



# Google and the Death of Books



**Michael I. Shamos, Ph.D., J.D.**  
Director, Universal Library  
Language Technologies Institute  
Carnegie Mellon University



# Judge puts off ruling on Google's proposed digital book settlement

By [Cecilia Kang](#)

Washington Post Staff Writer

Friday, February 19, 2010

NEW YORK -- Google confronted a barrage of criticism from opponents of its proposed digital book settlement Thursday as the Internet search giant tried to persuade a federal judge to approve a deal that would allow it to create the world's largest online library.

The opponents also argued that the [\\$125 million settlement](#) -- which would allow Google to scan and publish millions of out-of-print titles -- could give the company an unfair edge over other online publishers in the nascent but exploding market for digital books.

# Outline

- Books as physical artifacts
- Google's e-library replacement
- The Google "settlement"
- Piracy
- Causes of book death

# Disclosure

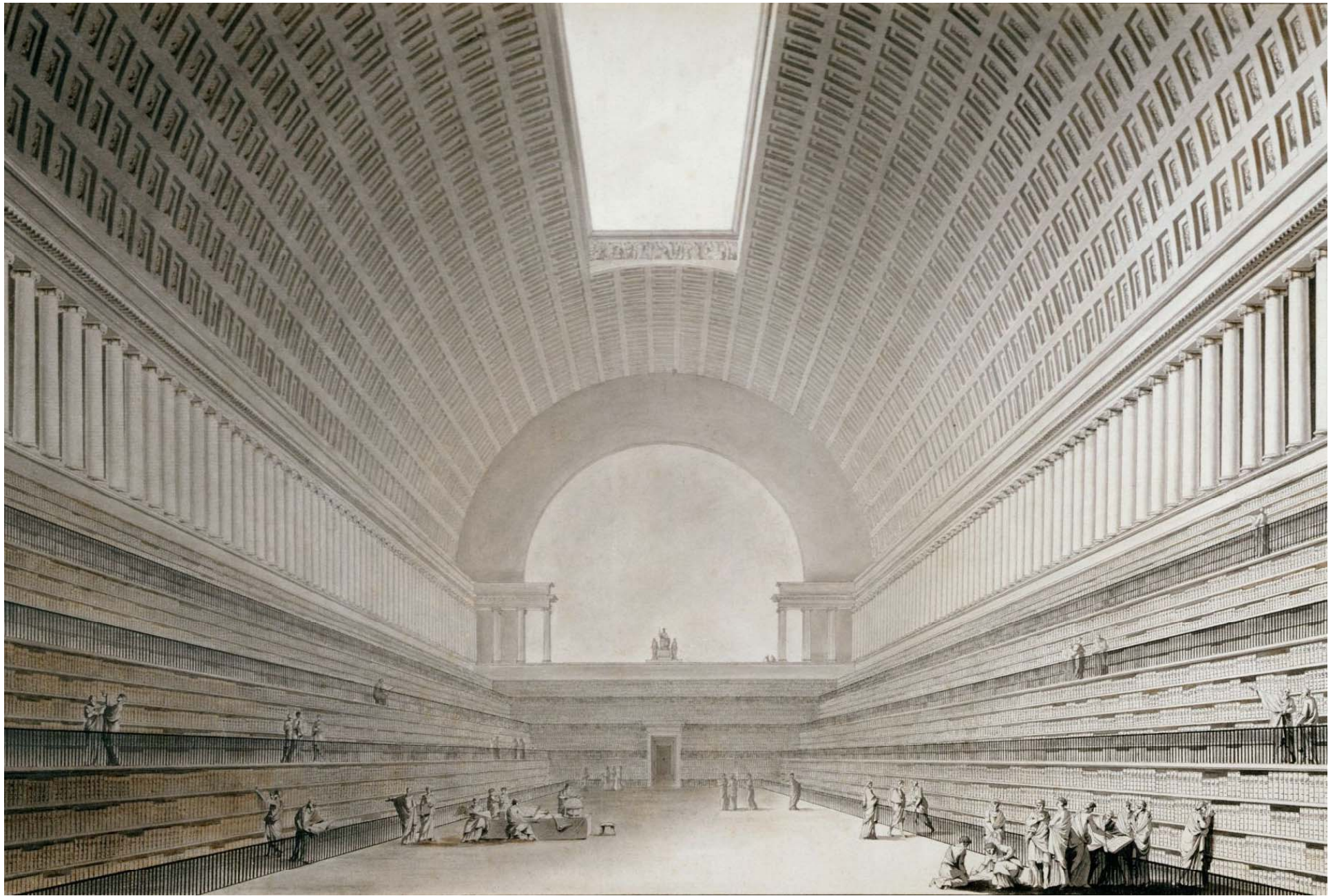
- Expert witness for Google in a patent case, *Performance Pricing v. Google*, in the United States
- Case involves the real-time algorithm Google uses to price its advertisements
- No relation to Google Books
- No copyright issues involved

# In the Beginning ...

- To distribute information:
  - People cut down trees to make paper sheets
  - Marked the paper with black ink
  - Sewed the sheets together with thread
  - Put sheets between pieces of cardboard
  - Pasted everything together
- 
- And called it a “book”

# In the Beginning ...

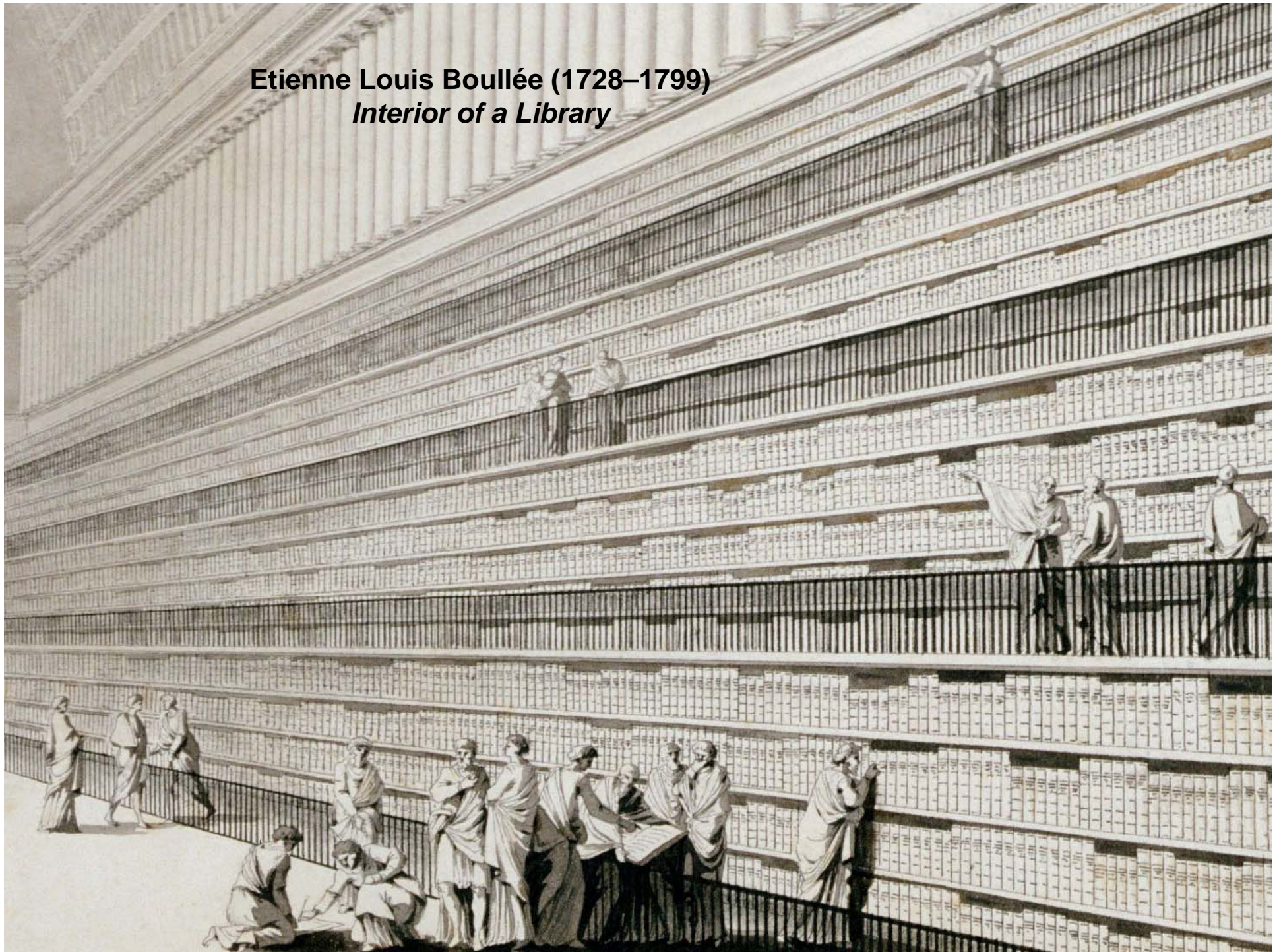
- Books became very valuable
- There was no other way to learn the information they contained
- Books were heavy and expensive
- Most people couldn't own many
- They were stored in centralized places called "libraries"
- From latin "librarium" – a chest for books



Etienne Louis Boullée (1728–1799), *Interior of a Library*



**Etienne Louis Boullée (1728–1799)**  
*Interior of a Library*





# Libraries Are Expensive

- It costs HK\$50 per year to store a book in a library
- Total number of books in all public libraries in the world is about 1 per person
- The world spends about HK\$300B per year storing books that fewer and fewer people access

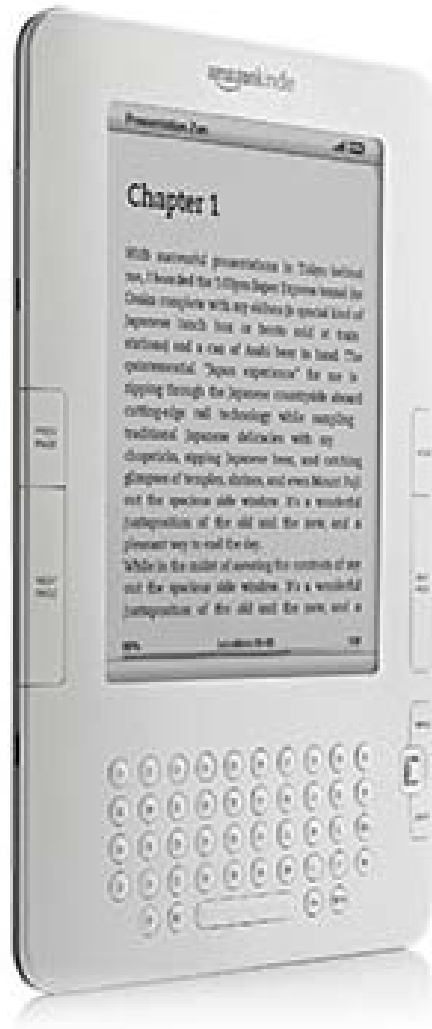
# The Problem with Books

- Heavy
- Expensive to produce and distribute
- Not searchable
- No hyperlinks
- No audio
- No video
- Can't be updated
- Can't be shared
- Easy to destroy

# In Modern Times ...

- Information became digitized
- Books are now a crude way to carry information
- e-books are becoming popular
- Even e-books imitate the old format
- Publishing will give way to networked forms

# Amazon Kindle



## Kindle

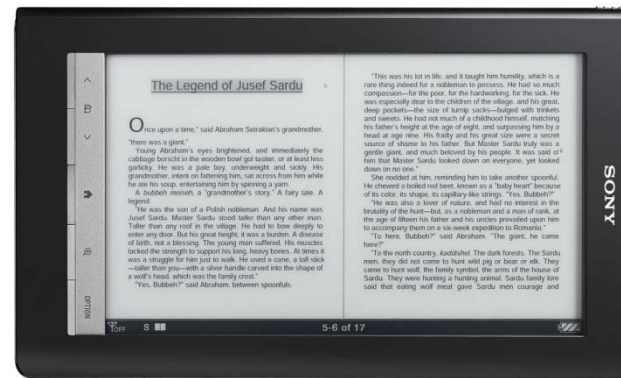
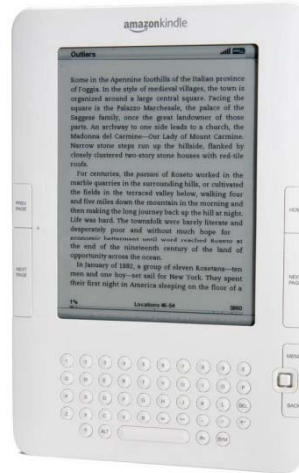
Over 420,000 Books  
to Choose From, Plus  
Thousands More  
for Free.

[Learn more](#) 

**amazon**kindle



# e-Book Readers



FEB. 25, 2010 HKU

GOOGLE AND THE DEATH OF BOOKS

©2010 MICHAEL SHAMOS

# Not All Information Is Digitized

- What to do about all the books that still exist only in physical form?
- Total number of books ever published
  - About 175 million
- Total number digitized
  - About 20 million
- Digitizing creates (big) copyright problems

# Determining Public Domain Status

- 1913: G.B. Shaw writes *Pygmalion* in the UK, died 1950
- 1938: U.K. movie version. Last author (screenplay) died 1997

UK term
  1<sup>st</sup> US term
  2nd or only term
  Public domain
  Restored term

	1913-1938	1938-1941	1942-1966	1966-1988	1989-1995	1996-2020	2021-2033	2034-2067	After 2067
Play (UK)			1950 +50		1950 +70				
Movie (UK)				1968 +50	1968 +70				
Play (US)	First term	28 years	2nd term			Match UK			
Movie (US)		First term	28 years	Not renewed but	play protects movie	47 years	from 1996		

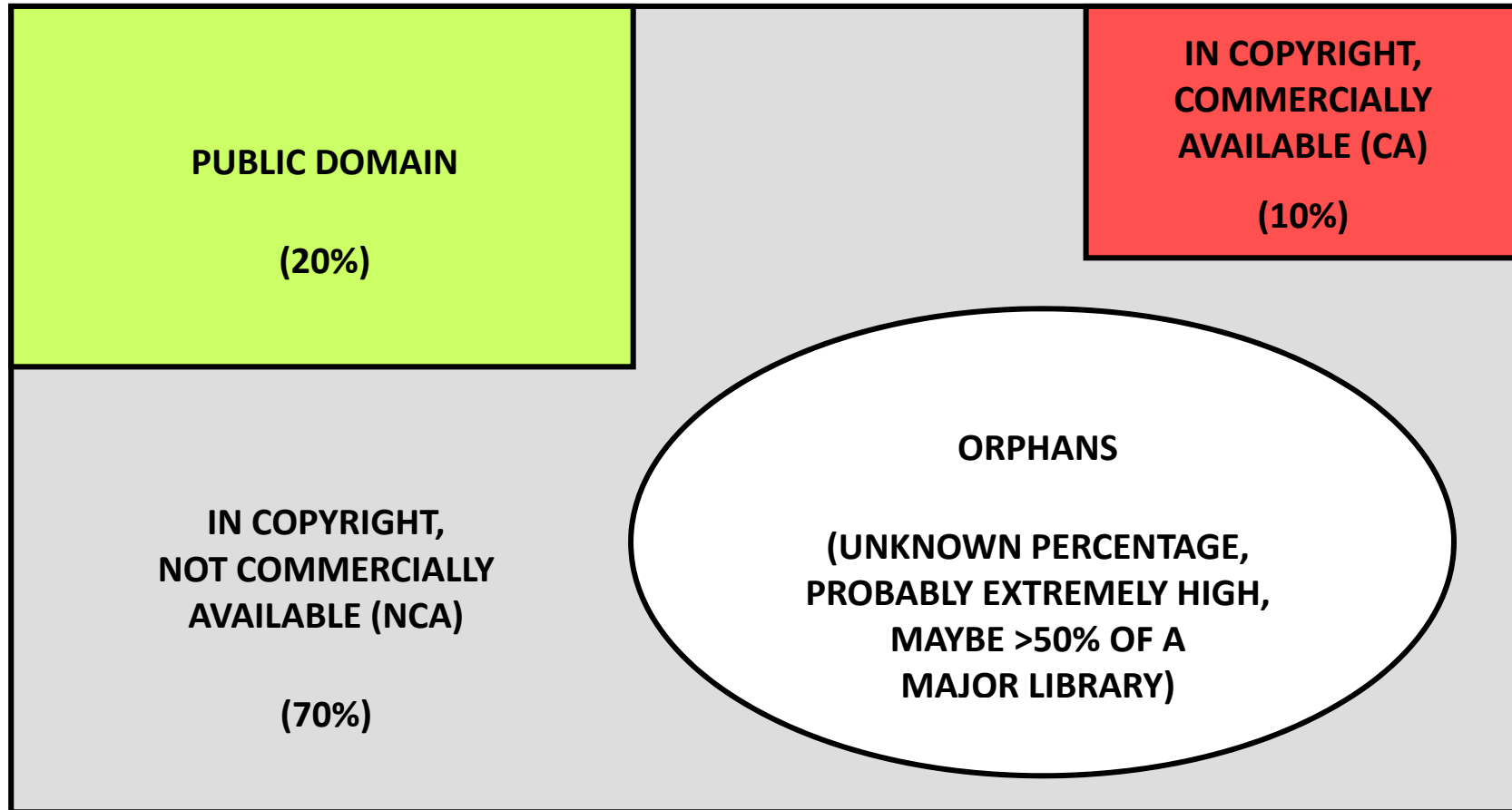
# Orphan Books

- 70% of published books are in copyright but “not commercially available” (NCA)
- A large percentage of books in copyright are “orphans”: in copyright, but no known copyright owner (e.g. publisher went out of business or can’t be found)
- Copyright owner, if it exists, might not know of its ownership
- No one to ask for permission
- How can we recognize an orphan? (Can’t)



# Orphan Books

ALL PUBLISHED BOOKS




# What Google™ Did

- Google scanned and indexed huge numbers of works, mostly in-copyright (estimated 15 million)
- When you search Google Books, your search hit includes a “snippet” of text surrounding your hit:

Google books   [Advanced Book Search](#)

**Speech recognition: invited papers presented at the 1974 IEEE symposium** By Dabbala Raj Reddy



› Overview  
[Reviews \(0\)](#)  
[Buy](#)

☆☆☆☆ (0) - [Write review](#)  
[Add to my library](#)

Get this book  
[Amazon.com](#)  
[Barnes&Noble.com](#)  
[Borders](#)  
[Books-A-Million](#)  
[Find in a library](#)  
[All sellers »](#)

3 pages matching raj reddy in this book

**Part Five. SYSTEMS ORGANIZATION AND ANALYSIS SYSTEMS**

**Tutorial on System Organization for Speech Understanding** 45

*D. Raj Reddy*  
*Lee D. Erman*

computer technology and a continued growth of speech language and hearing research to be integrated in our work, we need something like the diligence of the ants, the persistence of the mad scientist, plenty of time and enthusiasm of the kind exposed at this meeting to keep us on the right track. My congratulations to IEEE and to Raj Reddy and his

Page 457

**TUTORIAL ON  
SYSTEM ORGANIZATION FOR SPEECH UNDERSTANDING**

*D. Raj Reddy*  
*Lee D. Erman*

# World's Largest Libraries

1. U.S. Library of Congress (29M)
2. National Library of China (22M)
3. Russian Academy of Sciences (20M)
4. National Library of Canada (18M)
5. Deutsche Bibliothek (18M)
6. British Library (16M)
7.  (15M)
8. Institute for Scientific Information (14M)
9. Harvard (13M)
10. Vernadsky Institute (Ukraine) (13M)

# What Google™ Did

- Fair use as an information locating tool?
- But when Google announced its scanning project, it said it would make “brief excerpts” available
- “Excerpts” are larger than “snippets”
- Google was sued by authors in the class action *Authors Guild v. Google*
- Google was also sued by publishers in *McGraw Hill v. Google*



# What Google™ Did

- *Editions du Seuil v. Google*
- In December 2009, a French court fined Google 300,000 euros, ordered it to stop scanning copyrighted French works

# What Google™ Did

- Dec. 2009: Chinese Author Mian Mian sued Google for scanning her novel “Acid Lover”
- China Written Works Copyright Society said Google had scanned 18,000 books by 570 Chinese writers without authorization



Google says sorry to Chinese authors

10:56, January 11, 2010

- “Sorry” is not going to be enough

# The Google U.S. “Settlement”

# The Google Settlement

- Revenue-based model
  - Advertising + subscription sales
- Establishes a Book Rights Registry (BRR)
  - In copyright, commercially available (Google will not display)
  - In copyright, not commercially available (Google may display)
  - Public domain (no restriction)
- Google funds BRR HK\$350M to start
- BRR gets 70% of revenue; Google gets 30%

# Settlement for In Copyright, NCA

- FOR FREE VIEWING:
- Always at least 3 “snippets”
- Display up to 20% of text
- Non-fiction, no more than 5 adjacent pages
- Fiction, no more than 15 adjacent pages
- No display for anthologies
- Reference works: fixed 10% preview
- No printing or copy-and-paste (ha!)

# Settlement for In Copyright, NCA

- FOR PAID VIEWING:
- User can buy access to a book
- Owner sets price or allows Google to use a pricing algorithm
- Restrictions on printing and copy/paste
- Very complicated rules for free libraries and institutional subscriptions
- “Non-consumptive research” allowed



# The Google Settlement

- Applies only to books published in US, UK, Australia and Canada
- HK\$500 one-time payment per work for past unauthorized use
- Immunity for Google
- Independent trustee to supervise licensing of orphan works

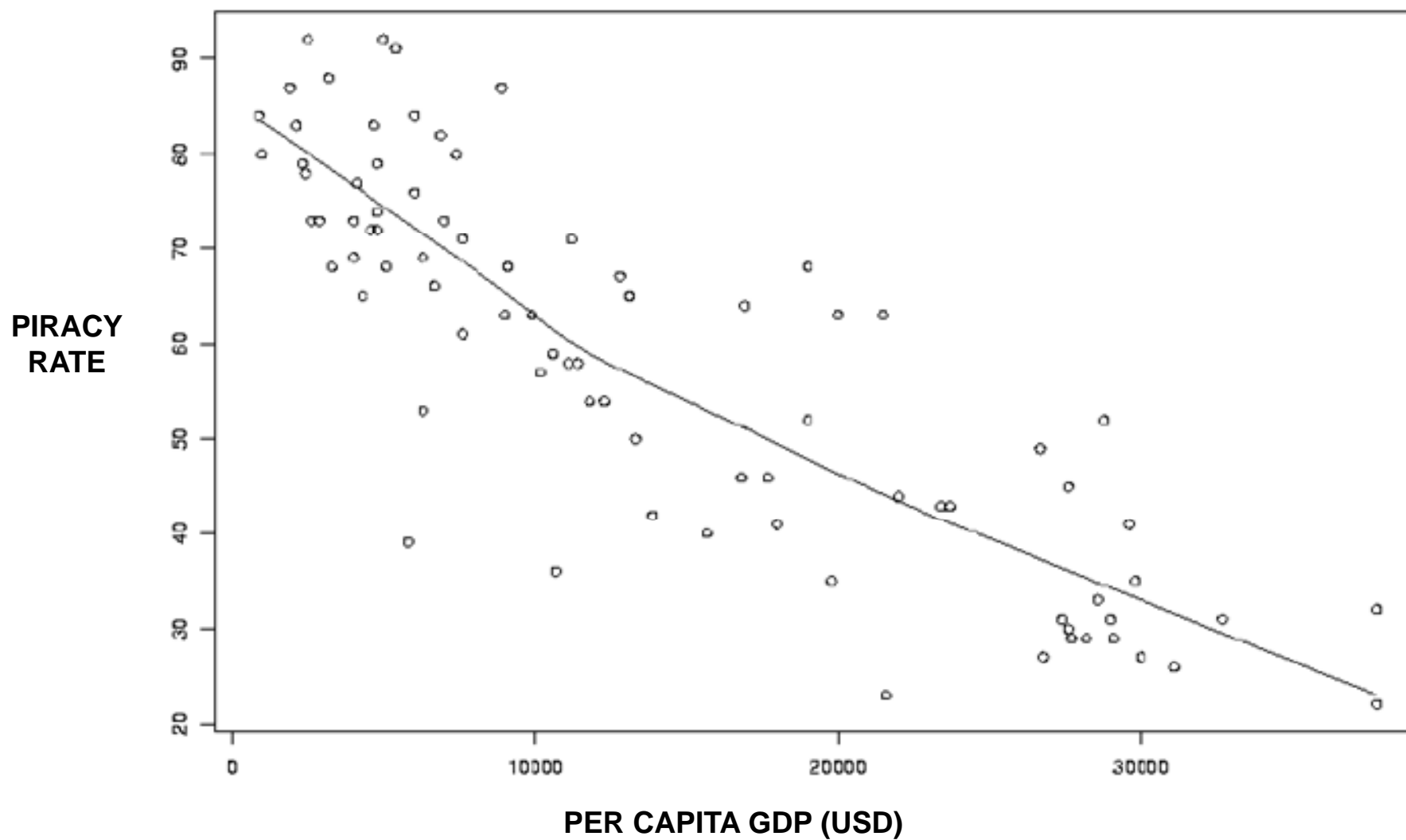
# Piracy



THE INTELLECTUAL PROPERTY  
WARS FROM GUTENBERG TO GATES

*Adrian Johns*

# Piracy rate v. per capita GDP

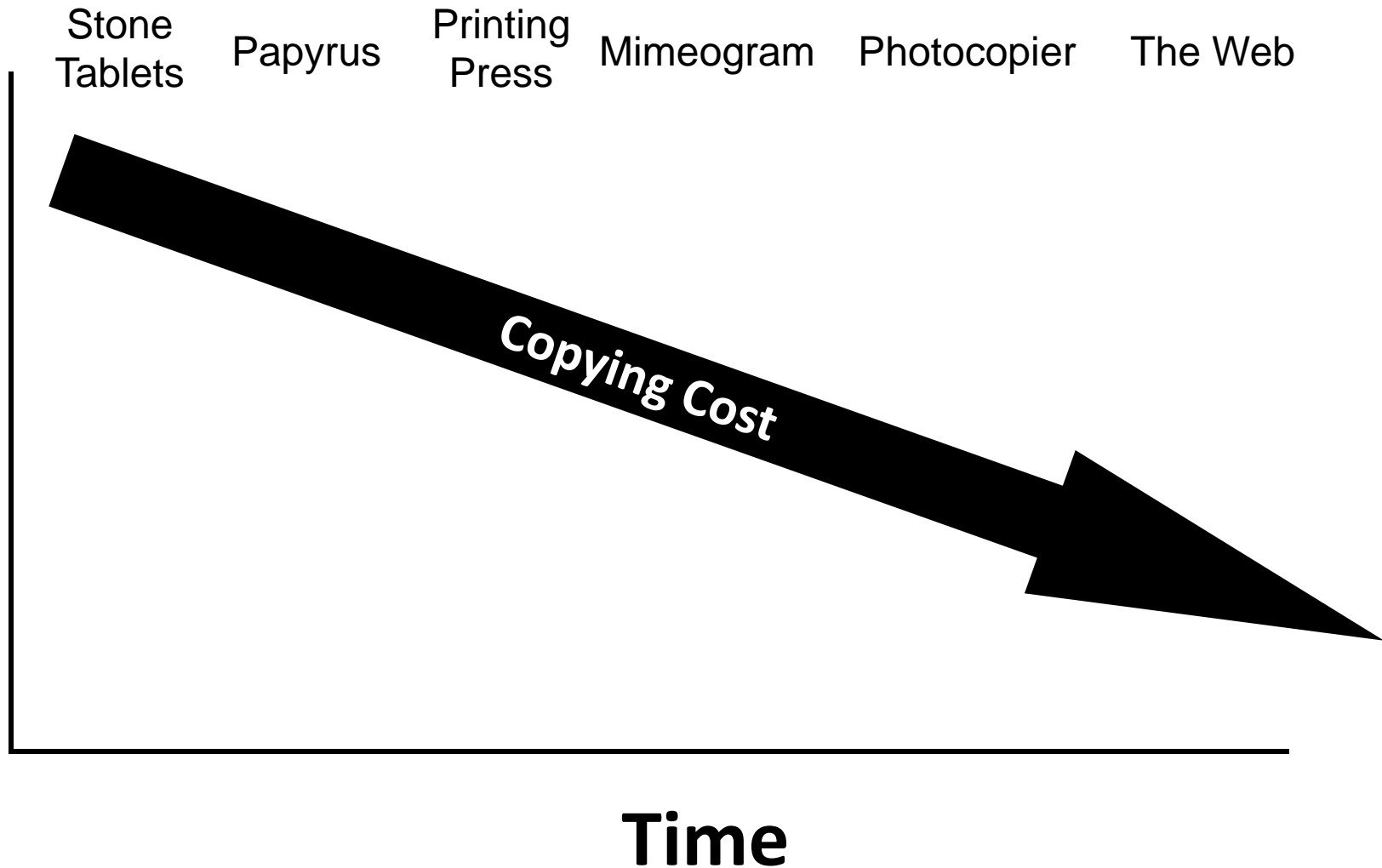


SOURCE: HAL VARIAN, GOOGLE

# Cost of Copying

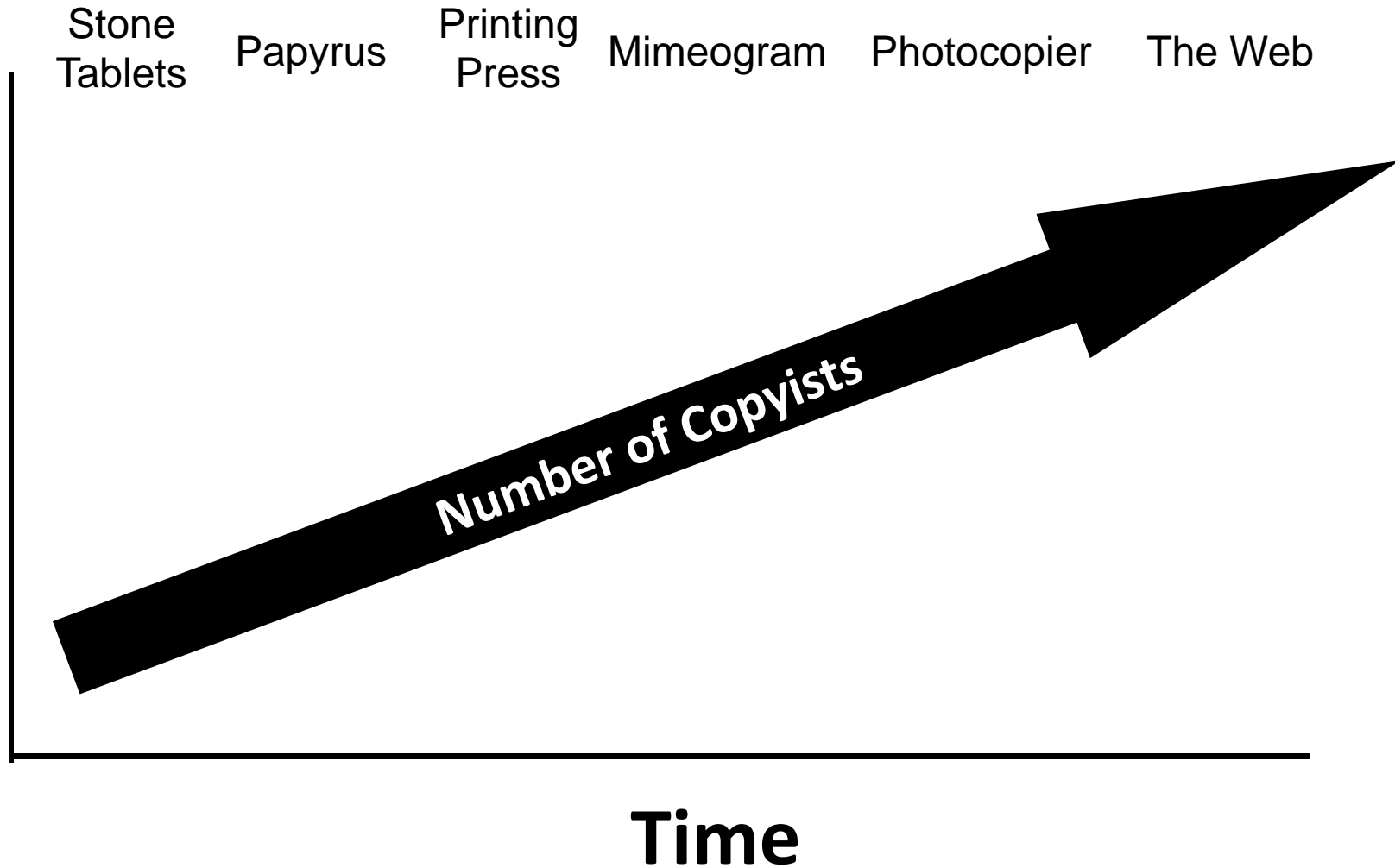
- As cost of copying declines, so does the value of content
- Profit from selling copies declines, taking down the “copyright industry”
- Piracy keeps prices low

# Copying Cost Goes Down

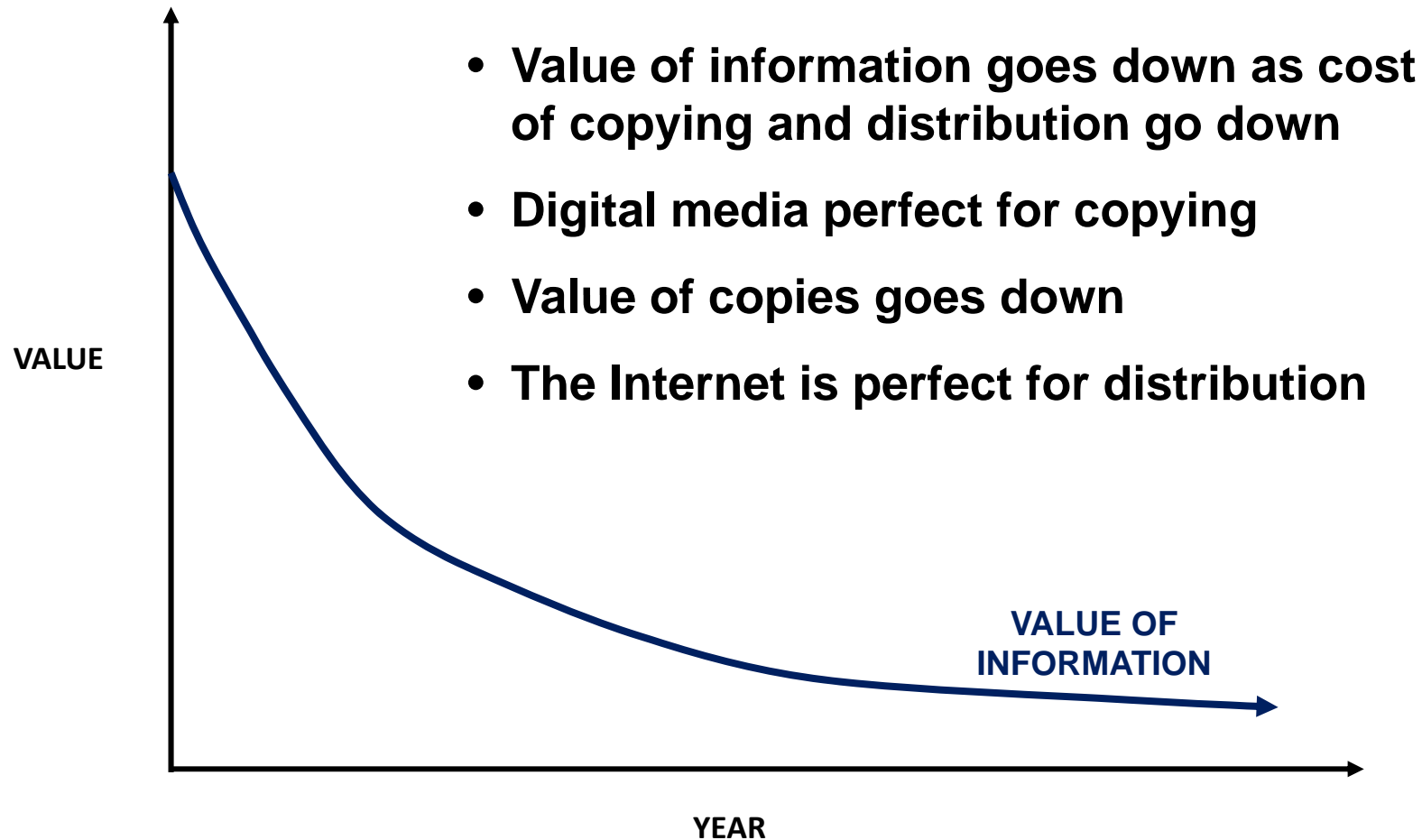


SOURCE: RAMEZ NAAM

# Everyone Can Copy



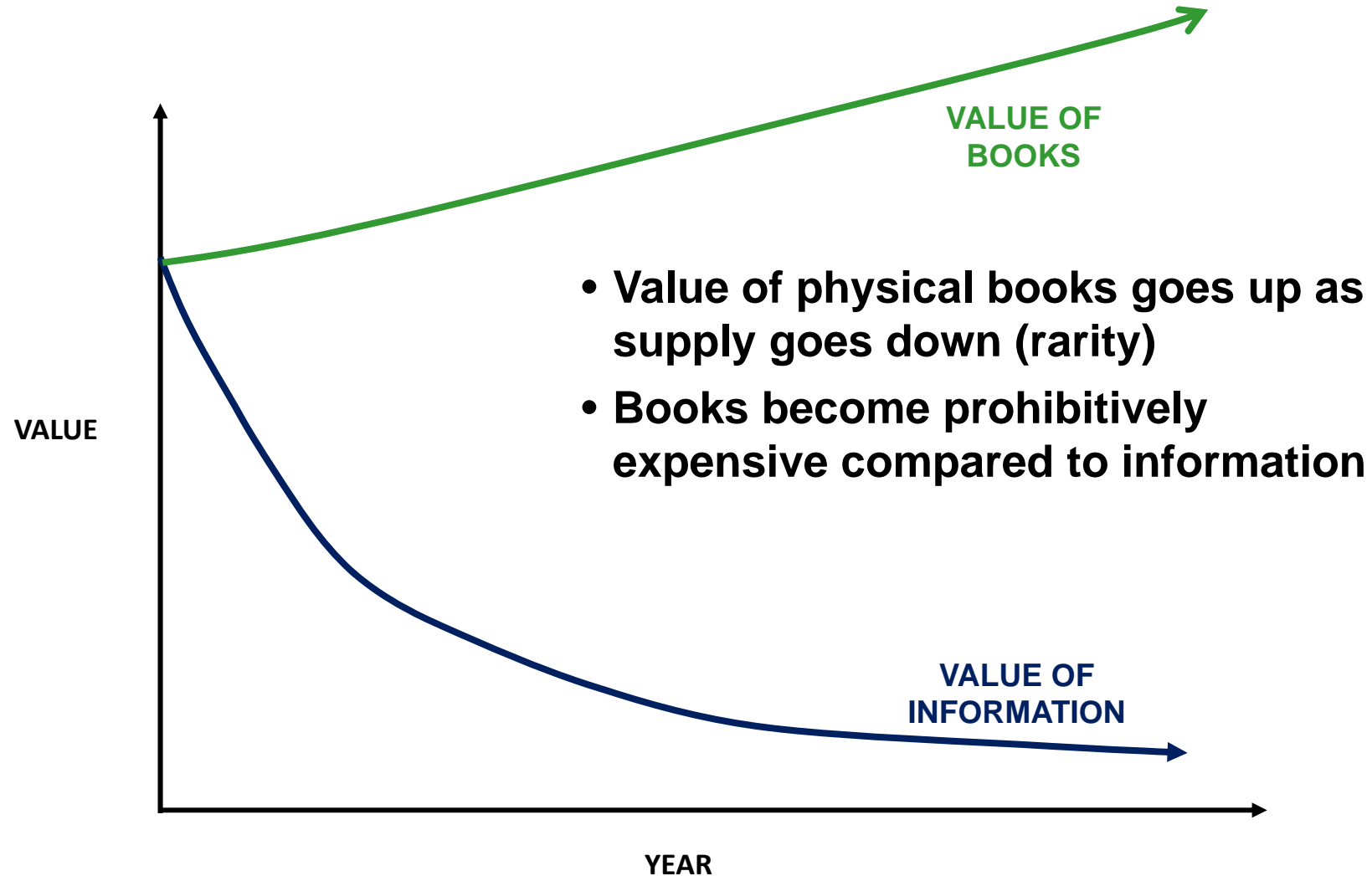
# Value of Information (Price that can be charged)





# Value of Books

(Price that can be charged)



# Shamos's Law of Information

- Free bits drive out costly ones (regardless of quality)
  - Like Gresham's law (1558): “bad money drives out good.”
- People will use free sources instead of pay sources, even if inferior
- People will use convenient sources over inconvenient ones, even if inferior
- Drives down the market value of information

# **The Book Autopsy: Determining the Cause of Death**

# In Olden Times ...

- Most information was ephemeral
- Created to be thrown away or impossible to preserve
  - Newspapers
  - Magazines
  - Live performances
  - Radio
  - Television

# In Modern Times ...

- Less and less information is from books
- Much information is digitized, stored, indexed and instantly available
  - Newspapers (Web)
  - Magazines (Web)
  - Live performances (YouTube)
  - Radio (Internet Radio)
  - Television (YouTube)

# Wikipedia

- Massive, distributed, volunteer effort to create a free encyclopedia
- Huge: 3,000,000 English entries; 10,000,000 overall
- More than 2 billion words, 1,000,000 contributors
- Would occupy >3000 printed volumes
- Wikipedia size: ~20GB, fits on a flash drive
- Encyclopedia Britannica has 120,000 entries; 5 million words (1-2% the size of Wikipedia)
- For-profit publishing of encyclopedias has ended
- Teachers mass against Wikipedia

# Causes of Book Death

- Vanishingly low cost of copying
- Unstoppable piracy
- Removal of publishers
- Changing social norms (personal piracy acceptable)
- Broadcasting vs. Publishing
- Mass collaboration
  - Wikipedia
  - Photobucket
  - YouTube



# When Books Die

- Libraries become mausoleums
- Books kept as historical artifacts, memories of an old civilization
- What is the useful lifetime of libraries?
- With fewer accesses, budgets will go down
- Libraries will be unable to survive



# Publishing v. Weblishing

- In publishing (books, CDs), copying is an expense
- Incomplete indexing, no delivery on demand
- In broadcasting (radio, TV), copying is not an expense
- Incomplete indexing, no delivery on demand
- No text content
- In Weblishing, copying is no expense
- Full indexing, worldwide delivery on demand
- Text content
- Everyone can be a weblisher

# The Weblish Model

- There has never been a medium like the Internet
- Publishing is miniscule by comparison
- Radio, TV vanish in real-time
  - Highly limited content, no indexing



# Scientific Publishing

- Internet posting is replacing bound journals
- Fast publishing
- Better indexing
- Universal access
- Free
- Reflects the fact that journal publishers do not add value. (Referees do, but they work for free.)

# Bio

- Professor at Carnegie Mellon University, Pittsburgh, Pennsylvania
- Director, Universal Library
- Visiting Professor, University of HK
- Collector of billiard books, top 3 in world
- Formerly, software executive
- Intellectual Property attorney
- Expert witness in computer cases

# Photo Sharing

-  facebook
  - 10 billion photos
  - 15 billion served per day, 3TB of new photos per day
-  photobucket
  - 6.5 billion photos
-  flickr
  - 2 billion photos



