

DATA 136 - Intro to data engineering

Spring 2025

Indexing and costs

17,700 Movies

in the

Netflix Competition

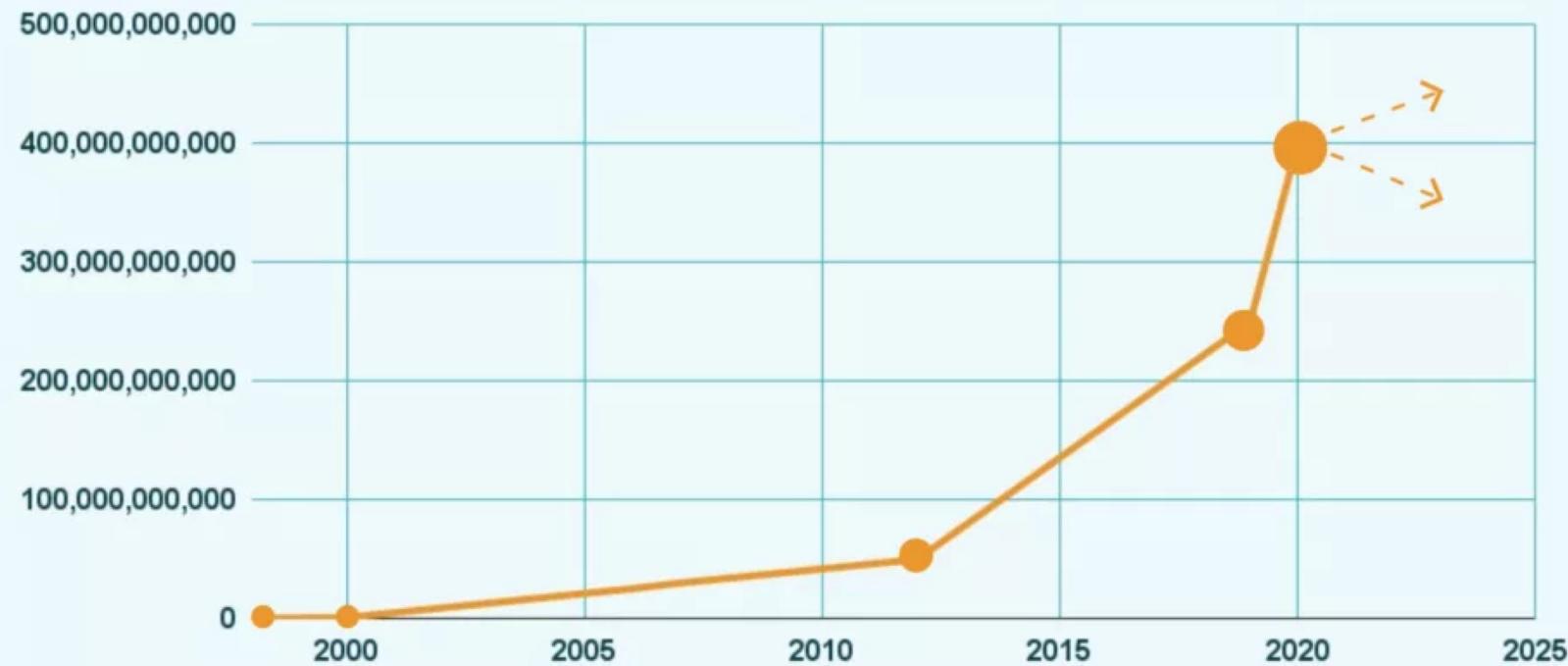
Todd.Holloway@gmail.com 03/25/2007

How many steps to look something up?



Motivation: search is getting harder

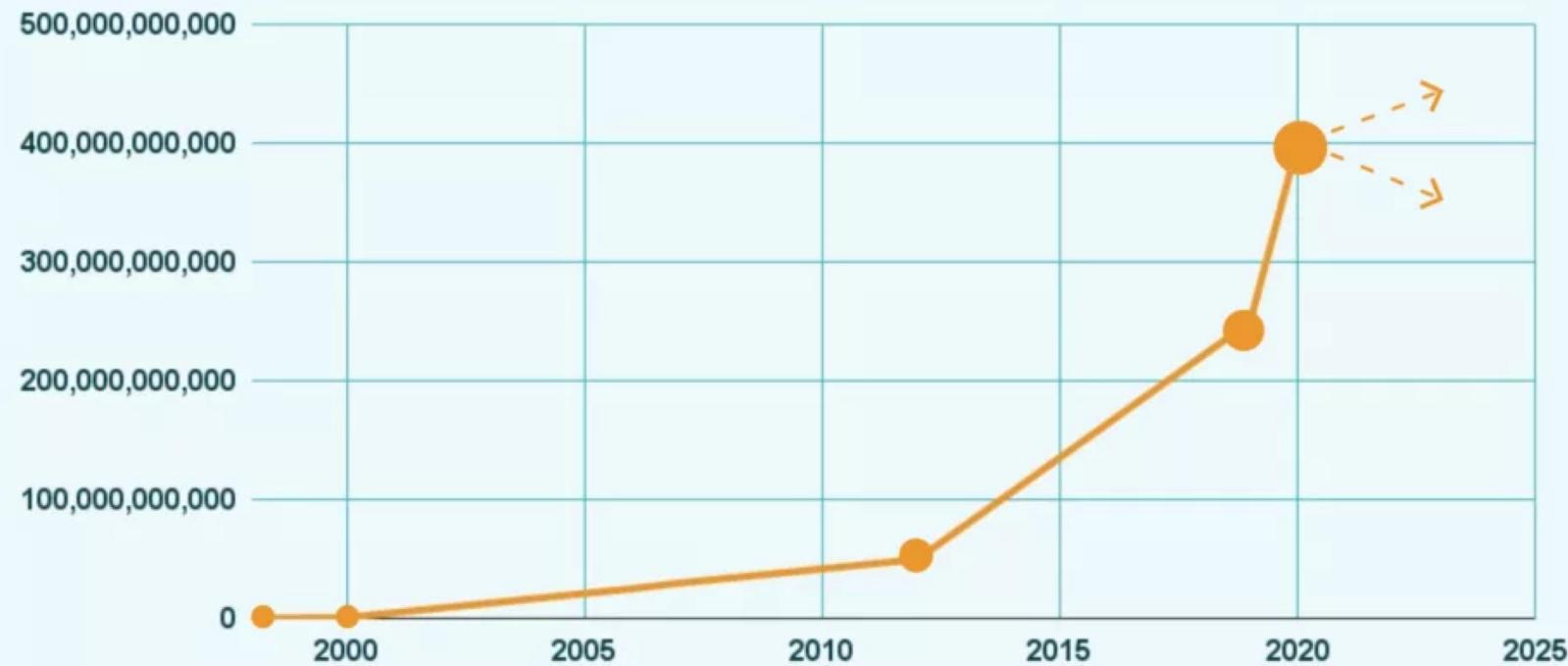
Google's Index Size Over Time



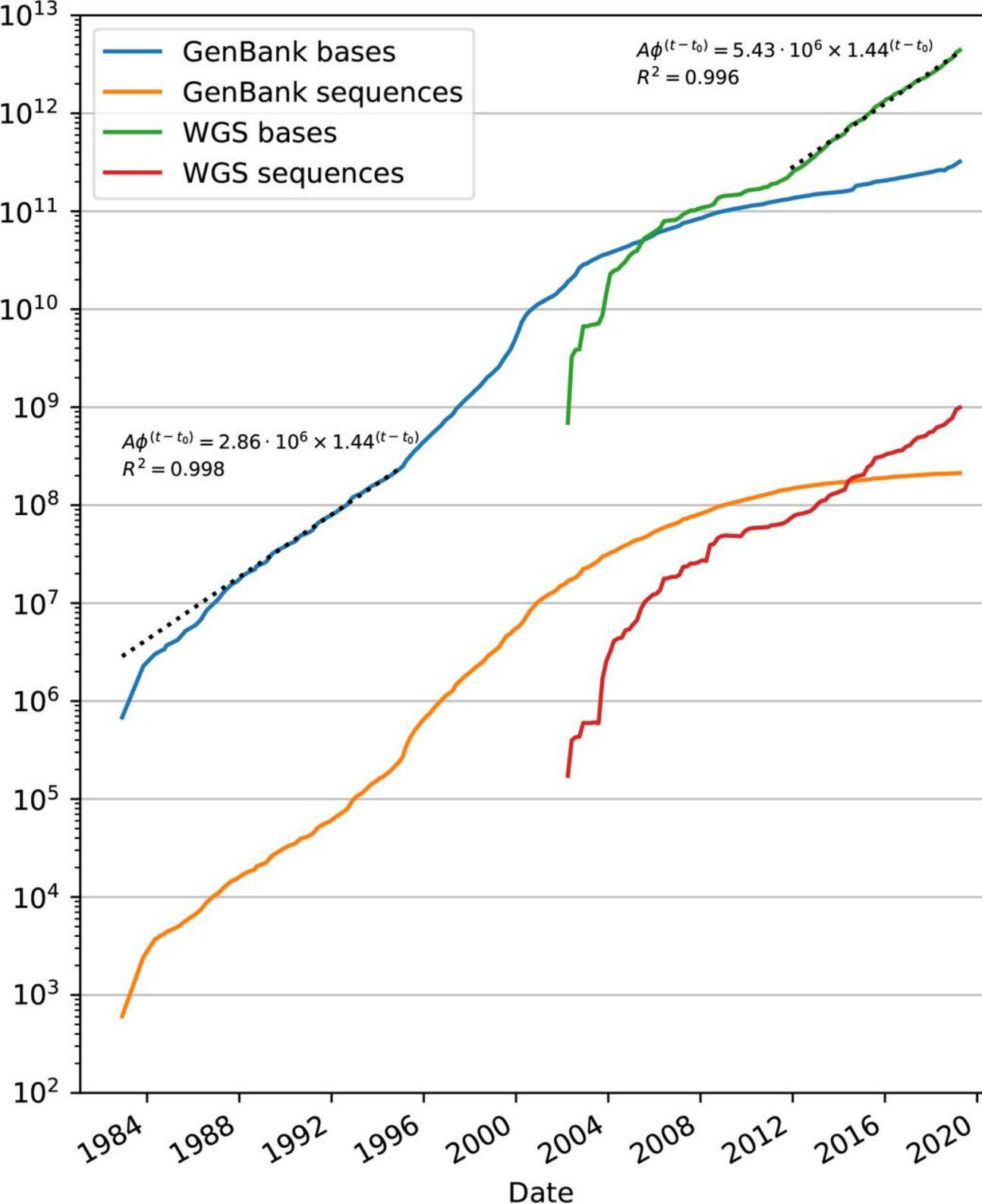
ZYPPY SEO

Motivation: search is getting harder

Google's Index Size Over Time

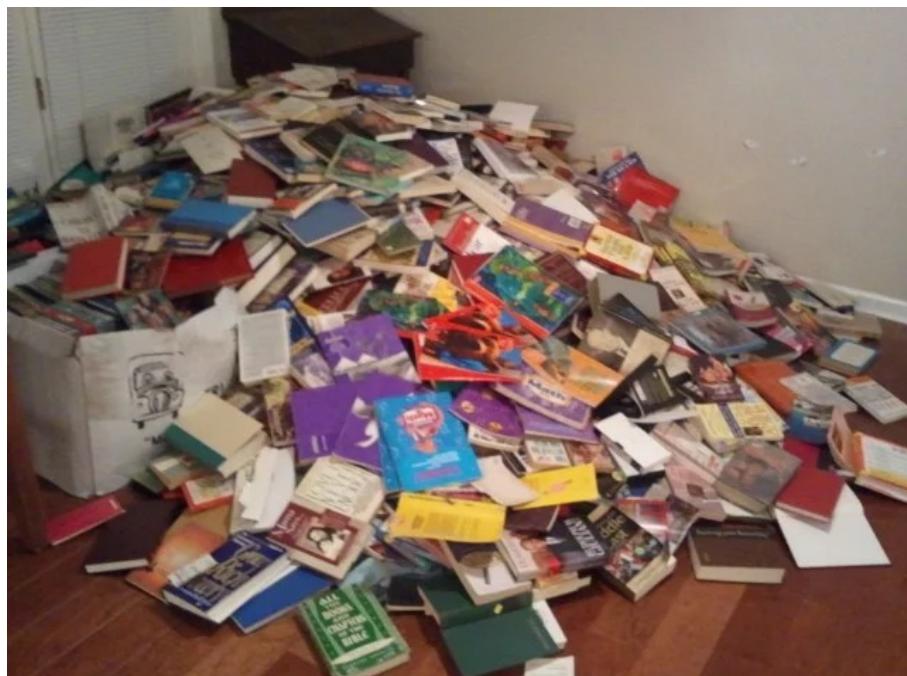


ZYPPY SEO



Genome
database
searching
is getting
harder, too.

How many steps to look something up?



Suppose I have a pile of books, in random order, and where my only access is by choosing a book (randomly).

Contrived? Yes...

How many books do I need to look at on average in a pile of size N to find a specific book (the one with the dedication from my high school biology teacher) ?

pseudocode for random search (an ineffective search strategy)

```
while sample_book != target_book:  
    sample_book = get_new_book()
```

What happens if it can't find my book?

pseudocode for random search (with MAXTRIES)

```
sample_book = get_new_book()  
n = 0  
  
while sample_book != target_book and n<MAXTRIES:  
    n +=1  
    sample_book = get_new_book()
```

How many steps to look something up?



Suppose I have a pile of books, in random order, but I can make sure I never check the same book twice

How many steps to look something up?



Suppose I have a pile of books, in random order, but I can make sure I never check the same book twice

My number of books to check is Uniform ($1/N$) which has a mean of $N/2$

```
while sample_book != target_book:  
    sample_book = get_new_book()
```

How many steps to look something up?



What if the books are sorted?

How many steps to look something up?



What if the books are sorted?

Binary search:

1. Look in the middle of the bookstack
2. compare the call number to target
3. If this is your book, terminate
4. else look in the upper (lower) half of the bookstack and repeat step 1.

How many steps?

How many steps to look something up?



```
def bisect(MIN, MAX, target):
    n_book = get_nearest_book( (MAX + MIN) / 2)
    if n_book.index < target_book.index:
        bisect (MIN, (MAX+MIN)/2, target)
    if n_book.index > target_book.index:
        bisect ((MAX+MIN)/2, MAX, target)
```

What if the books are sorted?

Binary search:

1. Look in the middle of the bookstack
2. compare the call number to target
3. If this is your book, terminate
4. else look in the upper (lower) half of the bookstack and repeat step 1.

How many steps?

How many steps to look something up?



```
def bisect(MIN, MAX, target):
    n_book = get_nearest_book( (MAX + MIN) / 2)
    if n_book.index < target_book.index:
        bisect (MIN, (MAX+MIN)/2, target)
    if n_book.index > target_book.index:
        bisect ((MAX+MIN)/2, MAX, target)
```

What if the books are sorted?

Binary search:

1. Look in the middle of the bookstack
2. compare the call number to target
3. If this is your book, terminate
4. else look in the upper (lower) half of the bookstack and repeat step 1.

How many steps?

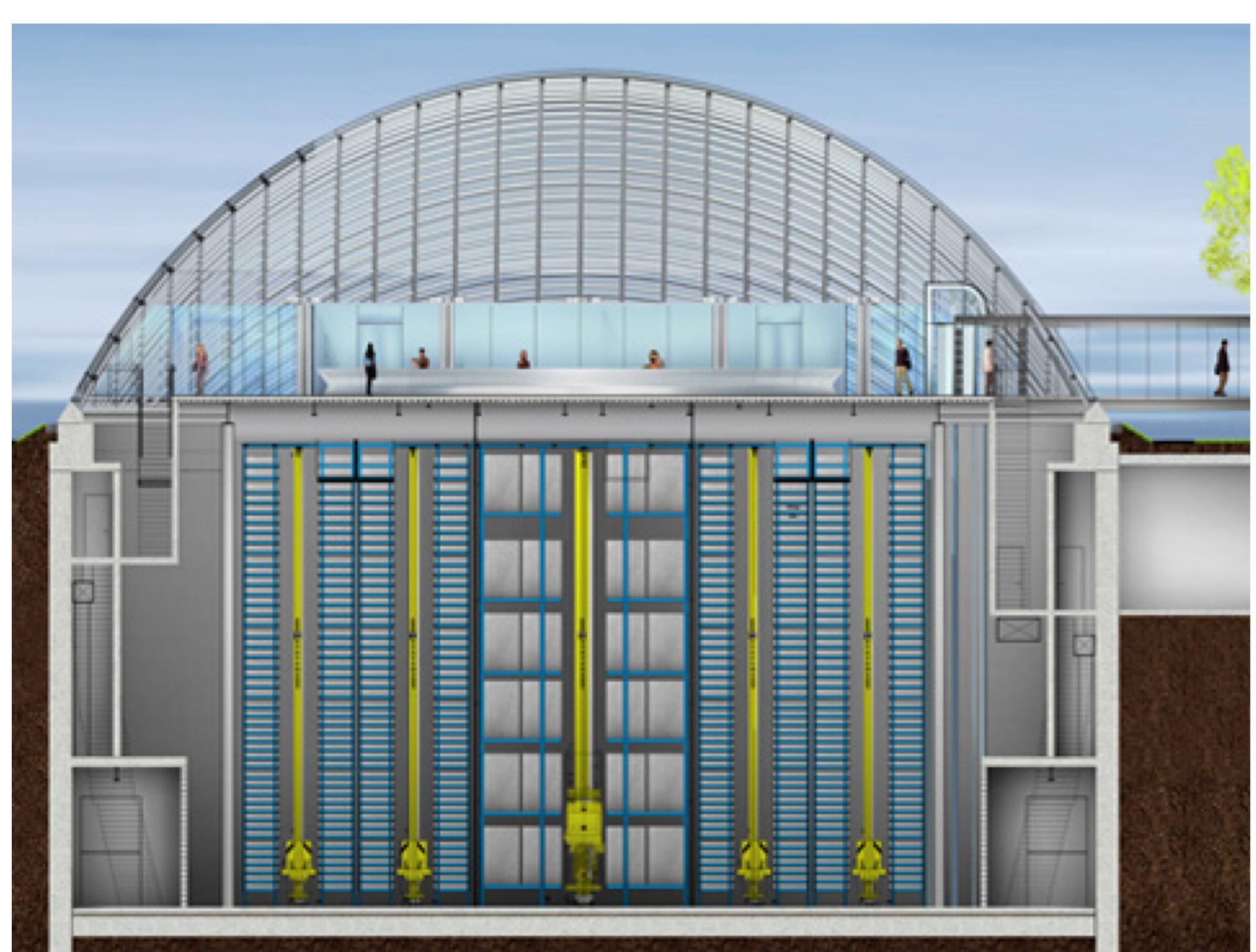
$\text{ceiling}(\log_2(N))$



How does Mansueto retrieve books?

Imagine that the location of every book is known.
Retrieving a book has just two steps. Find out the location, and go to the location and get the book.

How does the cost of this procedure depend on the number of books in the library?





Big-O notation

- Way of expressing the functional dependence, in the limit of very large N, of **the number of computing operations** to achieve an analytical result on the sizes of the query (k) and of the database (n)
- Some analytical operations may become easier with larger query sets, like birthday problem or "crack the first password"
- Describes the behavior of the performance function in the limit of large N and large k.

Amortizing indexing costs

```
cat database.csv | grep -i "Harry Potter and the  
Plot Hole"
```

We can solve this (responding to a search query) faster by an initial investment in an index.

Lookup cost = cost of indexing + k cost per query

Wk4-indexing.py

What tricks do we have?

Type	setup	query time
Unordered list (naive search)	None	$k O(N)$
Ordered list	$k O(N \log(N))$	$k O(\log(N))$
Data at known address	$O(N)$	$O(1)$



Google books has ~40 million books.

Books have ~100k words.

So, we're looking at 4×10^{12}
words in database.

The entropy of printed English is about 10 bits per word, corresponding to a "vocabulary size" averaged over usage of about $2^{10} = 1024$.

The search engine parlor trick

- So how long does a phrase need to be "on average," whatever that means, to uniquely specify a location in the ~~genome~~ google books corpus?

$$\log_2 (4 \times 10^{12}) = 41.9 \text{ bits.} = \log(\text{words in db})$$

$$H(\text{single word}) \sim 10 \text{ bits} = 1/1024$$

Would you believe, 5 words??

You can't hide embarrassing utterances from the search engine by disappearing into the crowd.

was not in charge this thing is over[.] Trump is in to the end.”²⁷ The next day, consistent with

P1 description, the defendant announced his staff change by Tweet, writing, “I look forward to CC1 spearheading the legal effort to defend OUR RIGHT to FREE and FAIR ELECTIONS! CC1 P10 P11 CC3 and P12

a truly great team, added to our other wonderful lawyers and representatives!”²⁸

was not in charge this thing is over[.] Trump is in to the end.”²⁷ The next day, consistent with

P1 description, the defendant announced his staff change by Tweet, writing, “I look forward to CC1 spearheading the legal effort to defend OUR RIGHT to FREE and FAIR ELECTIONS! CC1 P10 P11 CC3 and P12

a truly great team, added to our other wonderful lawyers and representatives!”²⁸

1:23-cr-00257-TSC Document 252 Filed 10/02/24



Donald J. Trump 
@realDonaldTrump

...

I look forward to Mayor Giuliani spearheading the legal effort to defend OUR RIGHT to FREE and FAIR ELECTIONS! Rudy Giuliani, Joseph diGenova, Victoria Toensing, Sidney Powell, and Jenna Ellis, a truly great team, added to our other wonderful lawyers and representatives!

9:11 PM · Nov 14, 2020

38K

45K

220K

631

↑

- Twitter: 200 billion tweets per year, 50 words, each, gives a database of the same scale, 10×10^{12} words.
- Little snippet of text in a screenshot of an article "conclusions that someone else's" will lead you to <https://doi.org/10.1038/nature.2016.19269>
- **Your content**, unless very short or derivative ("It's only a scratch!") **is probably unique**. If it is searchable, scrubbing out the username is ineffective at hiding your mistakes.
- Search for snippets of the fulltext to look for copies of a [book, article, etc.] in search engine.