

## **EPO Academic Research Programme**

### **Final Report “Insights from Product-Patent Correspondence”**

Author: Gaétan de Rassenfosse

#### **Introduction**

The stated objectives of the project were twofold. First, it proposed to rely on data science tools to build a database that tracks innovations into the marketplace. Specifically, it planned to develop software capable of harvesting data on products from the web, which can be directly traced to identifiable EPO patents. Second, it proposed to exploit the data to advance knowledge on two topics: patent valuation and the patent prosecution process.

The EPO gave us the following feedback at the start of the project:

“We would like to draw your attention on the high ambition of your project within the available timeframe. Against this background, the main expected contribution should be the demonstration of the feasibility of your novel methodology, and we therefore invite you to pay special attention on the fine tuning and calibration of the web search results.”

Consequently, we have invested more energy into developing the software than initially planned. Nevertheless, we have managed to produce all the deliverables mentioned in the original submission (to various degrees of completion).

The following lists the deliverables (as stated in the grant application) as well as the status at the time of the grant termination (indicated by '>>').

- 1. Raw database of product-patent correspondence for EP patents.** We are committed to sharing the data to maximise dissemination. However, we also want to secure primacy. To balance these competing interests, we will release the data once we have been able to publish a first, citeable reference paper and have other papers at fairly advanced stage. We plan to release the data one year after the completion of the project.

>> The data will be made available on the project website, available at [www.iproduct.io](http://www.iproduct.io). Interested parties can subscribe to the mailing list in order to be notified when the data will be released.

- 2. Descriptive paper**, released on SSRN 14 months after the start of the project.

>> A descriptive paper is available at <https://ssrn.com/abstract=3114637>. This paper analyses the factors that push firms to adopt virtual patent marking. In brief, we find that firms are more likely to mark their products if they have a higher likelihood of

being infringed, if they pursue an active branding strategy, and if they are in greater need of external financing.

3. **Infographics** from task 2.1, released on various communication channels (LinkedIn, Twitter, mailing list, etc.) 15 months after the start of the project (target date).

>> Infographics are provided in Appendix A.

4. **Empirical analysis of patent value** from task 2.2, released on SSRN 19 months after the start of the project (target date).

>> Our analysis on patent value has sought to associate to drop in product price around the time of patent expiry. Analysis suggests that patent expiry leads to an average drop in product price of about 2%. The drop in price is stronger in more competitive fields and for more important patents (as proxied with patent citations). We haven't released the working paper yet because we want to test further the robustness of our results. A working paper will be released on SSRN during the first half of 2020.

5. **Empirical analysis of patent pendency** from task 2.3, released on SSRN 24 months after the start of the project (target date).

>> We have collected data on the market release date of about 500 products from the database. We have performed a preliminary econometric analysis of the extent to which patent pendency affects market release date. Tentative estimates suggest that a 10-percent increase in office-induced patent pendency retards product market introduction by about 4 percent. The analysis is still in progress, and we expect to release a working paper in the second half of 2020.

The remainder of the report explains the software that we have developed to build the database.

### **Overview of software (data capture pipeline)**

The database is called *IPProduct*, formed by the contraction of the terms "Intellectual Property Rights" and "product".

#### Virtual patent marking webpages

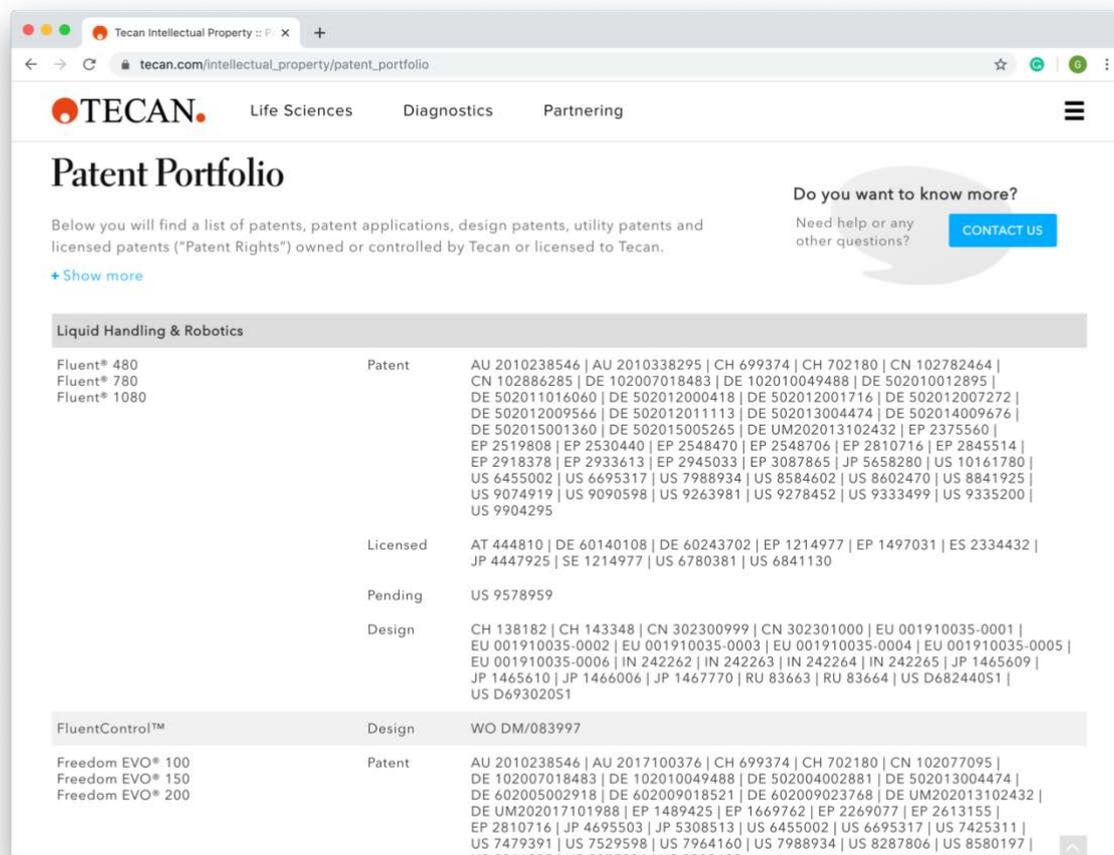
We have explained at length at the intermediary progress meeting that our starting material consists of virtual patent marking (VPM) webpages, which are self-reported product-patent lists available on company websites.

Figure 1 below provides an example of a VPM webpage. We have previously emphasized the variety of formats and supports that company rely on to deliver VPM information. This heterogeneity makes the data extraction task particularly challenging.

We have produced a paper that studies factors that affect VPM adoption by firms (<https://ssrn.com/abstract=3114637>). This work is important to understand potential selection bias in the data. In short, the results are consistent with the following mechanisms:

firms are more likely to use VPM if they have a higher likelihood of being infringed, if they pursue an active branding strategy, and if they are in greater need of external financing.

**Figure 1.** VPM webpage of Tecan Trading AG



We estimate that there must be between 10k and 50k VPM webpages. Finding these pages is a challenge, because there is no repository available and no unique keywords to identify them. We spend a great amount of energy trying to find the best ways of identifying VPM webpages. The common denominator between these pages is that they should all contain the triplet (patent(s) keyword, patent number(s), product names or code). However, many non-VPM pages also contain this triplet—such a naïve filter therefore produces a large amount of false positives.

First, we co-developed a large-scale web crawler in collaboration with SCITAS, the high-performance cluster at EPFL, and we carefully selected a list of seeds (*i.e.*, entry points into the web for the crawler). The seeds speed up crawling time and allow us to avoid collecting too many false positive. We crawled about 250m webpages, and about 2.5m of them contain the triplet.

Second, we trained a supervised VPM classifier, that is an algorithm that identifies VPM pages among our set of 2.5m candidates pages. We suspect that about 1% of candidates web pages are VPM webpage so we would need an extremely precise classifier. However, we did not push the development further as we had identified with certainty more than 1000 VPM

webpages thanks to a fairly simple classifier. This number was large enough to proceed to the next stage of the development.

### Data platform

A substantial amount of energy went into the development of a web platform for storing, structuring and managing the data.

We started with a simple PHP-based web interface that we used to manually extract data on product-patent correspondence. These manually curated data represent our ‘gold standard dataset’; they are crucial for the development and training of automatic extraction algorithms.

However, we quickly realized that manually cleaning the data was a daunting task and we developed a plugin for the Chrome web browser to assist in data annotation. The plugin turned out to be really useful and allowed to speed up manual data annotation and minimize encoding errors.

Next, we used the manually labelled data to identify to broad typologies of pages, with the aim of developing a family of fully automatic extraction algorithms. We came up with extraction algorithms but these algorithms often needed some manual input to function properly. In light of this, we have decided to integrate the extraction algorithms directly into the plugin. In other words, we have developed a tool that dramatically facilitates human labeling. We plan to continue in this direction in the future, by further refining the extraction algorithms and the automatic recognition of patent numbers. A video showing how the plugin works is available on YouTube at <https://youtu.be/bxiRyaMRFhc>.

Finally, we have decided to refactor the platform totally. We are now developing a Scala-based platform that directly embeds the plugin. The new platform will be more stable than the previous version and it will be also easier to open it to external users. Ideally, we would like to allow external users (either paid RAs or volunteering enthusiasts) to contribute to labeling, with specific access rights. Figures 2 and 3 provide a preview of the new version of the platform. The platform contains structured marking information as well as metadata about the company (e.g., LinkedIn URL). It will also be possible to observe different snapshot of the webpages and the marking information, allowing us to build a panel dataset of patent coverage in the near future. (At the moment, we are only able to store cross-sectional data).

The present report is an abridged, public version of the full report submitted to the EPO (95 pages)

**Figure 2.** Print screen of platform (under development) – directory listing

id	edited	last editor	# entries	company	comment
3021	2018-06-16	gderasse	1	kampmann.co.uk Kampmann GmbH	<a href="https://www.kampmann.co.uk/products/door-air-curtains/protector.html">https://www.kampmann.co.uk/products/door-air-curtains/protector.html</a>
1501	2017-12-08	Jkozak	46	appliedmedical.com Applied Medical	<a href="http://www.appliedmedical.com/Legal/Patents">http://www.appliedmedical.com/Legal/Patents</a>
4201	2018-08-05	gderasse	2	flatfrog.com FlatFrog Laboratories AB	<a href="https://www.flatfrog.com/patents">https://www.flatfrog.com/patents</a>
1411	2018-08-06	gderasse	3	kimberly-clark.com Kimberly-Clark	<a href="http://www.cms.kimberly-clark.com/umbracoimages/UmbracoFileMedia/GOOD">http://www.cms.kimberly-clark.com/umbracoimages/UmbracoFileMedia/GOOD</a>
629	2019-01-13	gderasse	1	doserite.info Surgin	Done, to be checked by Gaétan <a href="http://doserite.info/patents/">http://doserite.info/patents/</a>
5561	2019-01-09	gderasse	1	energetiq.com Energetiq - A Hamamatsu Company	<a href="https://www.energetiq.com/DataSheets/EQ105XR_DataSheet.pdf">https://www.energetiq.com/DataSheets/EQ105XR_DataSheet.pdf</a>
4101	2018-07-24	gderasse	23	emerson.com emerson.com	<a href="http://www.emerson.com/en-us/automation/brands/micro-motion/micro-motion">http://www.emerson.com/en-us/automation/brands/micro-motion/micro-motion</a>
2341	2018-06-22	gderasse	1	mysuspenders.com Holdup Suspender Company Inc.	<a href="http://www.mysuspenders.com/">http://www.mysuspenders.com/</a>
101	2018-06-08	gderasse	10	hoyasurgicaloptics.com HOYA Surgical Optics	<a href="http://hoyasurgicaloptics.com/global/about/patents/">http://hoyasurgicaloptics.com/global/about/patents/</a>
2031	2019-09-03	gderasse	42	stryker.com Stryker	product per patent -- To check aggregation of common entries... <a href="https://www.stryker.com/content/stryker/us/en/about/patents/endo-scopy-pater">https://www.stryker.com/content/stryker/us/en/about/patents/endo-scopy-pater</a>
479	2017-11-13	gderasse	1	digitalfactory3d.com Kraftwurx	<a href="http://www.digitalfactory3d.com/patent-notification">http://www.digitalfactory3d.com/patent-notification</a>
6851	2019-01-12	gderasse	25	epibio.com Epicentre, an illumina company	<a href="http://www.epibio.com/company/legal-information">http://www.epibio.com/company/legal-information</a>
333	2018-06-08	gderasse	1	corplug.com CorPlug, Inc.	<a href="http://corplug.com/?page_id=300">http://corplug.com/?page_id=300</a>
628	2019-11-19	none	30	tecan.com Tecan	<a href="http://www.tecan.com/intellectual_property/patent_portfolio">http://www.tecan.com/intellectual_property/patent_portfolio</a>
1031	2018-08-21	gderasse	33	callawaygolf.com Callaway Golf	<a href="http://www.callawaygolf.com/on/demandware.store/Sites-CG2-Site/default/Cus">http://www.callawaygolf.com/on/demandware.store/Sites-CG2-Site/default/Cus</a>
6431	2019-01-12	gderasse	1	cincinnatichildrens.org cincinnatichildrens.org	<a href="https://www.cincinnatichildrens.org/-/media/cincinnati%20childrens/home/ress%20molecular%20targeted%20combination%20therapy%20for%20autoimmun">https://www.cincinnatichildrens.org/-/media/cincinnati%20childrens/home/ress%20molecular%20targeted%20combination%20therapy%20for%20autoimmun</a>
	2017-10-25	Jkozak	26	monsantotechnology.com	<a href="http://www.monsantotechnology.com/Vegetables/Vegetables-Bean.aspx">http://www.monsantotechnology.com/Vegetables/Vegetables-Bean.aspx</a>

**Figure 3.** Print screen of platform (under development) – annotated webpage

company

id #1051 SELECT NEW

domain appliedmedical.com

name Applied Medical

sector Medical Devices

linkedin https://www.linkedin.com/company/a;

facebook

page

url http://www.appliedmedical.com/Legal/Patents

page type simple

updated

footer VPM

page belongs to company

comment

Products

US Patent Number(s)

Alexis Laparoscopic System 7,650,887; 7,815,567; 7,883,461; 7,892,172; 7,909,760

Alexis Laparoscopic System with Kai Fios First Entry 6,887,194; 7,083,626; 7,112,185; 7,650,887; 7,686,823; 7,771,395; 7,794,644; 7,815,567; 7,883,461; 7,892,172; 7,909,760; 7,947,058; 8,235,054; 8,430,851; 8,506,520; 8,608,768; 8,613,727; 8,940,009; 9,155,558; 9,314,266; 9,358,040

Alexis O C-Section Retractor 7,650,887; 7,883,461; 7,892,172; 7,909,760; 8,235,054; 9,101,354

Alexis O Wound Protector/Retractor 7,650,887; 7,883,461; 7,892,172; 7,909,760; 8,235,054; 9,101,354; 9,561,024

snapshots

2017-11-14 snapshot #4991

### Jupyter notebook for data enrichment

The final step in the data processing pipeline relates to data enrichment. In particular, we have linked the patent numbers available on the website to the following databases: PatentsView, PATSTAT and lens.org. Data enrichment takes place in a Jupyter notebook using Python. Figure 4 provides an overview of the notebook. We are only processing U.S. and EP patent so far, although we plan to extend the labeling to other jurisdictions.

One issue we need to deal with concerns the variety of format in which patents are reported. Table 1 illustrates four cases related to the EPO. Firms use a variety of markers to identify European patent documents, such as 'Europe', 'EP', 'EPO', 'EPO', etc. These labels are sometimes ambiguous, such as 'Europe' indicating both the EPO and the EUIPO, or used wrongly (such as EP to designate EUIPO). Firms also use different patent numbers, such as application numbers (with or without the final digit) or publication numbers.

The data are harmonized in three stages. First, we have developed a fairly sophisticated pattern recognition algorithm that accepts a variety of formats under which patent numbers can be reported. Second, a human supervisor corrects potential labeling mistakes from the previous step. The objective is to encode patent numbers into a standardized format, as shown in Table 2. Third, the matching of the encoded string to external patent databases allows us to verify that the patent document actually exists. If we cannot find a match, we perform a manual search to correct the encoding. Typical errors at this stage are rare and include wrongly reported jurisdictions or patent numbers on the company webpage. We manually correct such cases when we identify them.

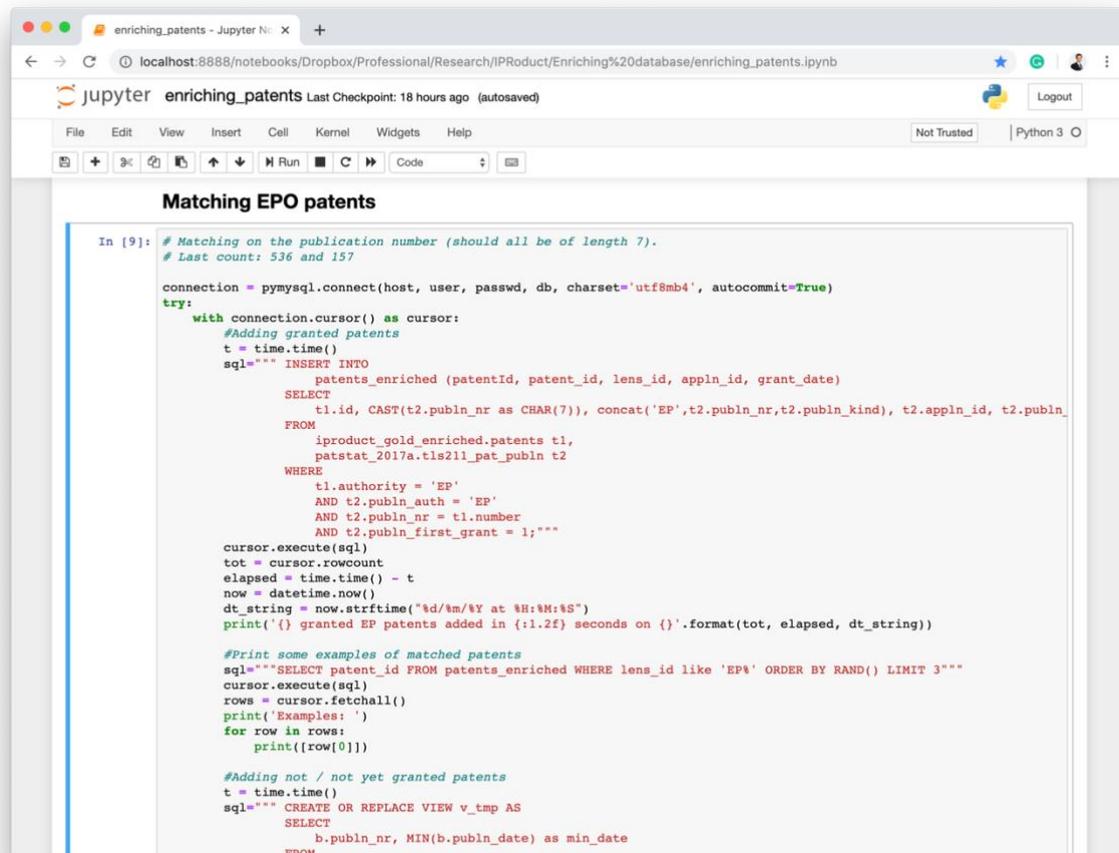
**Table 1.** Illustrations of some cases encountered for European patent documents

<b>Webpage</b>	<b>Encoded</b>
Europe 03756746.8	A---- EP 03756746.8
EPO 2768613A1	A---- EP 2768613 A1
EP506870	P---- EP 0506870
Europe 000116215-0001	P--D- EU 000116215-0001

The bottom line of the project is this:

1. We are confident that there are enough VPM webpages to build a database that would be useful to the scientific community;
2. We have prototyped the whole data collection and extraction pipeline;
3. We have a sound understanding of the computational and algorithmic needs to scale up the project; and
4. We have assembled a manually curated database that can be used to train ML algorithms to refine data collection and extraction.

Figure 4. Data enrichment notebook



```
In [9]: # Matching on the publication number (should all be of length 7).
# Last count: 536 and 157

connection = pymysql.connect(host, user, passwd, db, charset='utf8mb4', autocommit=True)
try:
    with connection.cursor() as cursor:
        #Adding granted patents
        t = time.time()
        sql=""" INSERT INTO
        patents_enriched (patentId, patent_id, lens_id, appln_id, grant_date)
        SELECT
        t1.id, CAST(t2.publn_nr as CHAR(7)), concat('EP',t2.publn_nr,t2.publn_kind), t2.appln_id, t2.publn_
        FROM
        iproduct_gold_enriched.patents t1,
        patstat_2017a.tls211_pat_publn t2
        WHERE
        t1.authority = 'EP'
        AND t2.publn_auth = 'EP'
        AND t2.publn_nr = t1.number
        AND t2.publn_first_grant = 1;"""
        cursor.execute(sql)
        tot = cursor.rowcount
        elapsed = time.time() - t
        now = datetime.now()
        dt_string = now.strftime("%d/%m/%Y at %H:%M:%S")
        print('{} granted EP patents added in {:.2f} seconds on {}'.format(tot, elapsed, dt_string))

        #Print some examples of matched patents
        sql="""SELECT patent_id FROM patents_enriched WHERE lens_id like 'EP%' ORDER BY RAND() LIMIT 3"""
        cursor.execute(sql)
        rows = cursor.fetchall()
        print('Examples: ')
        for row in rows:
            print([row[0]])

        #Adding not / not yet granted patents
        t = time.time()
        sql=""" CREATE OR REPLACE VIEW v_tmp AS
        SELECT
        b.publn_nr, MIN(b.publn_date) as min_date
        FROM
```

The present report is an abridged, public version of the full report submitted to the EPO (95 pages)

## **APPENDIX A**

**This appendix provides a series of illustrations of the *IP*Product data**

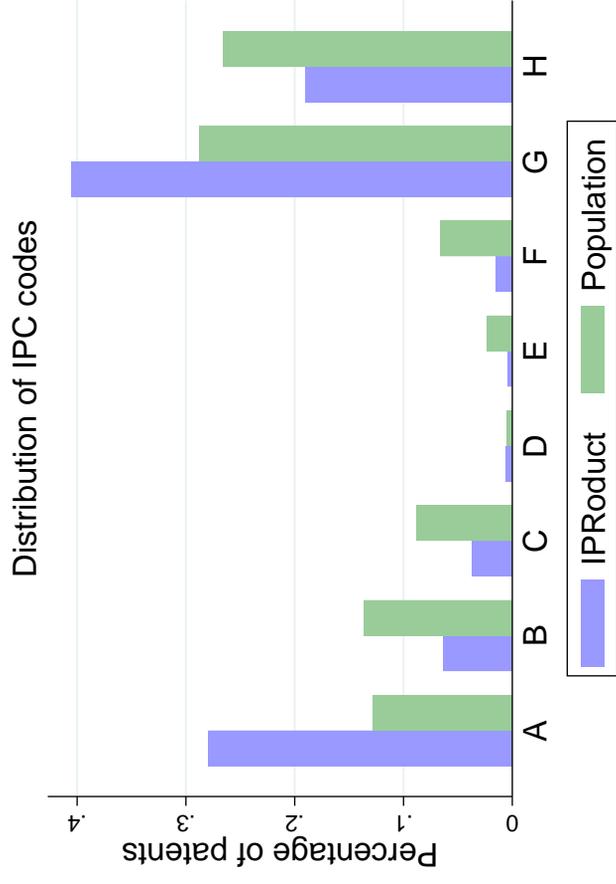
# *IP*Product: Infographics

Gaétan de Rassenfosse  
Ecole polytechnique fédérale de Lausanne



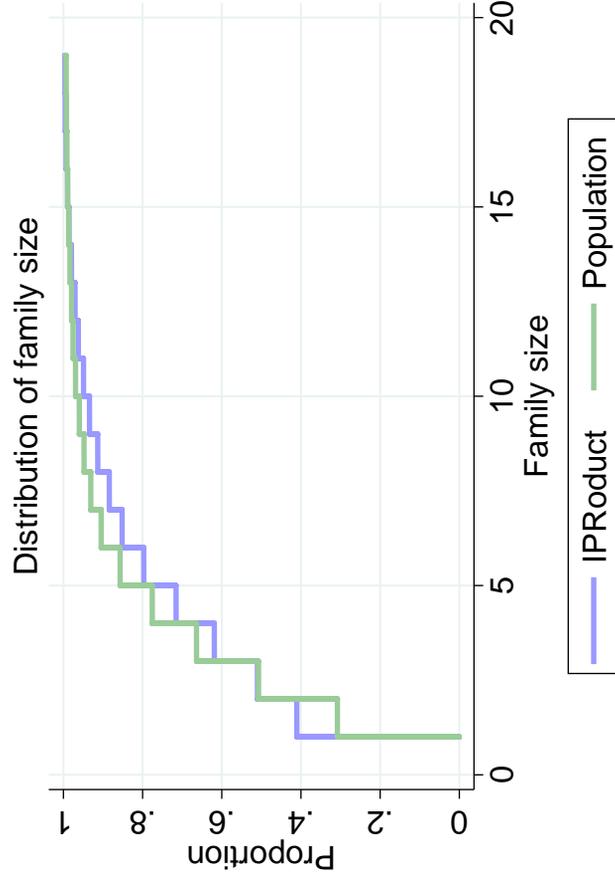
@gderasse

# Comparing with population of USPTO patents



- Patents in *IPRoduct* are spread across technologies, in a manner similar to the population of patents

# Comparing with population of USPTO patents



- Patents in *IPRoduct* are more likely to be singleton families.
- Conditional of not being a singleton, however, *IPRoduct* patents belong to larger patent families.

EPFL



SIEMENS



Alcon  
EPSON®



NEWTON  
running

BlackBerry



COREL™

Baxter



crocs™

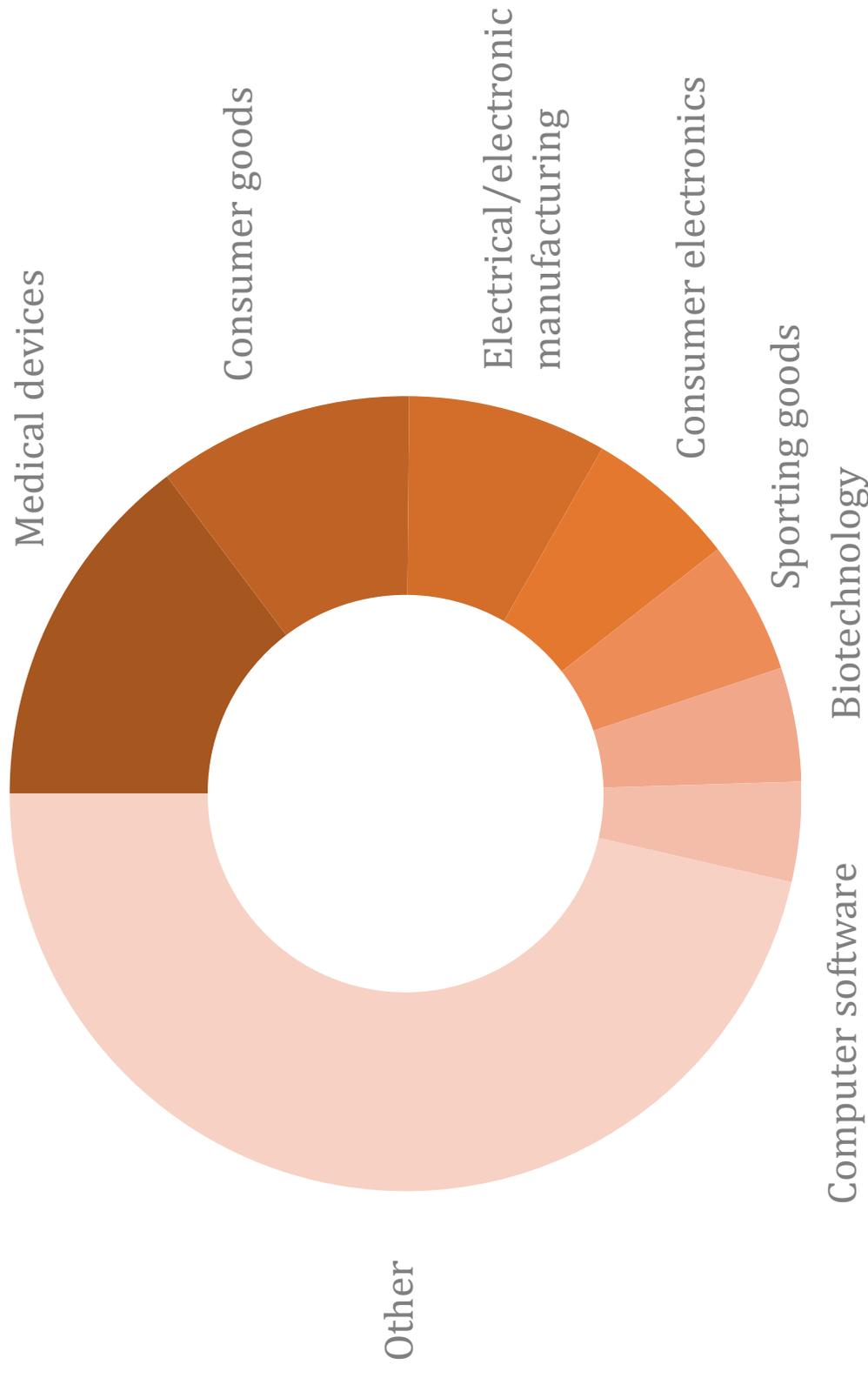


BOSCH

Panasonic

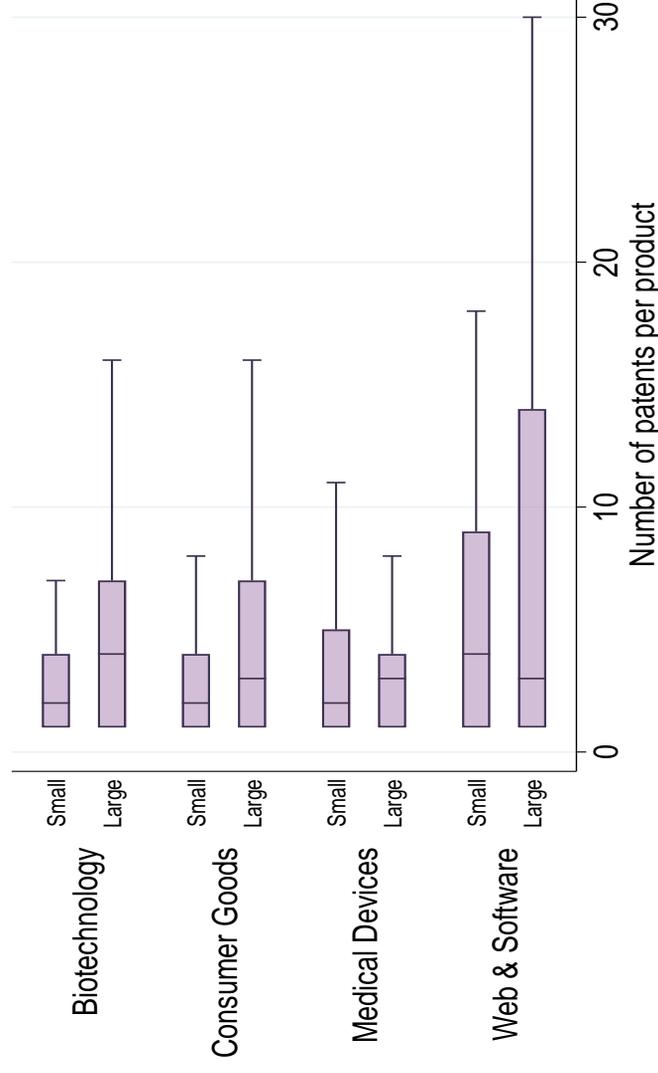
*About 25% of USPTO assignees with commercial products  
have a virtual patent marking webpage*

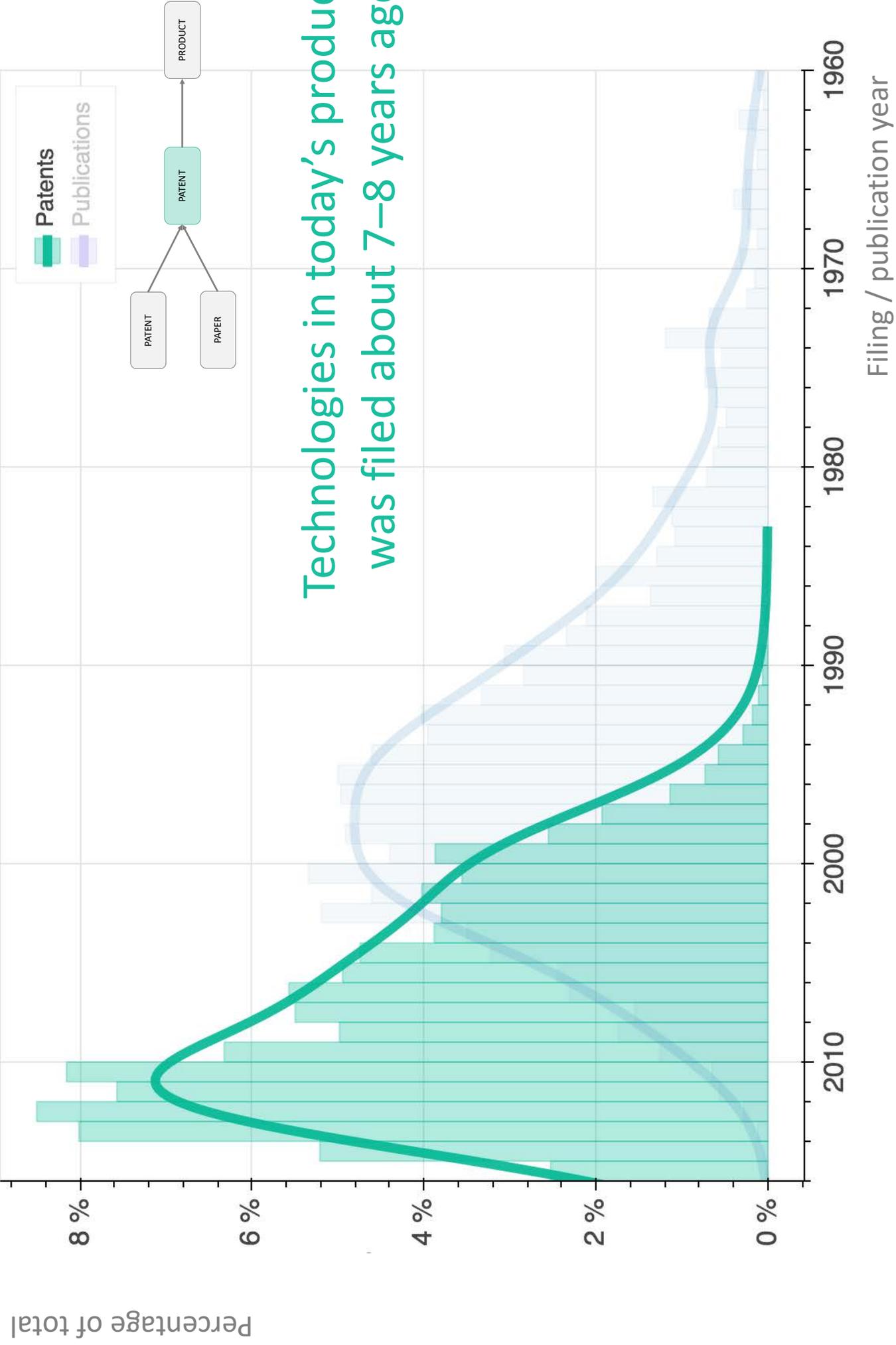
# 'Consumer goods' broadly defined form the largest group of marked items



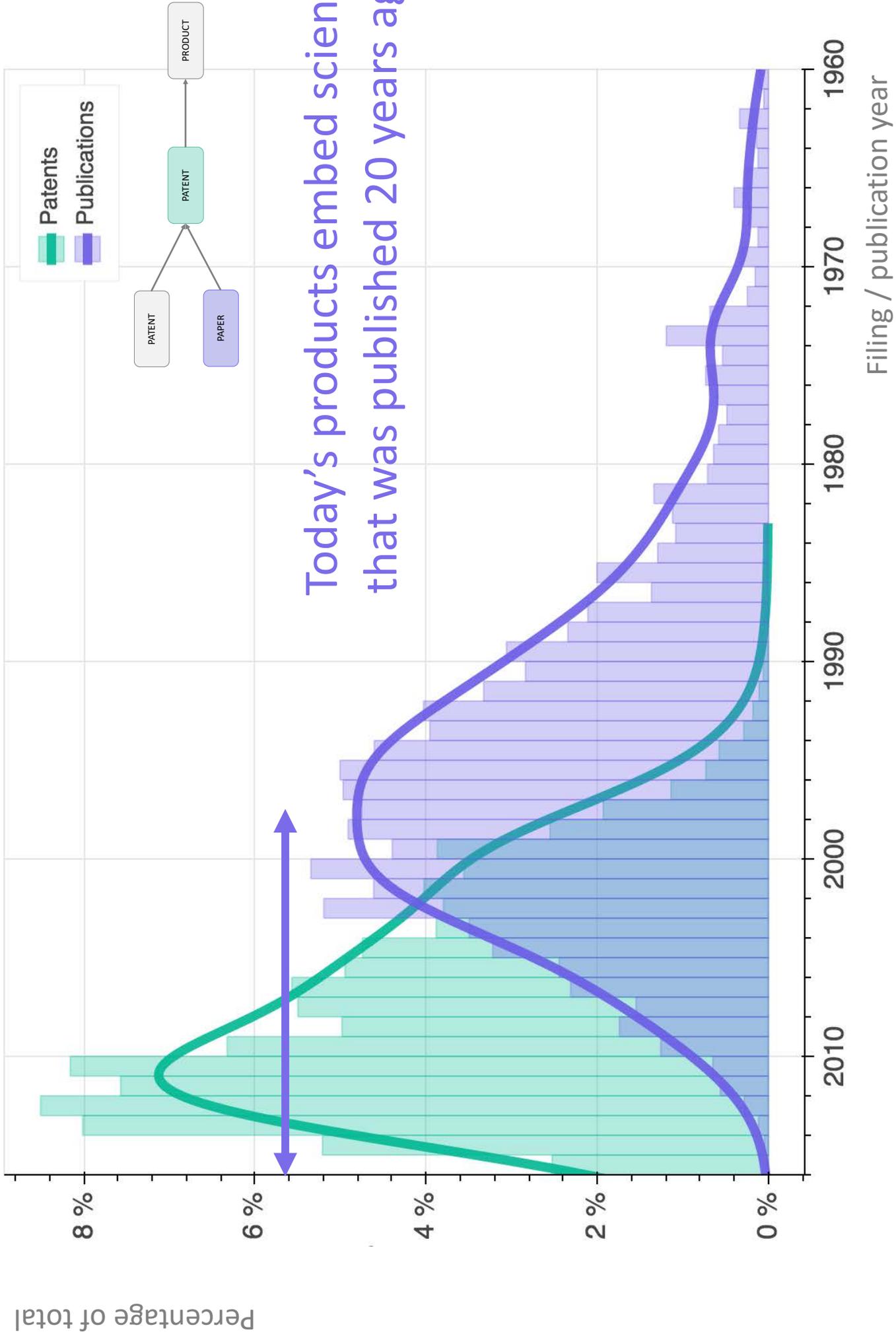
# The myth of single-patent products

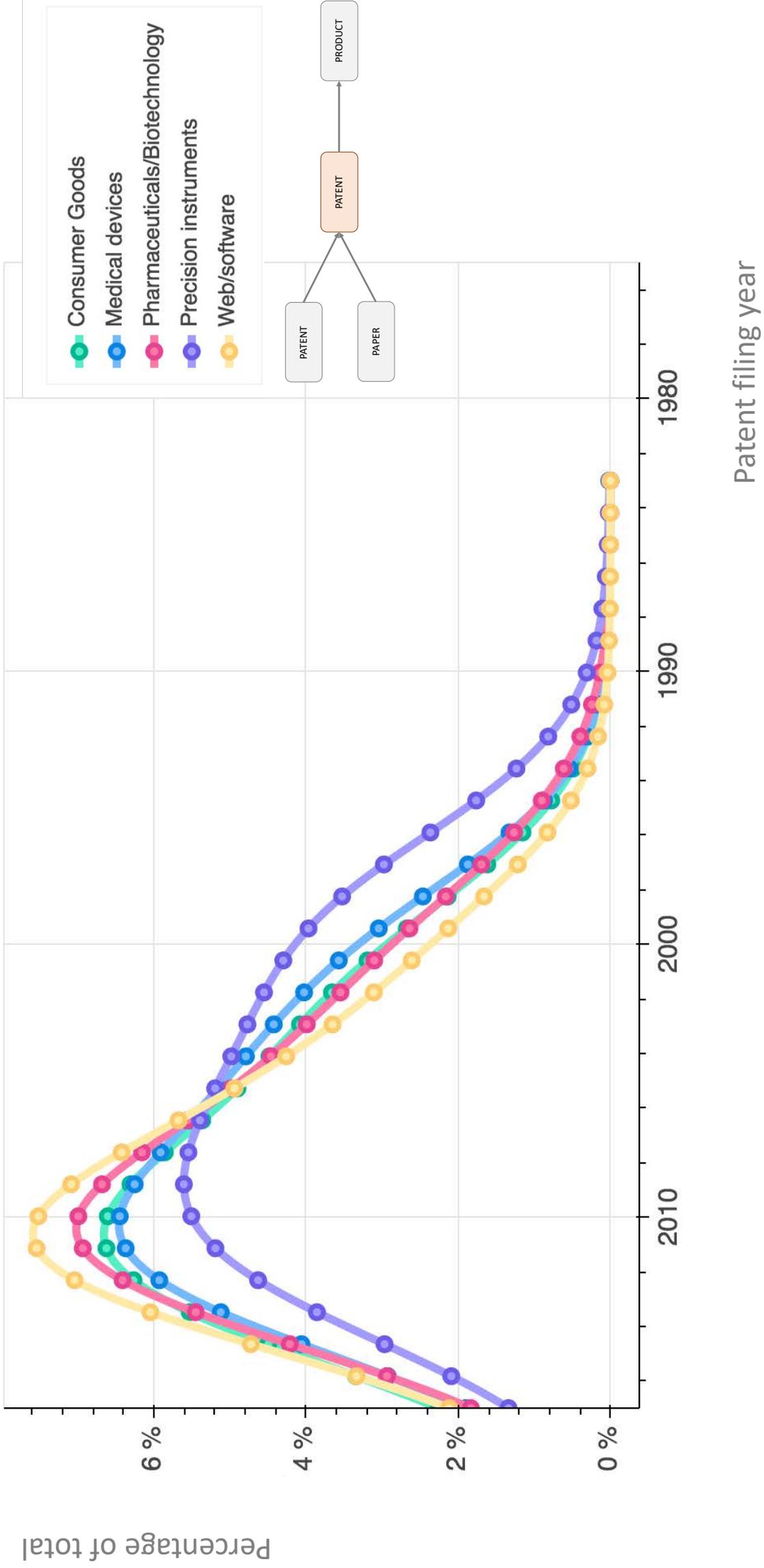
- The median number of patents per product is **three**, and the variable is highly skewed—10 percent of products are protected by 10 patents or more.
- In a typically discrete industry such as Medical Devices, about 7 percent of products are protected by 10 patents or more.
- This figure reaches 21 percent for consumer electronics (which forms a subset of Consumer Goods).





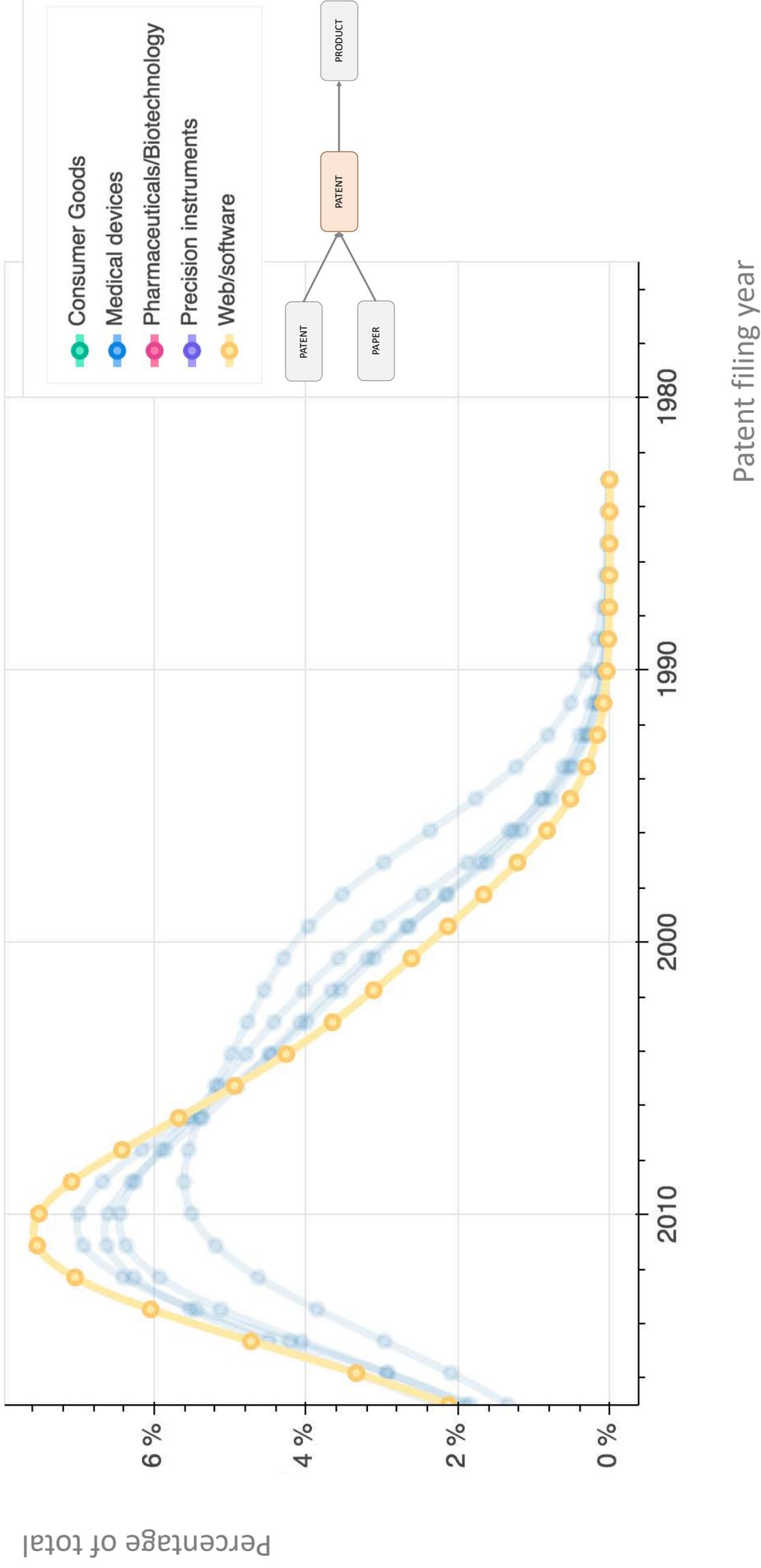
Technologies in today's products  
was filed about 7-8 years ago





Patent filing year

The pace of (patented) innovation is not (much) faster in web/software



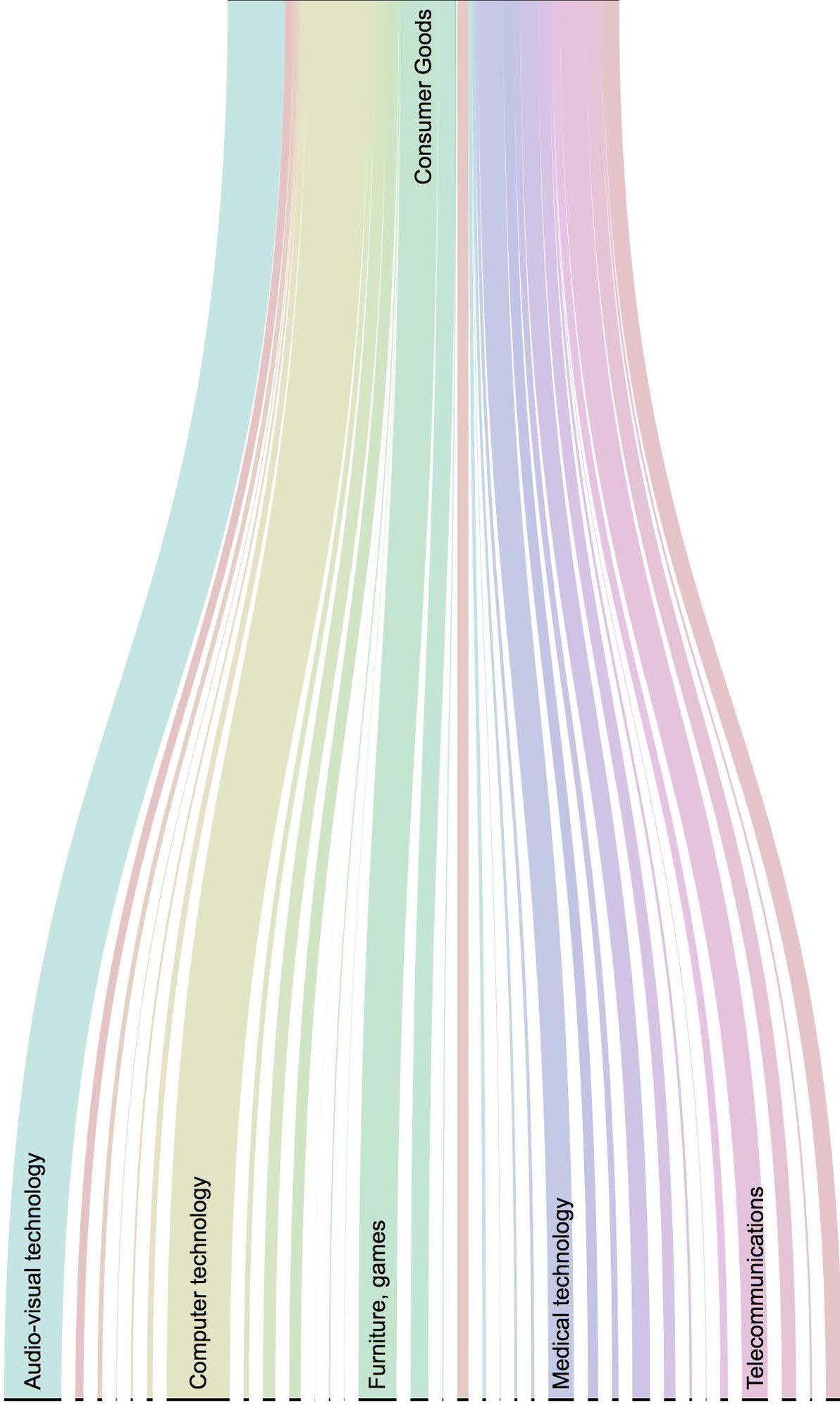
# Patent Technological Field

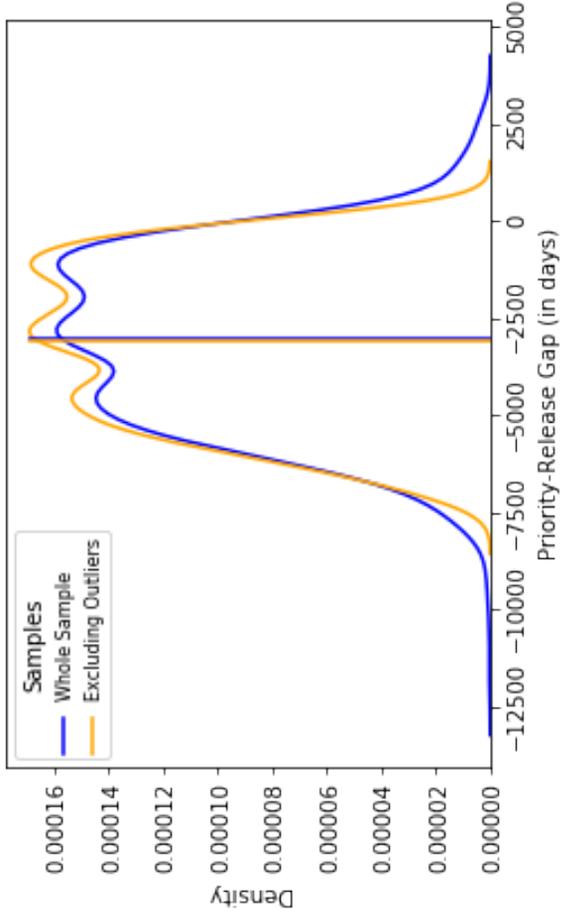
# Product Type



Patent Technological Field

Product Type





On average, patents are filed  
more than 7 years before  
product introduction

On average, products will be  
protected by at least one  
patent for 15 years

