Politecnico di Torino, 29 June 2019

Copyright, AI and concentration of power

Maurizio Borghi

Professor of Law

Centre for Intellectual Property Policy & Management

Bournemouth University

Summary

- Why copyright matters (for AI)
- How copyright can foster monopolies *viz*. concentration of power (in the AI world)
- How can we save the (AI) “world”

Some basic concepts

- Copyright (and relate rights)
- Public domain
- Fair use
- Exceptions and limitations

Some more basics

- “Open access”
- Open licenses: GPL, Creative Commons

The problem with copyright and AI

- Copyright is an “automatic” right
- Tech companies require repeated, systematic access to large amounts of data to train AI systems
- Good training data is crucial for creating accurate AI systems
- Some of the most valuable data are human-created works
- “Human” reading (viewing, listening) *vs* AI “reading”

Examples

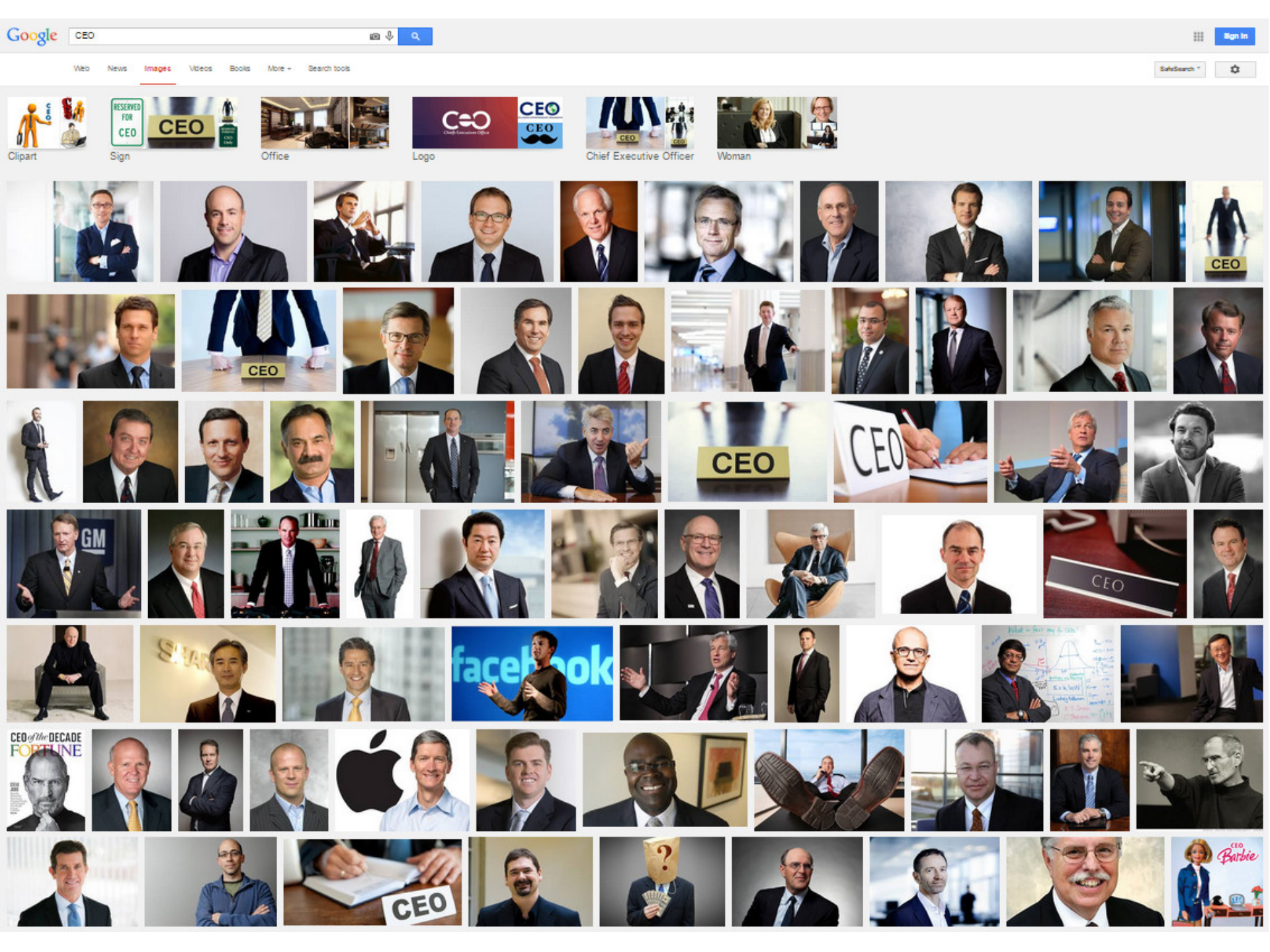
- Natural language processing – word embedding (word2vec)
- Machine translation
- Plagiarism detection
- Face recognition, voice recognition, etc…

How to feed the machine (with human-created works)

- Either you find a way to elude copyright, or…
- There are essentially two ways to secure access to a large corpus of © works: “buy-it” or “build-it”
- Buy-it approach: *e.g.* IBM’s Oncology Partnership, Google News, Google’s Patent Translate, ….
- Build-it approach: Facebook (and all social media)

Buy-it or build-it, but in both ways…

- Poorly accountable (and possibly biased) AI systems
- The “black-box” phenomenon
- “Propertization” of datasets and concentration of power
- Not only tech giants: the case of iParadigms


Google
CEO
Sign in
Web
News
Images
Videos
Books
More
Search tools
Clipart
Sign
Office
Logo
Chief Executive Officer
Woman
CEO
GM
facebook
CEO of the DECADE
FORTUNE
Barbie

The third way

- Permitted & permissible uses
- Examples: use of “open access” repositories, CC, Wikipedia,…
- Public bodies are obliged (by law) or encouraged (by policy) to make data/works created by them available for AI research
- The problem with the third way

The problem with the third way

- Risk-limiting solutions are always sub-optimal
- Quality of the data and possible bias due to use of “biased, low-friction data” (BLFD)
- Examples: Wikipedia, Open access journals, works in public domain – not always the best datasets to train AI systems

Where are we now?

- Large / dominant players owning both algorithms' code and datasets to feed them
- Entrants in the AI market left with two options: either use BLFD or become "vassal" of tech giants
- But (you may say) copyright law has the solution: fair use!

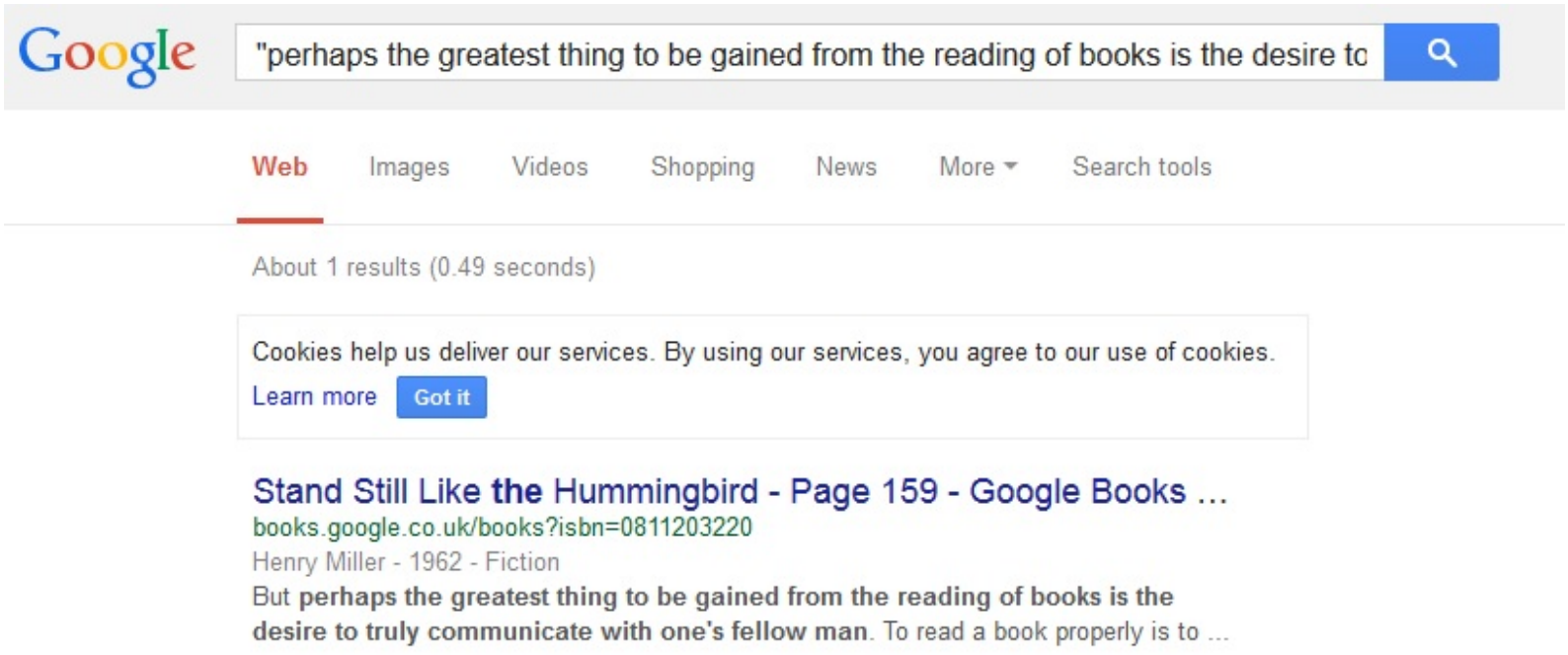
The fair use solution (and why it doesn't work)

- Not as "democratic" as it may seem: compare Google and Tixdaq
- But even if it were 100% predictable…
- The problem with "fair use exclusivity", *i.e.* when the first fair-user takes the whole benefit of the permitted use and precludes second-comers from availing of the same benefit.
- Examples?

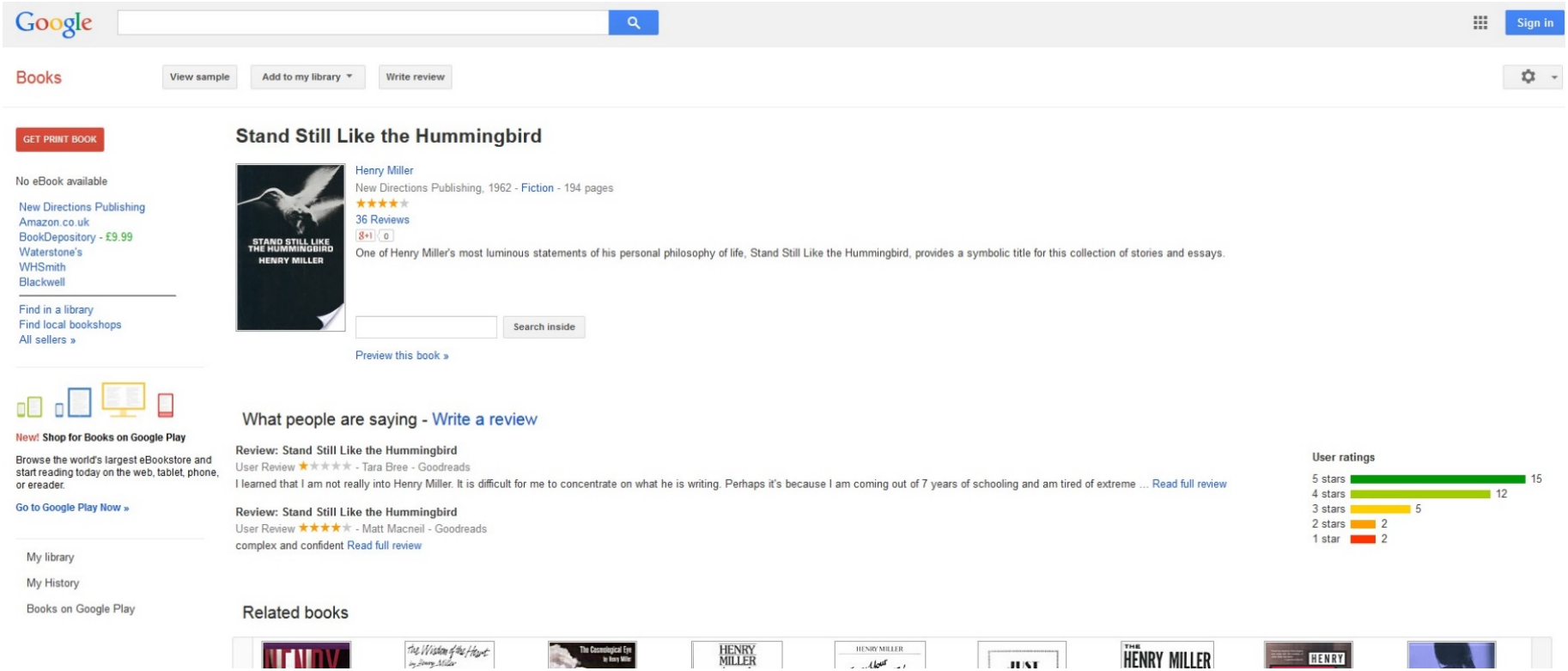
Google Books

- Started in 2004 as “Google Print” (then Google Book Search)
- Scanning and making available for search “all world’s books” (now featuring 30M + books in 45 languages), in partnership with libraries
- Sued in 2005 by the Authors Guild; $125M settlement agreement reached in 2009, but rejected by the court
- Finally decided as fair use (*AG v Google; AG v HathiTrust*)

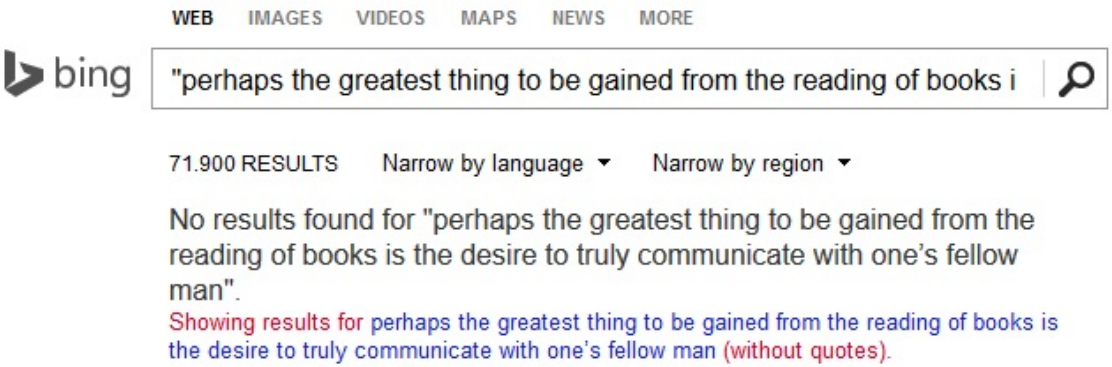
Google Books

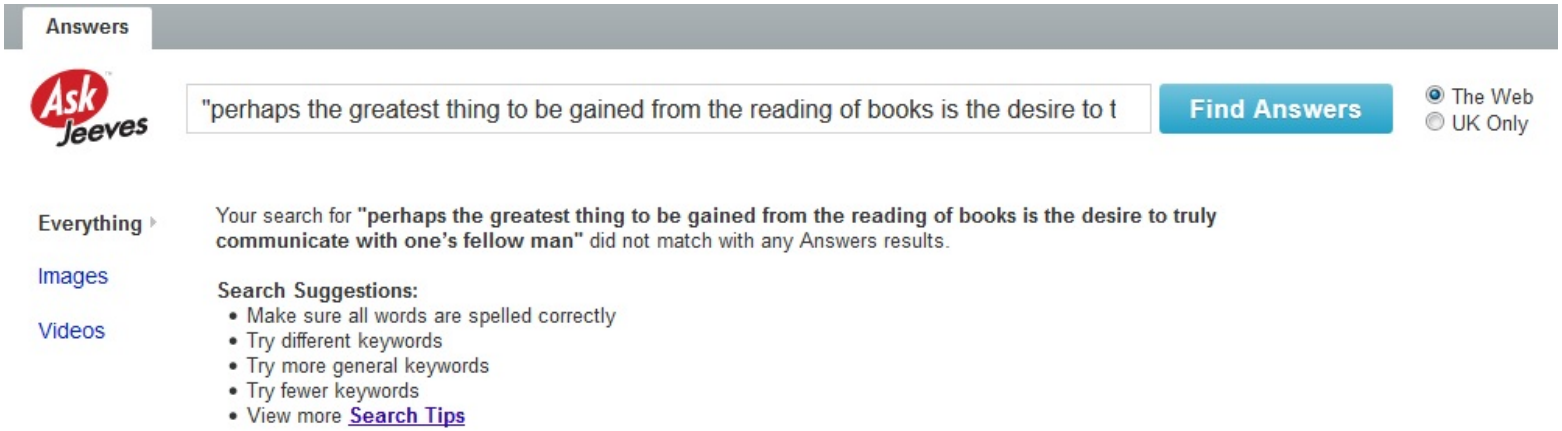


Google Books

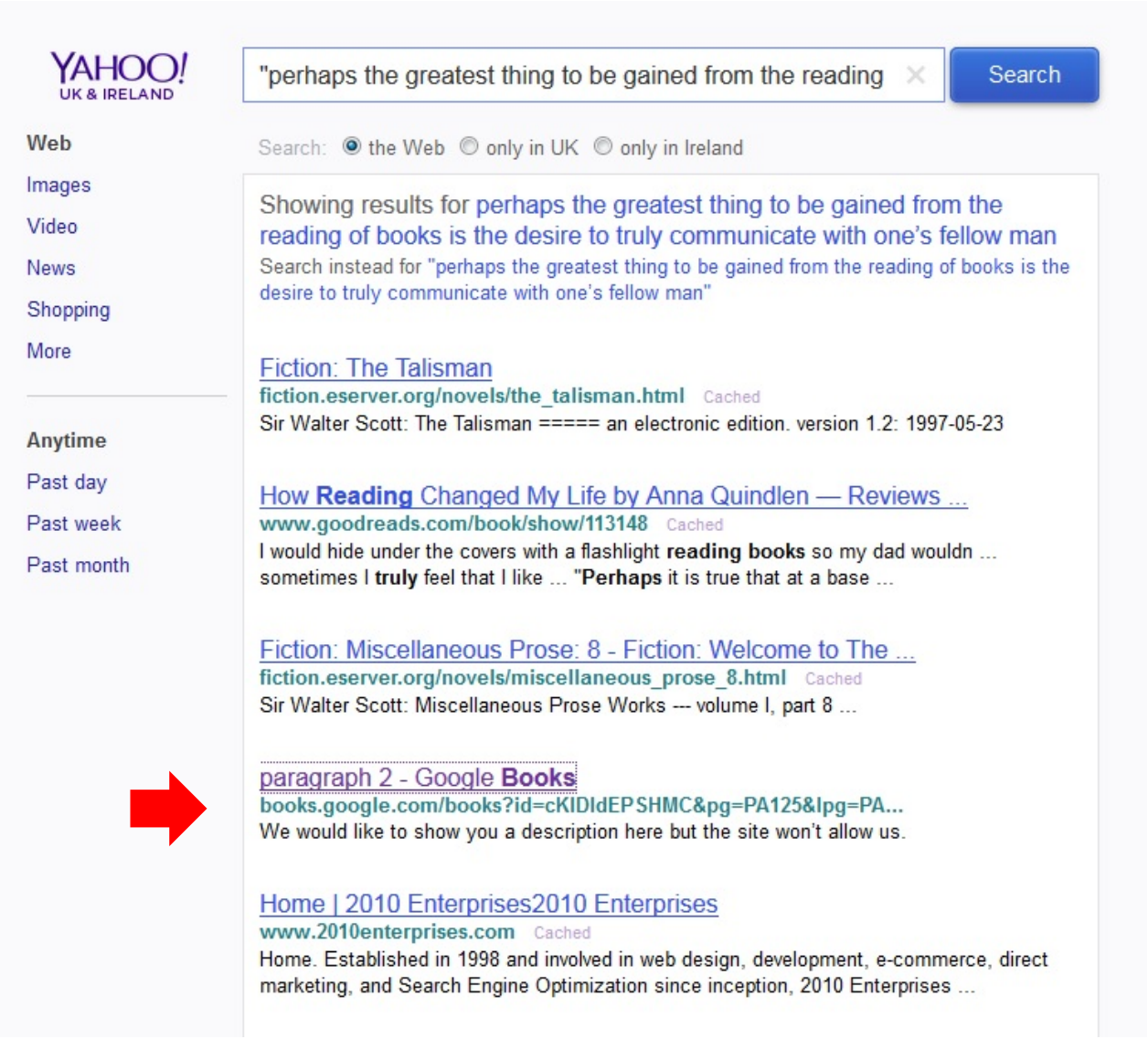


Google Books





Google Books



Turnitin

- Students' coursework submission → *wholesale copying and storage of copyright-protected content for plagiarism detection* (fair use in *AV v iParadigm*, 2009)

- Text mining from web pages (45+ billion), academic publications (130+ million) and papers previously submitted by students for marking (1 billion!)

- AI optimization + development of related services: online marking, "authorship investigation", automated marking (?)

AI monopolies

- Not just monopolies on the exploitation of digital objects (*e.g.* digital works), but monopolies over (certain) uses made possible by AI
- Mainly operating in two-sided markets
- Mainly non-replicable uses
- Similar (but not akin) natural monopolies

Unlocking AI monopolies

- Antitrust measures: doctrine of essential facilities (aka 'refusal to license' – ECJ *Magill*, C-241—242/91P) – *too slow or unavailable to remedy IP abuses*

- D.I.Y. measures – *e.g.* Internet Archive hosting ≈ 1M public domain books downloaded from Google Books and uploaded on their servers by users

- Use copyright as a basis to create "AI Commons". But how?

Towards AI Commons

- Legal basis: DSM Copyright Directive, art. 4: *mandatory exception for text and data mining*

Art. 4 DSM Copyright Directive

1. Member States shall provide for an exception or limitation to the rights [*of reproduction of works and extraction of database*] for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.
2. […]
3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine readable means in the case of content made publicly available online.