All the data and code necessary to reproduce the results are provided or described in this repository. Datafiles and code are **boldfaced**.

Figures from the article are reproducible from the year-level summary data in **us.govt.reliance.statistics.tsv**, generated from the patent-level data in **uspto.govt.reliance.metadata.tsv** using the script **command.generate.mastertable.sql**.

- The metadata table for years 1926-1975 was constructed from OCRed versions of the patent documents prior to 1976.[1] (These are uspto.ocr.1925-1976.zip along with **uspto.ocr.1925-1975.z0{1-5}**, all of which need to be put under one folder upon unzipping.)
- The 1976-2017 patent data can be obtained most easily from Patentsview,[2] more specifically the "patent" "rawassignee" and "rawlocation" tables.

A list of just the patents that are reliant on government-supported science is in **uspto.govt.reliance.category.tab**. For each patent, it lists the grant year and whether the patent is owned by the government, directly acknowledges government support, cites a patent owned by the government, cites a patent that acknowledges government support, cites a Web of Science paper authored by the government, or cites a Web of Science paper that acknowledges government support. This file is generated from **uspto.govt.reliance.metadata.tsv** using the script **build.patent.govt.reliance.category.sql** as well as the following data sources:

- Patent assignees are binned into government, university, or corporate using **assignee.type.binning.php**, using lists of governmental and educational institutions described in the Supplemental Materials.
- Code for identifying that a given patent directly acknowledges government support is **patent.acknowledging.government.funding.py.**
- Citations from patents to other patents are in two chunks: **uspto.citation.1976-2017.zip** and **uspto.citation.1926-1976.tab**.
- Citations from patents to Web of Science papers are not publicly available because the Web of Science data is property of Clarivate. The sales team at Clarivate can be contacted for licensing at https://clarivate.com/products/web-of-science/contact-us. Code for resolving Non-Patent References (also contained in **uspto.citation.1976-2017.zip**) to Web of Science papers is in two chunks.
    - Title-based matches are in **npl.wos.matching.sql** and assume that the Web of Science XML files have been ingested into an SQL database, with matching table and variable names.
    - Non-title-based matches are in a series of files executed as follows:
        1. Run **wosparsepubinfo.sh** (which calls **wosparsepubinfo.pl**) in a directory containing Web of Science XML files. This generates a number

[1] CPC classifications for pre-1976 patents were obtained by individually querying the USPTO Patent Search webpage for individual patent numbers and then scraping the HTML to extract the technology classification. We strongly recommend against replicating this procedure and do not provide code for doing so.

[2] We created the data from USPTO-provided XML data on granted patents since 1976 (https://bulkdata.uspto.gov/) and parsed these files. Rather than taking this duplicative path, we recommend using the PatentsView files, as they have performed this step.

of files that summarize the WoS XML files.[3] They need to be combined into a single file called **wosplpubinfo.txt.**

2. Run **construct_raw_wos_with_issue.sh** such that it uses **wosplpubinfo.txt** as an input.
3. Run **buildsplitregexVOLISSPAGE.pl wos** such that it uses the output of #2 as an input, named **wosplpubinfo1955-2017_filteredISS.txt**. This will create a perl script for each year from 1955 to 2017. Run each of those perl scripts and save the output. Combine those outputs into single file.
4. Run **score_matches.pl.2018-11-08-volissuepagescoring** and save the output.

- Whether cited Web of Science papers are government-owned or government-assigned is determined by **extract.WoS.affiliaton.funding-acknowledgement.php**. Again, Web of Science XML files need to be ingested into an SQL database.

Other analyses mentioned in the article, including CPC or federal agency breakdowns as well as quality comparisons of reliant and non-reliant patents, are performed by **patent.cpc.match.sql** and require having ingested into an SQL database[4] the following in addition to **uspto.govt.reliance.metadata.tsv**:

- For comparison of patent quality based on forward citations, **uspto.citation.1976-2017.zip** and **uspto.citation.1926-1976.tab**.
- For comparison of patent quality based on patent renewals, data available at (https://bulkdata.uspto.gov/data/patent/maintenancefee/).
- For comparison of patent quality based on counts of novel words, **lexical.novelty.tsv**.

Calculating the percentage of reliant startup patents relies on data publicly available but which can be requested from Michael Ewens, (ewens@caltech.edu).

[3] This code must be executed in a Sun Grid Engine (or similar) environment that supports parallelization. Moreover, these programs may contain fully qualified pathnames that need to be either replicated or altered to work in the user's environment.

[4] The optimal method of ingesting required files into an SQL databases may depend on the user's local configuration. Variable names for ingested data and tables must match those in **patent.cpc.match.sql**. For example, the file **lexical.novelty.tsv** must be ingested into a table named "novel" with the patent number as "patno" and count of novel words as "cc."