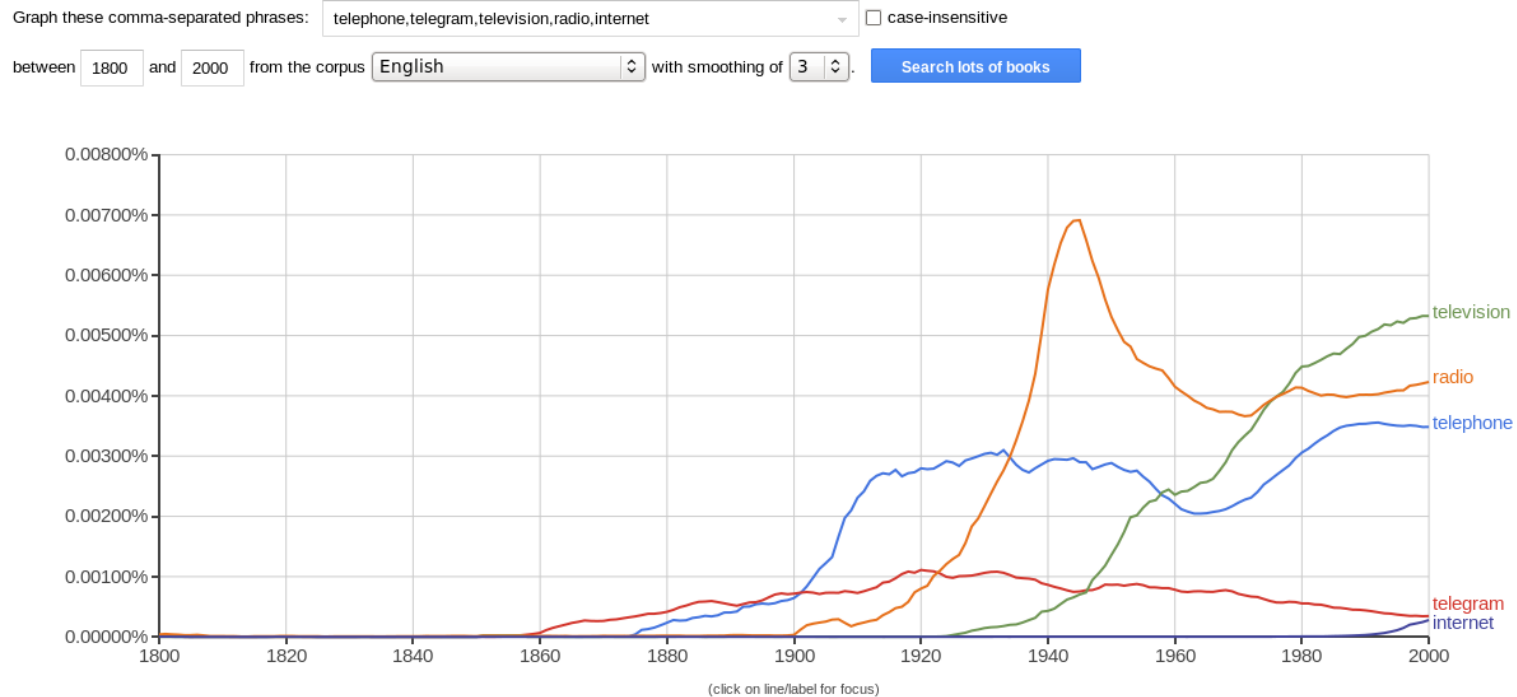
Also, Star Trek has peaks around 1978, 1980, 1985, 1991, 1995, 1999, 2002, 2004. Star Trek movies were released in 1979, 1982, 1984, 1986, 1989, 1991, 1994, 1996, 1998, 2002...

Star Wars shows peaks at around 1980, 1983, 1987, 1999 and 2002. Movies were released in 1977, 1980, 1983, 1999, and 2002... Notice that the Ngram Viewer seems to have completely ignored the first movie!

However, note that if there are less than 40 mentions of a topic in a given year, this is actually rounded down to zero.

**Example.** Now suppose we want to compare the relative usage of the words “telegram”, “telephone”, “television”, “radio” and “internet.” We enter these words in the search box separated by commas.

## Google Books Ngram Viewer



Some things make sense:

- telegram pops up first, and decreases over time
- telephone is next
- radio zooms up into the 1940's and then drops to a plateau
- television shows up early, in the 1920's, and goes up and up

But what's going on with the internet? Why are its numbers so low?



Ngram Viewer is case sensitive unless we tell it not to be.

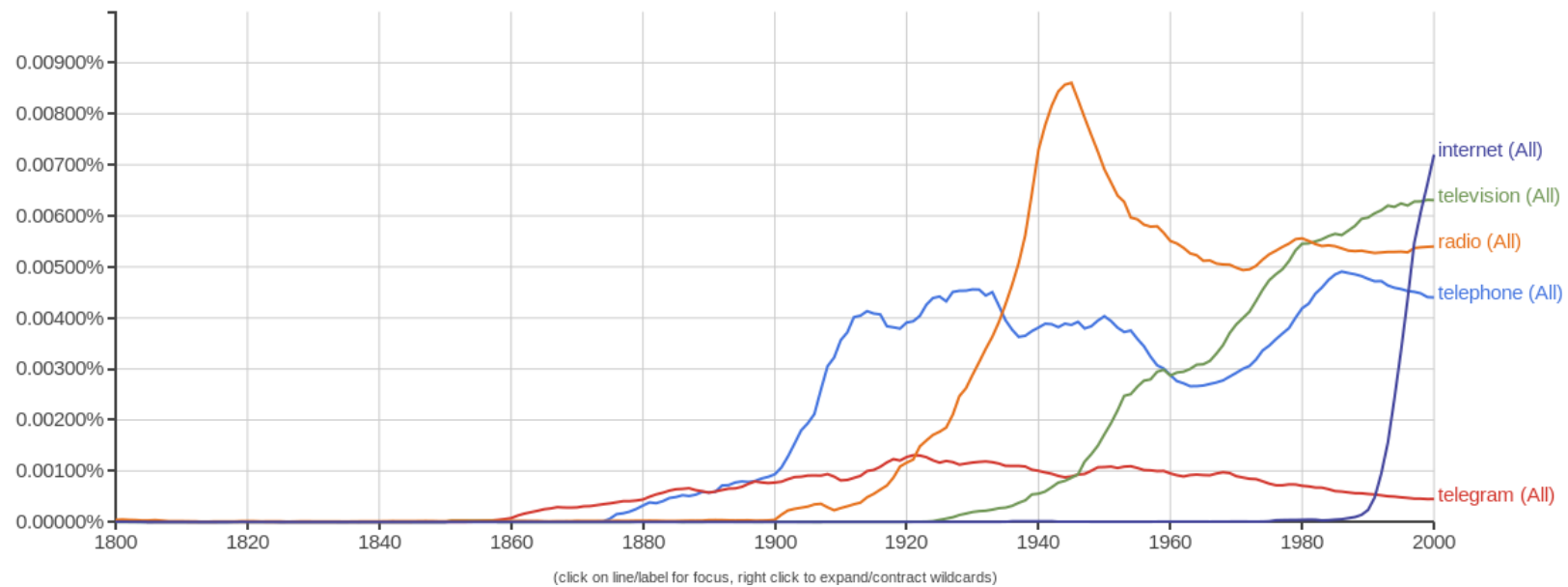
That means that if we search for `internet` then `Internet` will not count.

Simply by checking the case-insensitive box, we can repeat our search, but now look at the result. Apparently, “radio” is rarely capitalized , but almost all mentions of the internet capitalize it as “Internet”! Note that the case-insensitive box is to the right of the box containing the search words; simply click it.

# Google Books Ngram Viewer

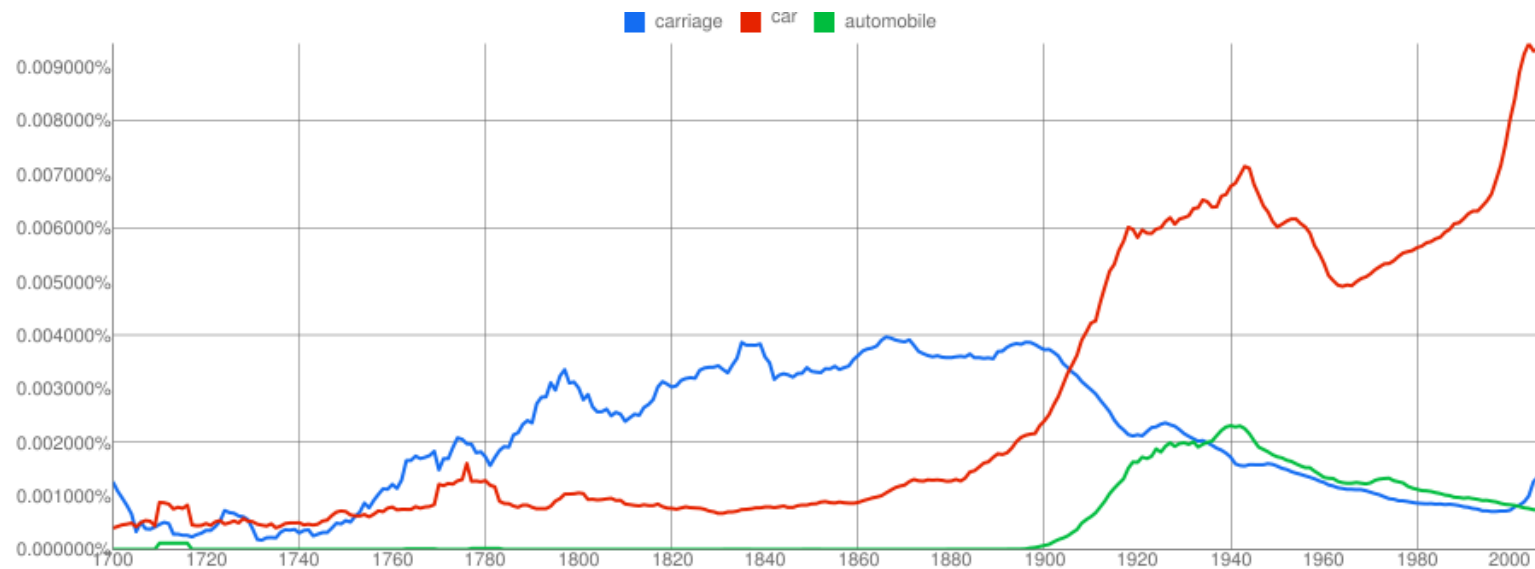
Graph these comma-separated phrases:  ☒ case-insensitive

between  and  from the corpus  with smoothing of  [Search lots of books](#)



**Example.** Compare the usage of the words “carriage”, “automobile” and “car”.

Since carriages were in use before 1800 we extend our search dates to 1700 to 2000.



We might suspect that “carriage” would disappear around 1900, just as “automobile” and “car” suddenly pop up.

The actual graph is more complicated. “carriage” doesn’t want to die out...but then again, “carriage” has other meanings besides something a horse pulls.

But did people in the 1800’s actually have cars? I don’t think so! Did they call a carriage a car back then? Seems unlikely! What is going on?

If we look at the text from some of the books from the 1800s we can see what is happening.

Specification of a patent for a Pendulous Rail-road Car

Isabel Trevithoe, a poem by C.A.R., 1879

A Key to the Classical Pronunciation of Greek, 1830, including Car-nus, Car-nu'tes, Car-pa'si-a, ...

CAR, verb, to cover the cop

the case of Habeas Corpus in the 3d of Car. 1 (*the third year of the reign of King Charles I*)

Carcase (car-case), n. dead body, body,

instructions for making a flying car, in which a man may sit

Rahla turned the car on and started to back out of the parking place (*1881? No way! Mistaken date!*)

Unfortunately, when we look at the data, we see some very peculiar things!

A train consists of railroad cars, so there were things called cars back in the 1800's.

A poem, written by a person with initials C.A.R., is listed as a hit.

A list of Greek words, hyphenated, produces a string of "car" hits.

An OCR mistake reads "CAP" as "CAR"

English laws are dated by the king's name, in Latin, abbreviated, so Charles I is Car. I

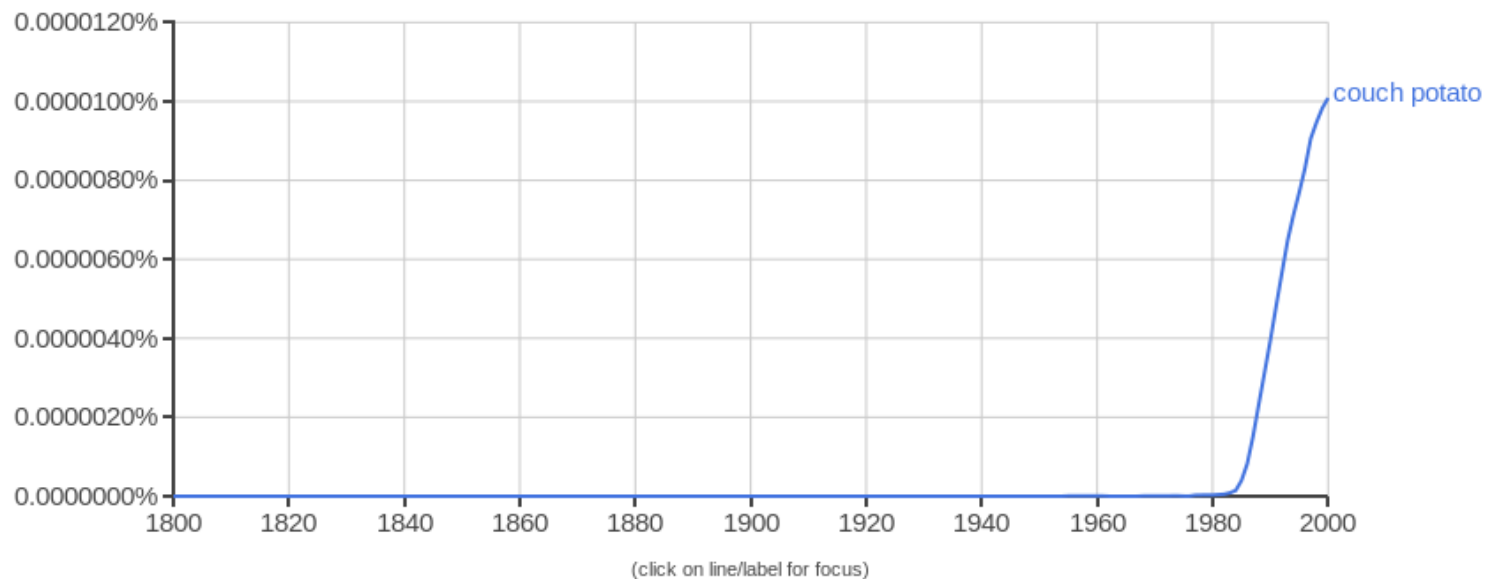
And worst of all, a book published in 1981 was mistakenly listed as 1881, and so we have people driving around in a car back then!

**Example.** Where and when did the term “couch potato” arise?

The phrase “couch potato” seems to be very modern. Can we estimate when it appeared?

## Google Books Ngram Viewer

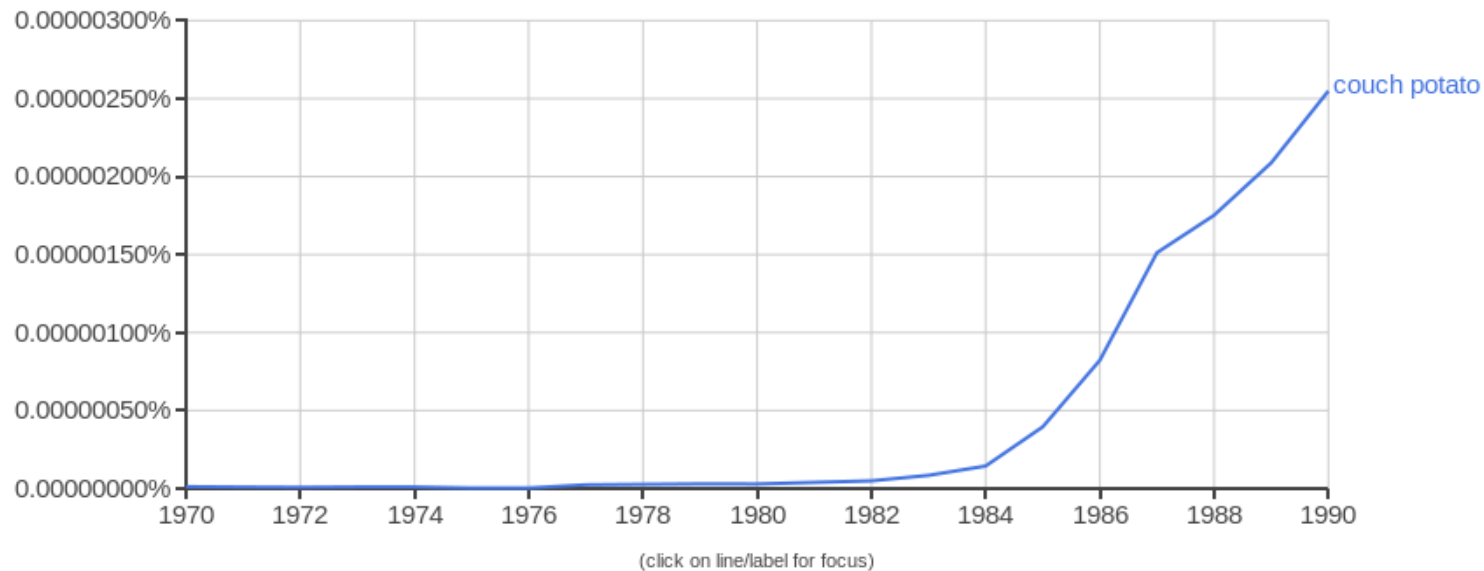
Graph these comma-separated phrases:  ☐ case-insensitive  
between  and  from the corpus  with smoothing of  [Search lots of books](#)



Our first search shows a dramatic growth in usage, but this can make it hard to see the origin, since the usage would typically be very low early on. Let's narrow the year range.

## Google Books Ngram Viewer

Graph these comma-separated phrases:  ☐ case-insensitive  
between  and  from the corpus  with smoothing of  [Search lots of books](#)





When we go into the book search, the closest thing we can find is an article in Texas Monthly, for April 1986.

Reading the text gives us some assurance that the phrase is being used in the way we expect.

Even though Ngram Viewer seems to show usage before 1980, the 1986 reference is the earliest I could spot in the database.

Because this is a magazine, we have confidence that the date is given correctly.

If it was a book, we would want to go to the first couple pages and look at the copyright page, because Google Books is full of incorrect date information!

Of course, another way to search is to just use a browser, and in this case, it turns out we can find some information about a man who claims to have invented the phrase back in the 1970s.

So Internet browsers can be a quicker way to find some kind of facts...on the other hand, we never would have been able to trace the growth in popularity of the phrase, or all its occurrences, without using Ngram Viewer!

---

## Maybe we found him!

---

# Meet Tom Iacino, the Man Who Coined the Phrase 'Couch Potato'

WRITTEN BY [ROCHELLE BILOW](#)

 PRINT  RSS

When we explored the [history of vegetable metaphors](#), one brief investigation left us with the nagging feeling that we needed to know more. The [facts behind the phrase](#) “couch potato” seemed shrouded in mystery, so we dug a little deeper. In our quest to get to the bottom of things, we ran straight into **Tom Iacino**, the man credited with coining the phrase. We caught up with Iacino on the phone, from his home in California, to hear his side of the story.



Tom Iacino, back in the day

**You’re an elusive guy; there’s not a lot of information to be found about the origin of “couch potato.”**

Well, this doesn’t come up that much so I can see that there’s probably not a whole lot out there. But when the Oxford Dictionary was doing their research work, they discovered that [my friend] **Bob Armstrong** had the trademark for the term “couch potato.” My part in this was that it was just an utterance I made to close friends. Bob, who was a cartoonist at the time, was looking for something to sum up his feelings about, I don’t know, what he was doing and feeling, I guess, and the couch potato just seemed perfect to him. So he asked me if he could use it and draw it, and of course I said yes.

---

## A few more features of Ngram Viewer

---

- Currently, the earliest year you can enter is 1500.
- Currently, the latest year you can enter (using English books) is 2008.
- The corpus of books defaults to English, but a drop-down menu can change that.
- If you put the cursor on a point on a curve then a window will appear with the actual statistics for that year.
- For homework, you will be asked to create plots using the Viewer. To save a plot without including all the extra information on the page, the simplest way is to take a [screenshot](#). The procedure for this is different on Macs and PCs so see the TA if you need assistance (or just Google it).

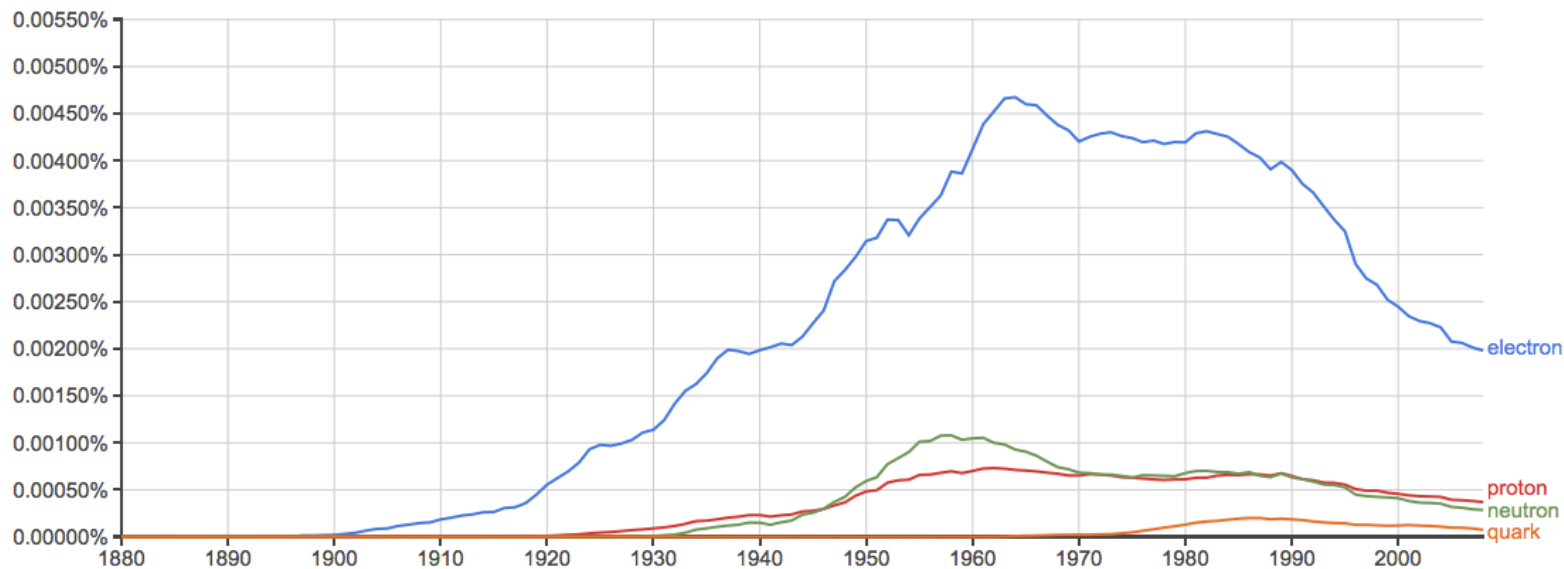
---

## Socrative Quiz Ngram\_Quiz2

CTISC1057

---

For the first 4 questions use the following Ngram Viewer results on the subatomic particles electron, proton, neutron, and quark.



1. Which of the four subatomic particles was named first?

2. Which of the four subatomic particles is the most recently discovered?
3. From 1950 to 1970 did “proton” or “neutron” occur most often?
4. From 1995 to 2005 did “proton” or “neutron” occur most often?
5. The *y*-axis on the Ngram Viewer gives the range of years for the publication date of books you are searching.
6. Every appearance of an n-gram is reported, even if it only appears once.
7. The range of years for publication dates of books can be adjusted.
8. Some occurrences of n-grams before their actual use are due to cataloging mistakes.
9. The phrase “You miss one hundred percent of the shots you never take” contains 11 distinct 1-grams.
10. The phrase “You miss one hundred percent of the shots you never take” contains 10 distinct 2-grams.