Online Appendix:

Estimating spillovers from publicly funded R&D: Evidence from the U.S. Department of Energy

Kyle Myers* Lauren Lanahan†

January 4, 2022

*Harvard University, Harvard Business School, kmyers@hbs.edu.
†University of Oregon, Lundquist College of Business, llanahan@uoregon.edu.

A Construction of CPC-level Data

This section describes how we convert the patent and SBIR data into an input-output data set. For reference, the level of observations and key variables in the raw data sets are as follows:

1. Patent Record
    - Observation level: Patent–Inventor or Firm Assignee–CPC group
    - Key variables: year application submitted and granted; inventor and firm assignee location
2. SBIR Award Data
    - Observation level: Year–Funding Opportunity Announcement (FOA) Topic–Firm
    - Key variables: dollar amount of grant
3. SBIR FOA Data
    - Observation level: Year–Topic
    - Key variables: text of Topic description

The overall flow of the data construction is as follows:

1. Standardize CPC groups
2. Compute similarity of patent abstract text to FOA topic text
3. Collapse patent-Topic similarity scores to CPC group – FOA topic similarity scores
4. Allocate funds awarded via each Topic (to firms) into each CPC group as a function of CPC group – FOA topic similarities
5. Collapse patent flows to CPC groups

A.1 Choosing a CPC Class Level

The first major decision is choosing a level of aggregation of the CPC. Each CPC code has what is referred to as a "main trunk", which consists of five units of the form:

[1 letter][2 numbers][1 letter][1-3 numbers](/)[2-6 numbers],

i.e., A01B33/08. We could, in theory, use the full code or splice the main trunk at any of the four breaks to generate different levels of aggregation of the hierarchy. To get a sense of the range of aggregation possible, there are nine different 1-digit codes, 128 different combinations for the three digit codes, 662 combinations for the four digit codes, about 10,000 codes if spliced at the "/", and over 220,000 codes if all digits are used. For simplicity, we refer to these as Level 1-5 codes, respectively.

For example, we could group all patents together if they are labeled with the Level 2 code for "Basic Electrical Elements." Or, we could separate these patents out into the fourteen Level 3 codes within, relating to "Cables", "Resistors", "Magnets", etc. Or, we could further separate out, for example, "Magnets" into another seventeen Level 4 codes, each of which has yet another dozen or so Level 5 codes within.

If we aggregate less, then we have a larger sample size and we rely less on the idiosyncrasies of the CPC hierarchy. Conversely, the traditional advantage to aggregation is related to the Stable Unit Treatment Value Assumption (SUTVA) necessary to make causal statements from statistical models. The idea is that if the researcher aggregates units together which, for instance, are most likely to experience spillovers or substitution from treatment, then the SUTVA should hold for the newly aggregated set of units.[1]

But because we are intent on identifying the magnitude of across-technology class spillovers and do not want to rely too much on the CPC's hierarchy, we lean towards less aggregation. We choose to work with the Level 4 codes of the CPC, which we term "groups." We think this suitably balances the need to avoid reliance on the CPC hierarchy in ways that can lead to misspecification (Thompson and Fox-Kean 2005), without dividing the data into units so small that patent counts are too rare to prove useful in our analyses.

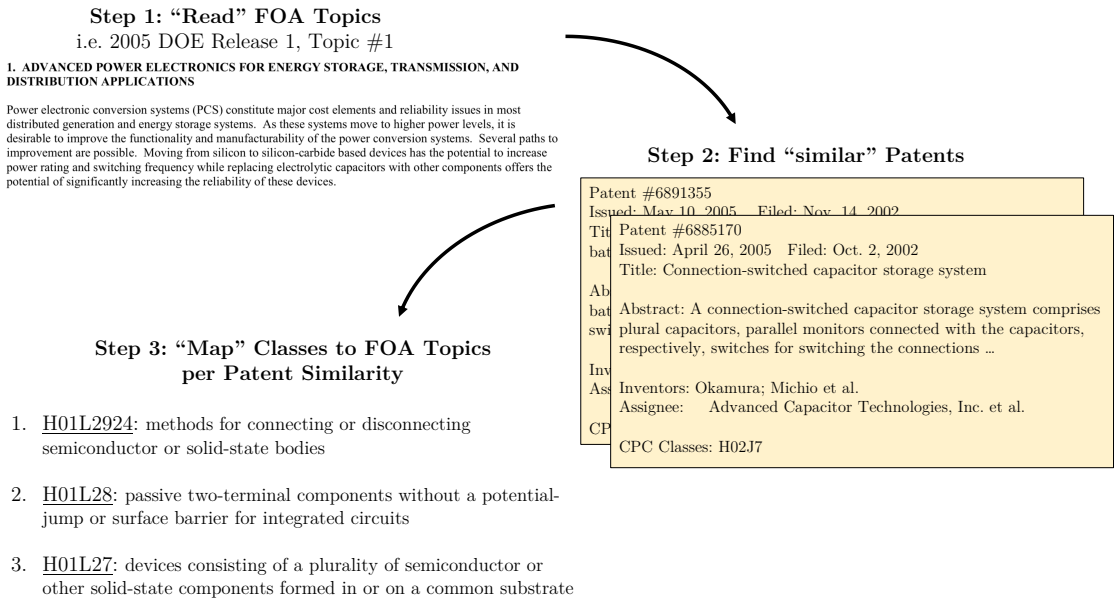
[1] However, this approach does not permit the researcher to tease out the extent to which treatment effects are "direct" or driven by spillovers across units within these aggregations.

A.2 Mapping Grants to CPC Groups

Thankfully, all patents are automatically labeled with CPC groups by the UPSTO. Our challenge then is to determine how each SBIR grant maps to a particular CPC code. We need to identify what technologies the government invests in. Our overall approach is outlined in Figure A.1.

We leverage the fact that we can connect each SBIR grant to the corresponding FOA topic that the grant application responded to. The text of these FOA topics is the key source of data we use to answer the following: if a grant recipient patented an invention that was in line with the stated goals of the topic they responded to, what CPC groups would that patent be assigned? If we know the answer to this question, then we know the dollar value of all grants awarded through each FOA topic has been "invested" in these corresponding CPC groups.

Figure A.1: Mapping Funding Opportunity Announcements to CPC Groups



We tackle this prediction problem in a series of steps that include textual similarity analysis and some simple summations and averaging. What follow are the steps of how we assign the grant dollars awarded via each FOA topic indexed by k, into the relevant CPC groups indexed by j. Following the outline, we motivate and describe the steps in further detail.

1. Estimate the text similarity S between the description of each FOA topic k and the abstract of each patent p: $S_{kp} \in [0, 1]$

2. Calculate the mean similarity $\overline{S}$ for each FOA topic to each group j based on the groups assigned to each patent, where $\mathcal{P}_j$ is the set of patents assigned group j and $N_{\mathcal{P}_j}$ is the number of patents in that set:

$$\overline{S}_{kj} = \frac{\sum_{p \in \mathcal{P}_j} S_{kp}}{N_{\mathcal{P}_j}} \tag{A.1}$$

3. Calculate percentile bins b of $\overline{S}_{kj}$, summing to create $\overline{S}^b_{kj}$, using either:
   - all $\overline{S}_{kj}$ values, or
   - $\overline{S}_{kj}$ values de-meaned at the FOA topic level
4. Calculate the twenty ventiles b (five percent groupings) of the $\overline{S}^b_{kj}$ distribution, and assume that below some percentile bin $\overline{b}$ threshold, spillovers do not occur across technology groups. (See Appendix D for more on this threshold.)
5. Evenly divide the total amount of SBIR awards given out via each FOA Topic, I_k, to all ventiles b above the percentile threshold $\overline{b}$, to give the topic-class level investment I^b_{kj}
6. Sum I^b_{kj} investments to the b and j level to obtain I^b_j – the total amount invested into group j from awards given via FOA topics that are in percentile b of similarity.

(Step 1) Text Similarity between FOA Topics & Patents: Our key assumption for this exercise is that if a patent abstract and an FOA use the same terminology, and especially if few other documents use that terminology, they are likely referring to the same technologies. This approach of exploiting the similarities between texts to link units of data has become increasingly common in economics. A number of studies leverage this approach to map "scientific space" by comparing the similarity of words used in publication abstracts (Azoulay et al. 2019, Myers 2020) and "product space" by comparing the similarity of words used in product descriptions (Hoberg and Phillips 2016). We follow the norms of modern natural language processing approaches. This includes removing "stop words" (i.e., a, the, and, etc.) and "stemming" words to remove common prefixes and suffixes. We use the commonly employed cosine function to calculate the similarity between text pairs. We also follow norms in using n-grams to identify terms, and weight these terms using the term-frequency-inverse-document-frequency (tfidf) method. We use 1-, 2-, and 3-grams in all of our specifications, which creates terms from all unique 1- 2- and 3- word combinations. Beyond 3-grams, we approach computational challenges given the size of the

matrices created.[2]

(Steps 2–3) Averaging and Shrinking Similarity Scores: Ideally, Step 1 would have been to relate FOA text directly to some description of each CPC group. However, the definitions in the CPC scheme tend to be very short pieces of text not well suited to this sort of similarity analysis. Hence, we rely on the patent abstracts. To account for the fact that these CPC-level average scores are generated from a wide range of patent documents (i.e., some CPC groups are assigned to one patent, some to thousands), we employ a standard Bayesian shrinkage estimator that compresses CPC-level means with high variances towards the overall mean.

(Step 4) Accounting for Spurious Text Correlations: Interpreting the cardinality of these scores leans rather heavily on assumptions about linguistic choices across FOAs. To avoid making inferences based on any spurious use of texts across FOAs, we undertake two alternative strategies. One is to demean the similarity scores ($\overline{S}_{kj}$) at the FOA topic level (k) and use the residuals to form the similarity percentiles. The other is to use the FOA-specific ordinal score rankings to determine the similarity percentiles. These approaches remove any variation in similarity connections that might arise purely based on how certain DoE program managers or offices write the text of the Topic descriptions. The demeaning process is our preferred approach as it does not eliminate all of the variation in scores. Though results shown in Appendix E show that using the rank approach yields very similar findings.

(Step 5) Setting the Spillover Threshold: This sort of assumption is necessary to identify the peer effects as shown by Manski (1993) and others. Appendix D details our data-driven approach to this decision in detail.

(Steps 6–7) Aggregating to Similarity Bins at the Group Level: Clearly, we need to aggregate investments to the CPC group level because this is the level of our outcome (patent flows). We use the similarity bins to identify heterogeneity in the degree to which across technology spillovers occur. We have good reason to think that, even within the assumed threshold of spillovers, the magnitude of these spillovers is likely to be an increasing function of the similarity between two groups. With these separate bins of investments, we can include multiple stocks in the production function and recover similarity-bin-specific estimates of the returns to investment, which can then be used to quantify the magnitude and shape of spillovers.

As a simple example, consider the following scenario:

[2]To avoid the endogenous use of terminology by patenters or the DoE, we use patents from before our sample, 2001-2004, to estimate the similarity scores.

- there is one FOA topic (k=1),
- there are one hundred unique CPC groups ($j = \{1, 2, ..., 100\}$),
- one grant of \$50,000 is awarded via the topic (I_k=\$50,000),
- we use two bins of size 50 ($b = \{100 - 51, 50 - 1\}$),
- we set the spillover percentile threshold to $\underline{b} = 51$, i.e., only the groups in $b = \{100{-}51\}$, the most similar 50% of groups, are allowed the possibility of spillovers.

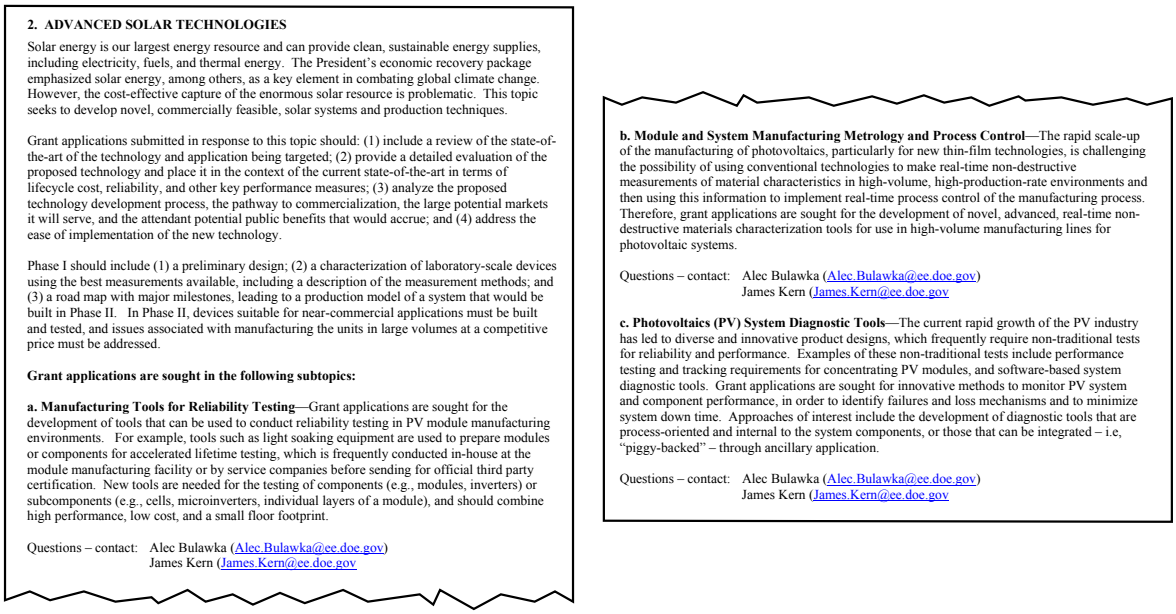
Here, I_j^b =\$1,000 for $b = \{100 - 51\}$ for the 50 groups most similar to the topic and \$0 otherwise, and I_j^b for $b = \{50 - 1\}$ is \$0 for all groups.

A.3 Crosswalk Examples

Figures A.2–A.4 each provide the text of a specific FOA topic as well as exemplary CPC titles that are matched to these topics per our methodology, sorted by their technological distance from the FOA topic description.

Figure A.2: FOA Example #1–Solar Energy

(a) FOA Text



2. ADVANCED SOLAR TECHNOLOGIES

Solar energy is our largest energy resource and can provide clean, sustainable energy supplies, including electricity, fuels, and thermal energy. The President's economic recovery package emphasized solar energy, among others, as a key element in combating global climate change. However, the cost-effective capture of the enormous solar resource is problematic. This topic seeks to develop novel, commercially feasible, solar systems and production techniques.

Grant applications submitted in response to this topic should: (1) include a review of the state-of-the-art of the technology and application being targeted; (2) provide a detailed evaluation of the proposed technology and place it in the context of the current state-of-the-art in terms of lifecycle cost, reliability, and other key performance measures; (3) analyze the proposed technology development process, the pathway to commercialization, the large potential markets it will serve, and the attendant potential public benefits that would accrue; and (4) address the ease of implementation of the new technology.

Phase I should include (1) a preliminary design; (2) a characterization of laboratory-scale devices using the best measurements available, including a description of the measurement methods; and (3) a road map with major milestones, leading to a production model of a system that would be built in Phase II. In Phase II, devices suitable for near-commercial applications must be built and tested, and issues associated with manufacturing the units in large volumes at a competitive price must be addressed.

Grant applications are sought in the following subtopics:

a. Manufacturing Tools for Reliability Testing—Grant applications are sought for the development of tools that can be used to conduct reliability testing in PV module manufacturing environments. For example, tools such as light soaking equipment are used to prepare modules or components for accelerated lifetime testing, which is frequently conducted in-house at the module manufacturing facility or by service companies before sending for official third party certification. New tools are needed for the testing of components (e.g., modules, inverters) or subcomponents (e.g., cells, microinverters, individual layers of a module), and should combine high performance, low cost, and a small floor footprint.

Questions – contact: Alec Bulawka (Alec.Bulawka@ee.doe.gov)
James Kern (James.Kern@ee.doe.gov

b. Module and System Manufacturing Metrology and Process Control—The rapid scale-up of the manufacturing of photovoltaics, particularly for new thin-film technologies, is challenging the possibility of using conventional technologies to make real-time non-destructive measurements of material characteristics in high-volume, high-production-rate environments and then using this information to implement real-time process control of the manufacturing process. Therefore, grant applications are sought for the development of novel, advanced, real-time non-destructive materials characterization tools for use in high-volume manufacturing lines for photovoltaic systems.

Questions – contact: Alec Bulawka (Alec.Bulawka@ee.doe.gov)
James Kern (James.Kern@ee.doe.gov

c. Photovoltaics (PV) System Diagnostic Tools—The current rapid growth of the PV industry has led to diverse and innovative product designs, which frequently require non-traditional tests for reliability and performance. Examples of these non-traditional tests include performance testing and tracking requirements for concentrating PV modules, and software-based system diagnostic tools. Grant applications are sought for innovative methods to monitor PV system and component performance, in order to identify failures and loss mechanisms and to minimize system down time. Approaches of interest include the development of diagnostic tools that are process-oriented and internal to the system components, or those that can be integrated – i.e, "piggy-backed" – through ancillary application.

Questions – contact: Alec Bulawka (Alec.Bulawka@ee.doe.gov)
James Kern (James.Kern@ee.doe.gov

(b) Titles of Relevant CPC Classes

Technology Distance ptile.	Example CPC Titles
1	Apparatus for processing exposed photographic materials
	Generation of electric power by conversion of infra-red radiation, visible light or ultraviolet light
	Plasma technique; production of accelerated electrically-charged particles
10	Electric heating; electric lighting
	Static electricity; naturally-occurring electricity
	Cyclically operating valves for machines or engines
20	Cranes; load-engaging elements or devices for cranes
	Locomotives; motor railcars
	Wireless communication networks

Notes: Topic #2 from the FY2010 Release 1 Funding Opportunity Announcement.

Figure A.3: FOA Example #2–Geothermal Energy

(a) FOA Text

4. GEOTHERMAL ENERGY TECHNOLOGY DEVELOPMENT

This topic is focused on the development and innovation required to achieve technical and commercial feasibility of EGS. Because of the complexity of these systems, grant applications are expected to focus on a component or supporting technology of EGS development that would enable improvements to the overall system. The unique function and innovation of the targeted subsystem or supporting technology must be clearly described and its function in relationship to the greater EGS system must be expressed clearly. Approaches can be targeted at any of the multi-step project stages for technology development: from design concept, through scale model development (if applicable), to laboratory testing, field testing, and commercial scale demonstrations.

Grant applications are sought in the following subtopics:

a. High Temperature Downhole Logging and Monitoring Tools—Challenging subsurface conditions are one of the barriers to an accelerated ramp-up of geothermal energy generation. To address this challenge, grant applications are sought to develop logging and monitoring tools that are capable of tolerating extreme environments of high temperatures and pressures. The instruments of interest include, but are not limited to, temperature and pressure sensors, flow meters, fluid samplers, inclination and direction sensors, acoustic instruments (high and low frequency), resistivity probes, natural gamma ray detectors, epithermal neutron scattering gauges, rock density gauges (gamma and sonic), casing monitoring devices (e.g. cement bond logs and casing collar locators), fluid conductivity, pH indicators and well dimension probes (caliper). The target temperatures and pressures for these logging and monitoring tools should be supercritical conditions (374° C and 220 bar for pure water), and the tools may be used at depths of up to 10,000 meters.

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov.

b. Cements for EGS Applications—While conventional geothermal wells experience large temperature rises during production, EGS wells experience large temperature drops at the bottom of the well during the stimulation process, due to the cooling effect of the injected water. This temperature drop may be in the neighborhood of 350°F. This unique situation causes significant stress and potential failure of the cement sheath if conventional cement systems are utilized. To address this issue, grant applications are sought for the research, design, development, testing, and demonstration of a cement system for the high temperature and stress conditions of an EGS wellbore. Proposed approaches may define cement formulations that would be used by the geothermal industry to place the cement within a long string of casings; such approaches should focus on preventing a premature set and maintaining a strong seal at the shoe (so that stimulations may be performed through the casing).

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov.

c. Drilling Systems—High upfront costs, largely due to high drilling costs, are a major barrier to expanded geothermal energy production in the United States. Therefore, grant applications are sought to reduce drilling costs by developing a drilling technology (horizontal and/or directional) that is capable of drilling three times faster than conventional rotary drilling. Approaches of interest include, but are not limited to the design and development of improved drilling fluids (to reduce frictional viscosity and remove cuttings), high-performance bottom-hole assemblies (e.g., collars, bent subs, drill bits), and downhole motors (to control wellbore orientation). Proposed approaches must demonstrate reliable operation and equipment durability that exceeds the performance of conventional equipment at depths up to 10,000 meters and temperatures up to 300° C.

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov.

d. Fracture Characterization Technologies—Subsurface imaging is an important part of creating a productive EGS reservoir, which requires visualization before, during, and after creation. In order to advance technology and reduce the upfront risk to geothermal projects, more robust subsurface imaging technologies must be developed. Grant applications are sought to develop improved downhole and remote imaging methods to characterize fractures. Fracture characterization includes prediction of fracture and stress orientation prior to drilling (needed to properly orient horizontal wells); determination of fracture location, spacing, and orientation (while drilling); and determination of the location of open fractures (after stimulation), in order to identify the location of fluid flow pathways within the enhanced geothermal reservoir. Proposed approaches should address robust methods for interpreting and imaging the subsurface, including but not limited to, the development of active or passive seismic, processing software, and joint inversion of geophysical techniques.

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov.

e. Working Fluids for Binary Power Plants—Binary power plants are rapidly becoming a major part of the geothermal industry, due to increased development of lower temperature geothermal resources. To address cost barriers associated with the working fluids in these binary power plants, grant applications are sought to (1) identify non-azeotropic mixtures of working fluids for improved utilization of available energy in subcritical cycles; (2) characterize the composition and thermophysical and transport properties of those mixtures; (3) identify working fluids for supercritical cycles and trilateral cycles; and (4) characterize the composition, thermophysical, and transport properties of those working fluids. Proposed approaches may address working fluids or mixtures of working fluids with the potential for greater energy conversion efficiency than conventional working fluids, such as isobutane or refrigerants.

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov.

f. GHP Component R&D—High initial costs have been identified as a key barrier to widespread GHP deployment. To address this barrier, applications are sought to improve GHP components to increase efficiency as well as energy savings as compared to conventional systems. Applications may address but are not limited to: variable-speed (VS) components, advanced sensors and controls (including water flow sensing), electronic expansion valves, heat exchange (HX) design and fluids, system optimization, unit control algorithms, and load management tools.

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov

g. Innovative System/Loop Designs—One of the main barriers in GHP technology is the high cost of drilling and loop installation. Applications are sought for innovative system/loop designs that reduce the costs of system and/or loop installation, through new design layouts, system components, materials, and/or methods.

Questions – Contact Raymond Fortuna, 202-586-1711, raymond.fortuna@ee.doe.gov

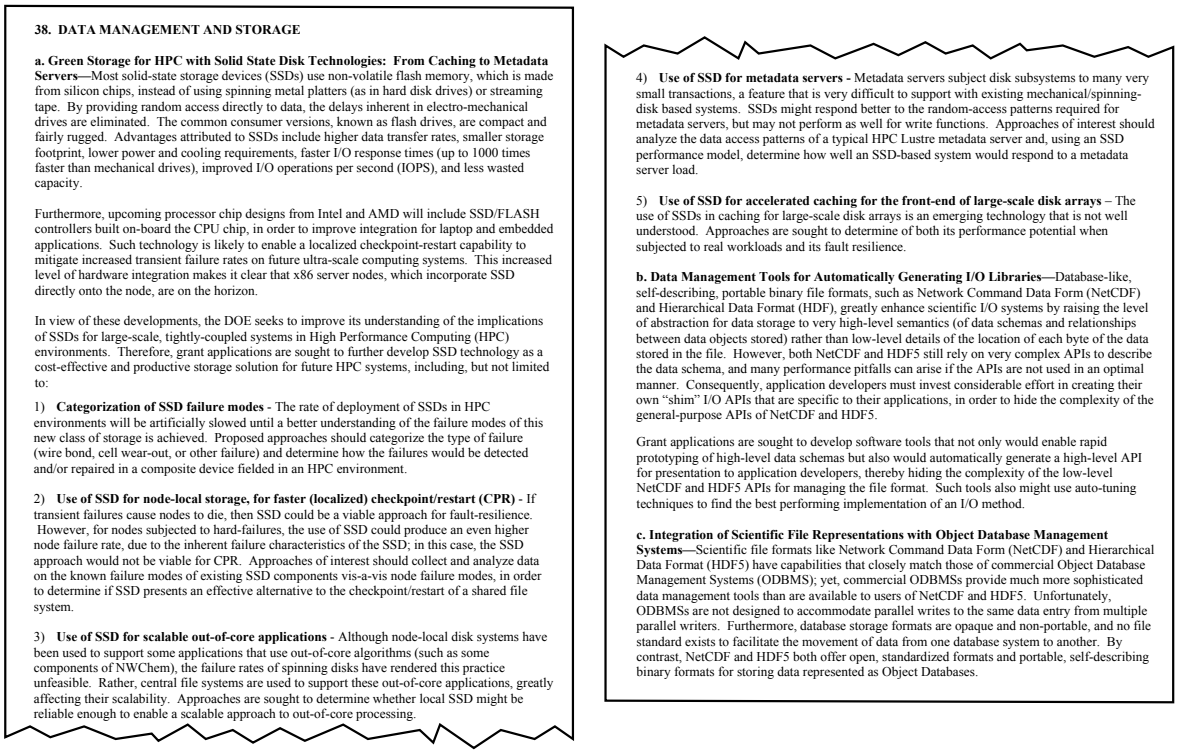
(b) Titles of Relevant CPC Classes

Technology Distance ptile.	Example CPC Titles
1	Geophysics; gravitational measurements
	Positive-displacement machines for liquids; pumps
	Collection, production or use of heat
10	Electric heating; electric lighting
	Static electricity; naturally-occurring electricity
	Cyclically operating valves for machines or engines
20	Installations or methods for obtaining, collecting, or distributing water
	Computer systems based on specific computational models
	Vehicles, vehicle fittings, or vehicle parts

Notes: Topic #4 from the FY2010 Release 1 Funding Opportunity Announcement.

Figure A.4: FOA Example #3–Data Management

(a) FOA Text



38. DATA MANAGEMENT AND STORAGE

a. Green Storage for HPC with Solid State Disk Technologies: From Caching to Metadata Servers—Most solid-state storage devices (SSDs) use non-volatile flash memory, which is made from silicon chips, instead of using spinning metal platters (as in hard disk drives) or streaming tape. By providing random access directly to data, the delays inherent in electro-mechanical drives are eliminated. The common consumer versions, known as flash drives, are compact and fairly rugged. Advantages attributed to SSDs include higher data transfer rates, smaller storage footprint, lower power and cooling requirements, faster I/O response times (up to 1000 times faster than mechanical drives), improved I/O operations per second (IOPS), and less wasted capacity.

Furthermore, upcoming processor chip designs from Intel and AMD will include SSD/FLASH controllers built on-board the CPU chip, in order to improve integration for laptop and embedded applications. Such technology is likely to enable a localized checkpoint-restart capability to mitigate increased transient failure rates on future ultra-scale computing systems. This increased level of hardware integration makes it clear that x86 server nodes, which incorporate SSD directly onto the node, are on the horizon.

In view of these developments, the DOE seeks to improve its understanding of the implications of SSDs for large-scale, tightly-coupled systems in High Performance Computing (HPC) environments. Therefore, grant applications are sought to further develop SSD technology as a cost-effective and productive storage solution for future HPC systems, including, but not limited to:

1) **Categorization of SSD failure modes -** The rate of deployment of SSDs in HPC environments will be artificially slowed until a better understanding of the failure modes of this new class of storage is achieved. Proposed approaches should categorize the type of failure (wire bond, cell wear-out, or other failure) and determine how the failures would be detected and/or repaired in a composite device fielded in an HPC environment.

2) **Use of SSD for node-local storage, for faster (localized) checkpoint/restart (CPR) -** If transient failures cause nodes to die, then SSD could be a viable approach for fault-resilience. However, for nodes subjected to hard-failures, the use of SSD could produce an even higher node failure rate, due to the inherent failure characteristics of the SSD; in this case, the SSD approach would not be viable for CPR. Approaches of interest should collect and analyze data on the known failure modes of existing SSD components vis-a-vis node failure modes, in order to determine if SSD presents an effective alternative to the checkpoint/restart of a shared file system.

3) **Use of SSD for scalable out-of-core applications -** Although node-local disk systems have been used to support some applications that use out-of-core algorithms (such as some components of NWChem), the failure rates of spinning disks have rendered this practice unfeasible. Rather, central file systems are used to support these out-of-core applications, greatly affecting their scalability. Approaches are sought to determine whether local SSD might be reliable enough to enable a scalable approach to out-of-core processing.

4) **Use of SSD for metadata servers -** Metadata servers subject disk subsystems to many very small transactions, a feature that is very difficult to support with existing mechanical/spinning-disk based systems. SSDs might respond better to the random-access patterns required for metadata servers, but may not perform as well for write functions. Approaches of interest should analyze the data access patterns of a typical HPC Lustre metadata server and, using an SSD performance model, determine how well an SSD-based system would respond to a metadata server load.

5) **Use of SSD for accelerated caching for the front-end of large-scale disk arrays** – The use of SSDs in caching for large-scale disk arrays is an emerging technology that is not well understood. Approaches are sought to determine of both its performance potential when subjected to real workloads and its fault resilience.

b. Data Management Tools for Automatically Generating I/O Libraries—Database-like, self-describing, portable binary file formats, such as Network Command Data Form (NetCDF) and Hierarchical Data Format (HDF), greatly enhance scientific I/O systems by raising the level of abstraction for data storage to very high-level semantics (of data schemas and relationships between data objects stored) rather than low-level details of the location of each byte of the data stored in the file. However, both NetCDF and HDF5 still rely on very complex APIs to describe the data schema, and many performance pitfalls can arise if the APIs are not used in an optimal manner. Consequently, application developers must invest considerable effort in creating their own "shim" I/O APIs that are specific to their applications, in order to hide the complexity of the general-purpose APIs of NetCDF and HDF5.

Grant applications are sought to develop software tools that not only would enable rapid prototyping of high-level data schemas but also would automatically generate a high-level API for presentation to application developers, thereby hiding the complexity of the low-level NetCDF and HDF5 APIs for managing the file format. Such tools also might use auto-tuning techniques to find the best performing implementation of an I/O method.

c. Integration of Scientific File Representations with Object Database Management Systems—Scientific file formats like Network Command Data Form (NetCDF) and Hierarchical Data Format (HDF5) have capabilities that closely match those of commercial Object Database Management Systems (ODBMS); yet, commercial ODBMSs provide much more sophisticated data management tools than are available to users of NetCDF and HDF5. Unfortunately, ODBMSs are not designed to accommodate parallel writes to the same data entry from multiple parallel writers. Furthermore, database storage formats are opaque and non-portable, and no file standard exists to facilitate the movement of data from one database system to another. By contrast, NetCDF and HDF5 both offer open, standardized formats and portable, self-describing binary formats for storing data represented as Object Databases.

(b) Titles of Relevant CPC Classes

Technology Distance ptile.	Example CPC Titles
1	Electric digital data processing
	Apparatus or arrangements for taking photographs or for projecting or viewing them
	Transmission of digital information, e.g. telegraphic communication
10	Information and communication technology adapted for specific application fields
	Radio-controlled time-pieces
	Secret communication; jamming of communication
20	Presses in general
	Production of cellulose by removing non-cellulose substances
	Methods of steam generation; steam boilers

Notes: Topic #38 from the FY2010 Release 1 Funding Opportunity Announcement.

B Other Estimation & Data Notes

B.1 R&D Stock Construction

We use standard perpetual inventory methods to construct the stock of R&D investments: $K_{jt} = I_{jt} + (1-\rho)K_{j(t-1)}$ where I_{jt} is the investment flow and ρ is the discount rate. In our preferred specifications we use no discounting, as it gives us the most conservative estimates. All dollar values are adjusted for inflation using the 2018 Consumer Price Index. Results under alternative discounting assumptions are shown in Appendix E.4.

B.2 Approximating Elasticities when Negative Numbers are Present

The following demonstrates the usefulness of the "demeaning" transformation to handle the fact that our focal independent variable, the stock of state match windfall investments, can take on negative values. For the purpose of clarity, consider the following "log-log" linear regression model that is analogous to our preferred specification:

$$\begin{aligned} \log(Y_j) &= \alpha + \log(X_j)\beta + \epsilon_j, \\ \beta &= \frac{\partial \log(Y_j)}{\partial \log(X_j)} = \frac{\partial Y_j}{\partial X_j}\frac{X_j}{Y_j}; \end{aligned} \tag{B.2}$$

$$\begin{aligned} \log(Y_j) &= \tilde{\alpha} + \frac{X_j}{\overline{X}}\theta + \tilde{\epsilon}_j, \\ \theta &= \frac{\partial \log(Y_j)}{\partial \frac{X_j}{\overline{X}}} = \frac{\partial Y_j}{\partial X_j}\frac{\overline{X}}{Y_j}. \end{aligned} \tag{B.3}$$

Below the regression models, which relate the dependent variable Y to the independent variable X, we also define the coefficients of interest: β and θ. Eq. B.2 shows the useful result that average elasticities are estimated directly when using the log-log transformation. In the case of Eq. B.3, a similar coefficient is estimated, although now instead of estimating a "mean elasticity" given by β, the θ coefficient describes the elasticity across all values of Y_j, but at the sample mean of X_j, denoted here by $\overline{X}$.

In practice, we assume that any difference between these two parameters is negligible. To motivate this assumption, the following shows that θ approximates β. Substituting a first order

Taylor series approximation of $\log(X_j)$ around $\overline{X}$, $\log(\overline{X}) + \frac{X_j - \overline{X}}{\overline{X}}$ into Eq. B.2 yields:

$$\begin{aligned} \log(Y_j) &\approx \alpha + \Big(\log(\overline{X}) + \frac{X_j - \overline{X}}{\overline{X}}\Big)\beta + \epsilon_j, \\ \log(Y_j) &\approx \big[\alpha + \log(\overline{X})\beta - \beta\big] + \frac{X_j}{\overline{X}}\beta + \epsilon_j, \end{aligned} \tag{B.4}$$

where $\big[\alpha + \log(\overline{X})\beta - \beta\big] \approx \tilde{\alpha}$, mapping to Eq. B.3.

B.3 Calculating Implied Marginal Products

Our regressions yield estimates of the output elasticity of SBIR funding based on variation across CPC groups over time. But this variation is not due to the DoE directly choosing to invest in a CPC group per se, but rather it is due to their decision to invest in certain FOA topics (which are in turn connected to different CPC groups). Thus, when we use these elasticities to estimate the marginal product of additional investments, we do so in a way that reflects the FOA topic-level source of variation.

Our approach to obtaining marginal products is as follows: first, note that given patent flows Y, R&D stocks K, and single elasticity estimate θ, the implied group-year (jt) level marginal product of a dollar is: $\theta \times Y_{jt} \times \frac{1}{K_{jt}} \equiv MP_{jt}$.[3]

Next, we make use of the fact that we know which CPC groups each FOA topic directs funding towards. Let w_{jk} be an indicator variable that equals one if CPC group j is deemed relevant to the funding from FOA topic k per our text similarity approach (as in, within the boundary of technological distances we consider). With these weights w_{jk}, we can estimate a k-specific weighted average of MP_{jt} for each FOA topic: $\sum_{jt} MP_{jt} w_{jk} / \sum_{jt} w_{jk}$. Then, by averaging over the roughly 1,000 FOA topics in our data, we can arrive at value that more closely approximates the true average marginal product of increased funding.

Lastly, we accommodate the fact that we have multiple θ estimates, one for each technological distance bin in Eq. 4, and also for the fact that our goal is to arrive at an estimate of the marginal cost per patent (not the marginal cost of increasing the annual patent flow rate by one). To the first point, the additive separability of the production function implies that we can simply sum over the multiple MP_{jt} values calculated for each bin. However, since our data construction approach evenly divides funding across all bins, we modify MP_{jt} to be: $\theta^b \times Y_{jt} \times \frac{1}{K^b_{jt}} \times \frac{1}{N^b}$, where b indexes bins and N^b indicates the number of bins. This captures the fact that when the DoE invests a marginal dollar, our data construction funnels equal portions of that dollar into each of the N^b bins. In the final step, we convert marginal increases in flow rates into total patent output by simply summing up the net increase in patent flows to be expected if we followed the observation for the remainder of the sample period. This mimicks a firm-level analysis of how SBIR funding would increase the stock of patents a firm produces (e.g., Howell 2017).

[3] This clearly is undefined for zero-valued patent flows or investment stocks, but elsewhere in the Appendix we report results where we estimate the θ parameters using only non-zero observations and obtain very similar estimates, which indicates that our assumption of a constant elasticity is reasonable.

B.4 Estimating Within-US Travel Costs

Although geographic distance separates inventors, the implication of much empirical work on the geographic distribution of invention (e.g., Agrawal et al. 2017) is that the costs of human travel, not geographic distance per se, constrains the flow of ideas. It is beyond the scope of this paper to estimate highly accurate travel costs given the high dimensionality of the data and numerous modes of transportation possible. However, we make strides in this direction by: (1) focusing on US county-to-county pairs as semi-dense yet computationally tractable set of regions to focus on; (2) using US Internal Revenue Service (IRS) driving mileage rates to approximate driving costs between all counties[4]; (3) using the Department of Transportation (DoT) Airline Origin and Destination Survey (DBIB) to obtain negotiated airfare rates between all US airports; (4) NBER Place Distance Database; and (5) solving for the minimum cost of traveling between each county pair in the US using the minimum of either these approximate costs of driving directly, or driving to the nearest airports and flying.[5] Then, by taking the set of counties where DoE SBIR grants (that focus on a particular technology group) are awarded as the focal set of counties, we can calculate the average cost of making a round trip to these focal counties for individuals in all other counties.

We define the mode of transit based on the proximity of the origin and destination county to an airport. This defines four possible paths: (i) origin county with an airport to destination county with an airport; (ii) origin county without an airport to destination county with an airport; (iii) origin county with an airport to destination county without an airport; and (iv) origin county without an airport to destination county without an airport. For (i), we simply use the average annual market fares reported in DB1B to compute the travel cost. For the paths that include a county without an airport, we add the cost of driving from the center of a county without an airport to the center of the closest county with an airport. We rely on the mileage – as reported by NBER – and the IRS standard mileage reimbursement rate to compute this driving cost. For the total cost, we add the ground and air transportation accordingly. For all cost measures, we adjust for inflation using the 2018 CPI adjusted index.

We impute fares using observed data as follows: for airports a and b, we estimate the cost of a round trip flight using the following regression:

$$\text{fare}_{ab} = \alpha_a + \beta_b + \text{geographic distance}_{c(a),c(b)}\delta + \epsilon_{ab}, \tag{B.5}$$

[4]As well as an estimate of the conversion between geographic distance and driving distance.

[5]DoT data is available at: http://bit.ly/2RSwkG1. NBER data is available at: http://bit.ly/2U6TtHk. IRS data is available at: http://bit.ly/37wuVvl.

where α and β are airport fixed effects, $c(\cdot)$ defines the county that an airport is located in, and the parameter δ describes how fares grow with the distance traveled. Then we use the estimated values of α, β, and δ to predict fares both in-sample (where we observe fare_{ab}) and out-of-sample (where we do not observe fare_{ab}, but we do observe at least two fares for a and/or b). We use these imputed fares, combined with population levels, to create the population-weighted geographic distance groupings that we use in our main analyses.

B.5 Additional Data Sources

The following outlines data used in the search for correlates of spillovers at the domestic and international levels as discussed in Appendix F.

B.5.1 BEA Economic Area Data

US Cluster Mapping: The majority of features we use to describe US economic areas was obtained from the US Cluster Mapping project (www.clustermapping.us; Delgado et al. 2016)

Universities and colleges: Location and average total annual R&D funding for major US universities and colleges was obtained from the National Science Foundation's Higher Education Research and Development Survey (HERD; www.nsf.gov/statistics/srvyherd).

Federally Funded R&D Centers (FFRDC): The locations of all FFRDCs was obtained from the National Science Foundation (www.nsf.gov/statistics/ffrdclist).

Nuclear reactors: The locations and types of all operating nuclear reactors was obtained from the Nuclear Regulatory Commission (www.nrc.gov/info-finder/reactors/index.html).

Travel costs: See B.4.

B.5.2 International Data

World Development Indicators (WDI): The majority of features we use to describe countries was obtained from the World Bank's WDI (datacatalog.worldbank.org/dataset/world-development-indicators).

Geography and travel: Geographic and travel distances between countries was obtained using the *GeoDist* database created by the Centre d'Études Prospectives et d'Informations Internationales (Mayer and Zignago 2011).

Trade: Bilateral trade flows between all in-sample countries and the US was obtained via the UN Comtrade Database (comtrade.un.org).

FDI: Data on Foreign Direct Investments (FDI) to/from the US was obtained from the US BEA (www.bea.gov/international/di1fdibal).

Migrant stocks: International Migrant Stocks of all US-country pairs was obtained from the UN Population Division (www.un.org/development/desa/pd/content/international-migrant-stock).

C State Match Policies

C.1 Isolating the Windfall Funding

The variation in funding across technology groups is due to the variation in funding across FOAs. The total amount I invested into each FOA k in a year t is the sum of federal awards I_{kt}^{fed} and any matching funds awarded by states I_{kt}^{match}. We are concerned that the federal investments are endogenous, in that they may be correlated with unobservable productivity or demand shocks. Thus, we cannot assume that the variation due to state matches (which are a function of federal investments) is exogenous. Thus, we isolate the variation in state match investments across FOAs that arises only due to the distribution of SBIR grant winners across states with and without matching policies.

We first write match investments I_{kt}^{match} as a function of federal investments I_{kt}^{fed}:

$$I_{kt}^{\text{match}} = \alpha + I_{kt}^{\text{fed}}\gamma_t + W_{kt} \text{ if } I_{kt}^{\text{fed}} > 0, \tag{C.1}$$

where α is a constant and the parameter γ_t captures the fact that all (year-specific) matching programs are a linear transformation of federal awards (i.e., if all have states a 50% matching rate and received the same level of federal investments, then $\gamma = 0.5$).[6] If we estimate Eq. C.1 via OLS, our estimate of γ_t reflects the average of match rates across the country in year t, weighted by the amount of I_{kt}^{fed} invested in each state. The residual W_{kt} – which we term the state match windfall – arises because firms working on different technologies are differentially concentrated in states with or without matching programs. Therefore, our key identification assumption is that the distributions of firms and state policies are not related to any unobservable productivity or demand shocks. In other words, it cannot be the case that more (less) productive firms working on technologies with larger (smaller) unobservable productivity shocks are concentrated in states with larger (smaller) matching rates.

We know the true grant-level match rates for each state, which is how we construct I_{kt}^{match}. But we must estimate the effective "FOA-level" match rates, the γ_t parameters, from the data. To ensure that each observation's windfall estimate is not driven by that observation's role in determining our estimate of these γ_t parameters, we take a jackknife approach in the spirit of Angrist et al. (1999) and construct our windfall estimates as:

$$W_{kt} = I_{kt}^{\text{match}} - \widehat{\alpha^{-k}} - I_{kt}^{\text{fed}}\widehat{\gamma_t^{-k}}, \tag{C.2}$$

[6] In practice, many states' programs operate as lump sum matches. But because the size of Phase I and Phase II awards is largely standardized, this is effectively equivalent to the states setting a match rate.

where $\widehat{\alpha^{-k}}$ and $\widehat{\gamma_t^{-k}}$ are our estimates of those parameters from estimating Eq. C.1 while excluding FOA k from the regression.

C.2 Comparison of States with SBIR Match Policies

In this subsection, we test two hypotheses: (1) state-years with SBIR match programs do not systematically differ in the amount of federal funding awarded to the firms located in that state-year, and (2) small firms are not more or less likely to relocate into or out of state-years that have enacted an SBIR matching program.

To test the first hypothesis, we estimate the following regression model at the state-year *st* level:

$$\text{SBIR Funding}_{st} = \text{Match Status}_{st}\beta + \tau_t + \epsilon_{st} \ , \tag{C.3}$$

where a significant estimate of the β would reject our hypothesis of the null. We explore three measures of a stat-years' match status: (1) whether there is any match, and then whether the match rate is (2) above the median or (3) below the median.

As shown in Table C.1, which uses two different measures of SBIR funding (total funding in cols. 1–4 and funding per capita in cols. 5–8), state-years with match policies generally appear to receive less federal SBIR funding, but in no specification do we estimate a statistically significant value for β. The point estimates generally suggest that state-years with any match have about 30-50% less federal funding, but these estimates have large standard errors such that we cannot reject a null of no difference. Furthermore, in the linear models we obtain very small R^2 values, suggesting these policies are likely not directly responsible for (or indirectly correlated with) larger shifts within states surrounding the SBIR program. This gives us confidence that there are not significant differences in the SBIR-involved firms residing in states with or without the match programs, as our identification strategy assumes.

To test the second hypothesis – that small firms are not moving into or out of states with matching programs – we use data from the National Establishment Time Series spanning 2000 to 2015 and covering 13,325 small firms to estimate the following model of firms' location:

$$\mathbf{1}\{\text{Location}_{ist}\} = \text{Match Status}_{st}\beta_{DoE(i)} + \alpha_i + \sigma_s + \tau_t + \epsilon_{st} \tag{C.4}$$

where firm-state-year observations are indexed by *ist*, $\mathbf{1}\{\text{Location}_{ist}\}$ is a dummy variable indicating that firm i resides in state s in year t, Match Status_{st} is again a dummy variable indicating whether state s has a matching program in place in year t, and α_i, σ_s, τ_t are firm, state, and year fixed effects sometimes included in the models.

All firms in the sample win an SBIR award at some time from the DoE or any of the other major federal agencies. We can observe firms' locations from the time they first appear in

Table C.1: Federal DoE SBIR Flows by State Match Status

	SBIR $M				SBIR $ per M capita			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any match	–0.366	–0.589			–0.719	–0.357		
	(0.280)	(0.402)			(1.131)	(0.305)		
High match			–0.375	–0.573			0.0913	–0.311
			(0.372)	(0.568)			(1.102)	(0.369)
Low match			–0.396	–0.633			–1.591	–0.393
			(0.358)	(0.473)			(1.491)	(0.453)
N	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.038		0.039		0.031		0.034	
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Model	OLS, ihs	PPML	OLS, ihs	PPML	OLS, ihs	PPML	OLS, ihs	PPML

Notes: Reports various estimates of β from Eq. C.3 using either state-wide annual SBIR funding in $-millions (cols. 1–4; mean=3.8, s.d.=6.9), or state-wide annual SBIR funding in $-per-million-people (cols. 5–8; mean=631,000, s.d.=973,000). Standard errors clustered at the state level. "OLS, ihs" indicates model is estimated using OLS with an inverse hyperbolic sine transformation of the dependent variable. "PPML" indicates model is estimated as a Poisson pseudolikelihood regression.

the National Establishment Time Series data onwards, and can match about 80% of ever-SBIR-winners. We allow firms' response to vary as a function of whether they ever receive a DoE award or not, as indicated by the $\beta_{DoE(i)}$ parameter.

Table C.2 reports the results of these regressions, which are economically insignificant. Regardless of how much we saturate the model (or not) with fixed effects, there is no meaningful association between the movement of these award winners and the state matching policies. In all cases, we cannot reject a null that this set of firms are neither more nor less likely to locate in states when a match policy is in place. This suggests that the firms winning awards in states with matching policies likely did not relocate their firm to the state because of the matching policies or any of the underlying economic or political forces that motivated the enactment of those policies.

Table C.2: Small Firm Locations per State Match Status

	(1)	(2)	(3)
Never DoE SBIR	–0.00964	–0.000216	–0.000201
× State Match	(0.00575)	(0.000441)	(0.000443)
Ever DoE SBIR	–0.0104	–0.000936	–0.00107
× State Match	(0.00561)	(0.00117)	(0.00126)
N	7,006,300	7,006,300	7,006,300
R^2	0.001	0.046	0.046
Year FE	Y	Y	Y
State FE		Y	Y
Firm FE			Y

Notes: Reports various estimates of the β parameters from Eq. C.4 using varying degrees of year, state, and firm fixed effects. The mean of the dependent variable is $1/50$ since the dependent variable is whether or not a given firm resides in a given state in a given year. Standard errors clustered at the state level.

D Technological Spillover Boundary Search

This section outlines our approach to arriving at an assumption about the boundary of technological spillovers. This assumption is of the sort motivated by Manski (1993), which is required to achieve identification of peer effects. For reference, our focal production function is as follows:

$$\mathbb{E}[Y_{jt}^{d} \,|\, W_{jtb}] = \exp\Big(\sum_{b\in\mathcal{B}} \frac{W_{jtb}}{\overline{W}}\theta_b^d + \tau_t^d\Big) \,. \tag{D.1}$$

In the context of this model, the question is what should be chosen for the maximum technological distance bin b in $\mathcal{B}$. In other words, what is the threshold of the textual similarity score between an FOA topic and a CPC group beyond which we can safely assume there is not effect of investment in that FOA topic on the patent flows in that CPC group?

Clarke (2017) tackles this exact question, framing it as a bandwidth selection problem (similar to the challenge of determining the optimal bandwidth for regression discontinuity designs). To illustrate the approach, let us instead consider a simple Poisson model:

$$\mathbb{E}[y_i \,|\, x_i] = \exp(\alpha + x_i\beta) \,.$$

Letting $\widehat{\alpha^{-i}}$ and $\widehat{\beta^{-i}}$ indicate the parameter estimates from a regression where i is excluded, we construct two predicted values of the outcome to use in the cross validation procedure:

$$\begin{aligned}\widehat{y}_i &= \exp(\widehat{\alpha^{-i}} + x_i\widehat{\beta^{-i}}),\\ \widetilde{y}_i &= \widehat{\alpha^{-i}} + x_i\widehat{\beta^{-i}}.\end{aligned}$$

But instead of obtaining parameter estimates for every i, we simplify the number of computations by following Clarke (2017) and use a "k-fold" approach, setting k to ten as is commonplace. This approach randomly splits the sample of N observations into ten equal-size groups and estimates a series of ten regressions, where each group is excluded from estimation and the resulting parameter estimates are used to form the expected values for that excluded group.

We test three different measures that are commonly used in practice for assessing the cross

validation fit of Poisson models:

$$\text{Mean Squared Error (MSE): } \frac{1}{N}\sum_i (y_i - \widehat{y}_i)^2 \ ,$$

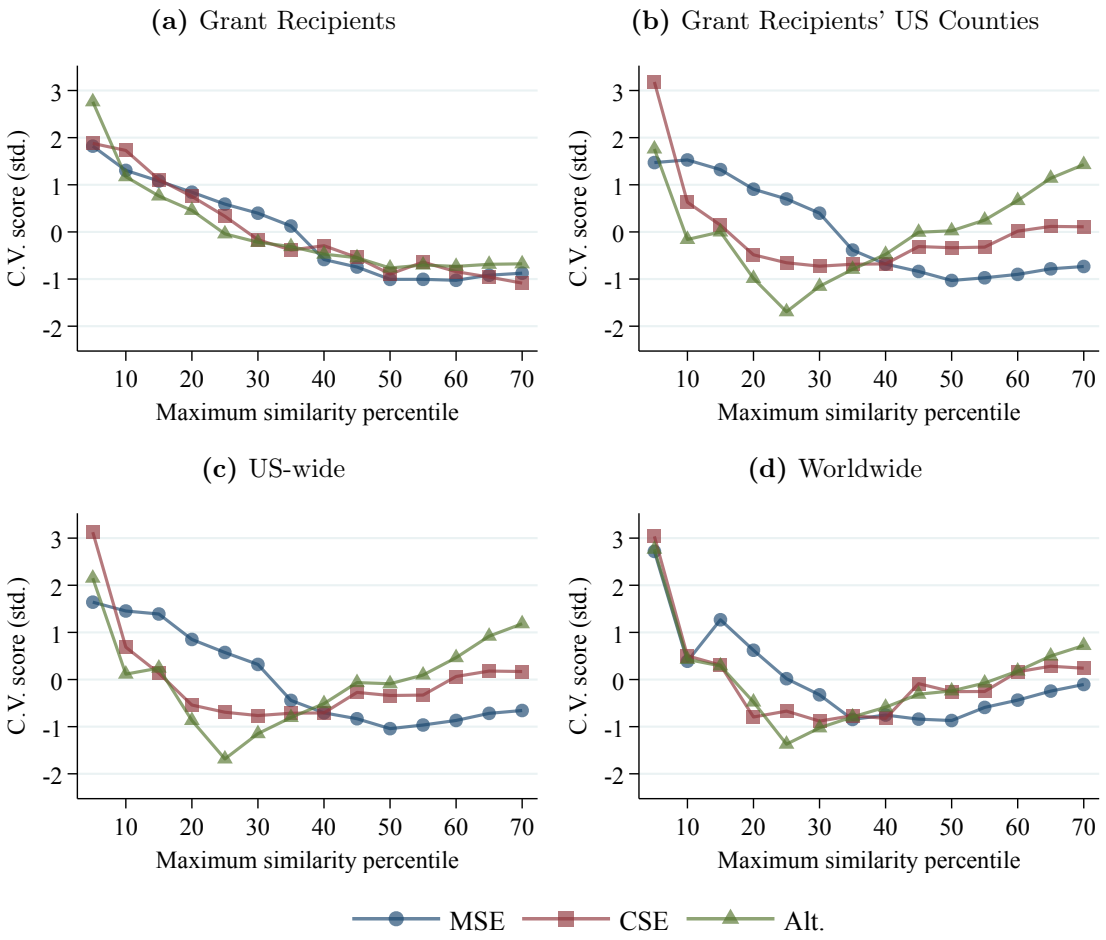
$$\text{Mean Chi-Squared Error (CSE): } \frac{1}{N}\sum_i \left((y_i - \widehat{y}_i)^2 / \widehat{y}_i \right) \ ,$$

$$\text{Mean Alternative Error (Alt.): } \frac{1}{N}\sum_i (-y_i \widetilde{y}_i + \widehat{y}_i) \ ,$$

where the "Alternative" function is the same function employed in the Stata© `lasso poisson` command.

Figure D.1 plots the results of the bandwidth searches. For the grant recipients, we observe a convergence to a minimum at approximately the 60^{th} percentile, and for the rest of the sets of patents we observe minimums at approximately the 40^{th} percentiles. Thus, we use these values for the threshold in our main specifications. Clearly, these minimums are not extremely sharp, and so we also report robustness tests that vary these thresholds but plus/minus 10 percentiles, but we obtain similar results in all cases.

Figure D.1: Cross Validation Technological Spillover Boundary Search



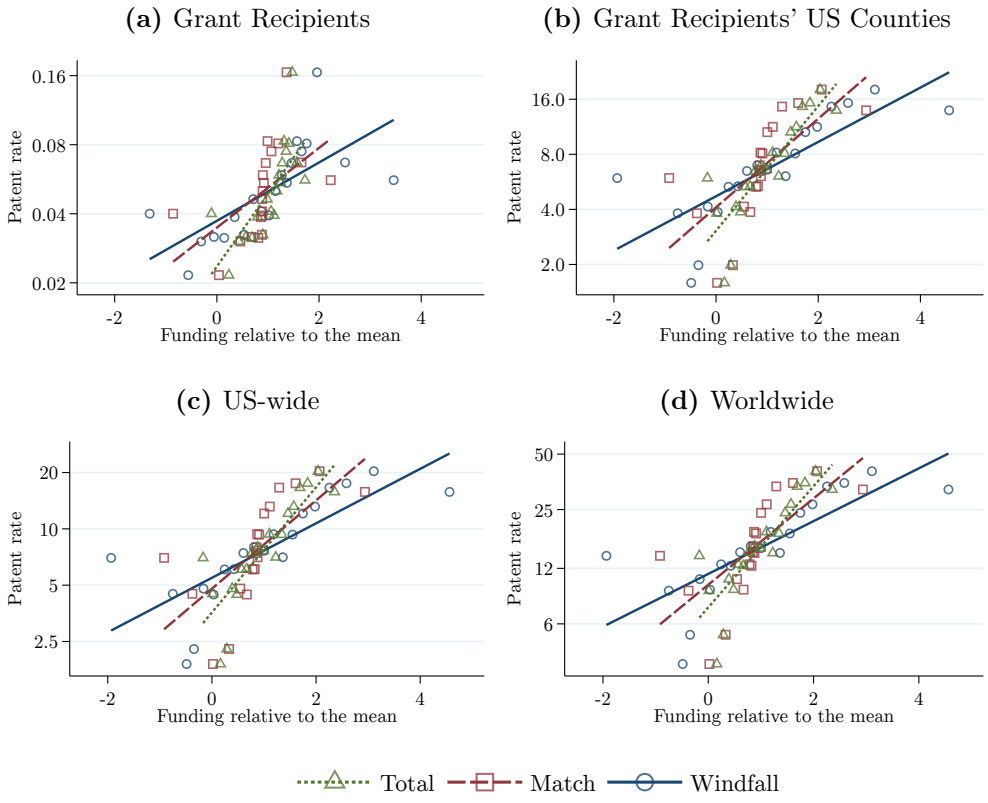
Notes: Plots the standardized cross validation scores from varying the maximum similarity percentile that determines the boundary of technological spillovers using three different penalty functions (MSE, CSE, Alternative). Each panel corresponds to a different set of patents used as the dependent variable.

E Additional Results, Specifications, & Robustness Tests

E.1 Binned Scatterplots

Figure E.1 contains binned scatter plots of the variation underlying Table 2 in the main text. The axes are scaled so that the linear fits mimic the constant elasticity assumption used throughout the paper. Except for outliers at the funding extremes, this constant elasticity assumption appears very reasonable. The downward shift in the slope of the fitted lines graphically depicts the pattern seen in Table 2 columns 1–3. We estimate smaller elasticities as we shift from using all of the funding variation (federal and state), to focusing on the state match variation, to finally focusing only on the windfall subset of the match-based variation.

Figure E.1: Patenting and Funding Conditional on Aggregate Time Trends



Notes: Plots the annual flow of patents within a CPC group (y-axis; log scale) as a function of the stock of SBIR funding in that CPC group (x-axis; scaled relative to the sample mean to approximate a log transformation) after conditioning out year fixed effects. The funding variation is always the same, but the set of firms and inventors whose patents are included in the dependent variable is different in each panel.

E.2 Alignment with Howell (2017)

When we focus only on the patents produced by grant recipients, our estimates suggest an average marginal cost of approximately $1.3 million per patent, with 95% confidence intervals spanning $1 million to $1.6 million. The estimates from Howell (2017) indicate that Phase I and Phase II grants lead to anywhere from 30%–80% more citation-weighted patents. Taking the lower bound of these estimates, and assuming that these magnitudes are similar for raw patent counts (which Howell 2017 does not report), which has a mean of 2.0 in that sample, would imply conservative average marginal costs per patent of $250,000 ($= 1/((2.0 \times 0.3)/\$150,000)$) for Phase I grants and $1.7 million ($= 1/((2.0 \times 0.3)/\$1,000,000)$) for Phase II grants. And since roughly one-third of total DoE SBIR dollars in that sample are awarded via Phase I grants, this suggests an average marginal cost of $1.2 million on a per-dollar basis. Using the less conservative estimates from Howell (2017) can yield a per-dollar cost closer to $750,000.

This close overlap supports our empirical approach. We are comfortable focusing on the most conservative estimates from Howell (2017) because (1) Howell (2017) focuses on just two of the more "applied" funding offices of the DoE (and we look at all offices, which may incorporate funding less likely to lead to patents), and (2) our estimate is an intent-to-treat, since we cannot observe the actual match-based funding awarded.

E.3 Robustness to Technology-specific Time Trends

Perhaps the largest threat to our identification strategy is that there are unobservable shocks correlated with both (1) which CPC groups receive windfall funding because of the state match policies, and (2) the supply of or demand for patents in those same CPC groups. This could arise if firms in states with match policies are more productive or tend to pursue technologies more in-demand, or vice versa in each case. Our findings in Appendix C.2 suggest this is likely not the case.

To further explore this possibility, we estimate a series of models of the form:

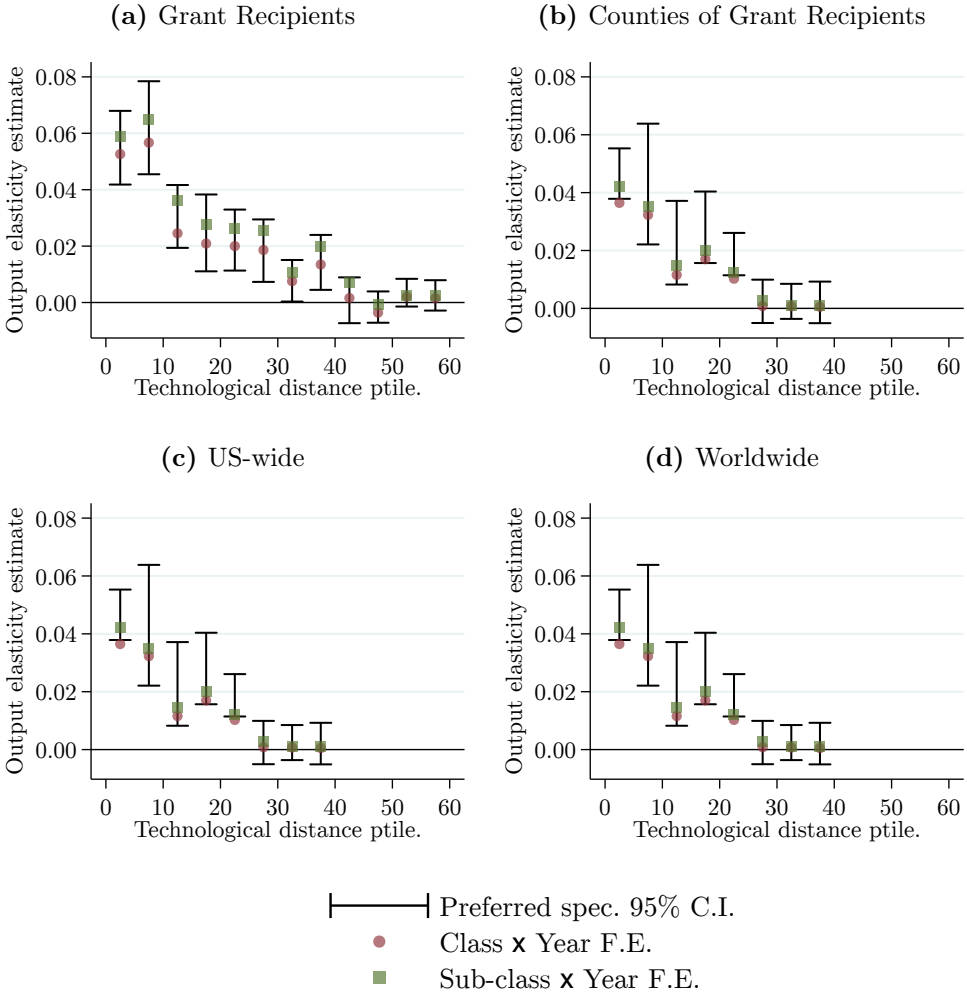
$$\mathbb{E}[Y_{jt}^{d} \,|\, W_{jtb}] = \exp\Big(\sum_{b \in \mathcal{B}} \frac{W_{jtb}}{\overline{W}} \theta_{b}^{d} + \tau_{g(j)t}^{d}\Big) \,, \tag{E.1}$$

which are identical to our main regression specification (Eq. 4), except now we allot the year fixed effects to be specific to different sets g of CPC groups (which are indexed by j). To construct these sets, we leverage the hierarchical nature of the CPC scheme and estimate two versions of equation E.1 aggregating the j-level CPC codes we use ("Main Groups") up either one level (to the level 3 "Sub-class") or two levels (to the "Class") in the hierarchy.

Removing this variation from the data with these fixed effects decreases the likelihood that our estimates are driven by any worrisome aggregate shocks that are unique to different sets of technologies. However, it also introduces the possibility of exacerbating measurement error in our independent variable and biasing our coefficients towards zero (Griliches and Mairesse 1995).

Figure E.2 plots the results of these regressions (along with the 95% confidence intervals from our preferred specification). We do tend to estimate coefficients closer to zero in many cases. But overall, the estimates from these models saturated with more fixed effects are broadly consistent with our main results. And given the large amount of variation in the data that these fixed effects remove, these results suggest that our identification strategy is not merely reflecting some unobservable trends that are covarying with patenting rates and the SBIR match policies

Figure E.2: Patent Output Elasticity Estimates with Technology-specific Time Trends



Notes: Plots the point estimates of $\theta^{d,b}$ from Eq. E.1 when using either group-time or sub-class-time fixed effects. The error bars plot the 95% confidence interval from our preferred specification (Eq. 4) based on standard errors clustered at the CPC group level.

E.4 Alternative Data Construction and Regression Specifications

Alternative Specifications

Figure E.3 reports estimates from a range of specifications that use alternative choices at the data construction, sample inclusion, or estimation stages. We present the results here only for our two geographic edge cases for brevity (only grant recipients in Figure E.3a and then all firms and inventors in Figure E.3b), though we find very similar patterns when focusing on other sets of firms and inventors. In all cases, we obtain point estimates and patterns in coefficients that are very similar to our preferred specification which gives us confidence that no single choice we make is driving our results in particular.

"FOA FE in sim calc." includes FOA fixed effects in the text similarity analyses, which removes variation across FOAs that may be erroneous (e.g., due to different writing styles of DoE program managers). However, it also imposes an assumption that the latent patentability of all FOAs is the same, which is an assumption we would rather not make. This is why our preferred specification estimates the percentiles of technological distances using the full distribution of raw similarity scores. Regardless, the estimates are very similar.

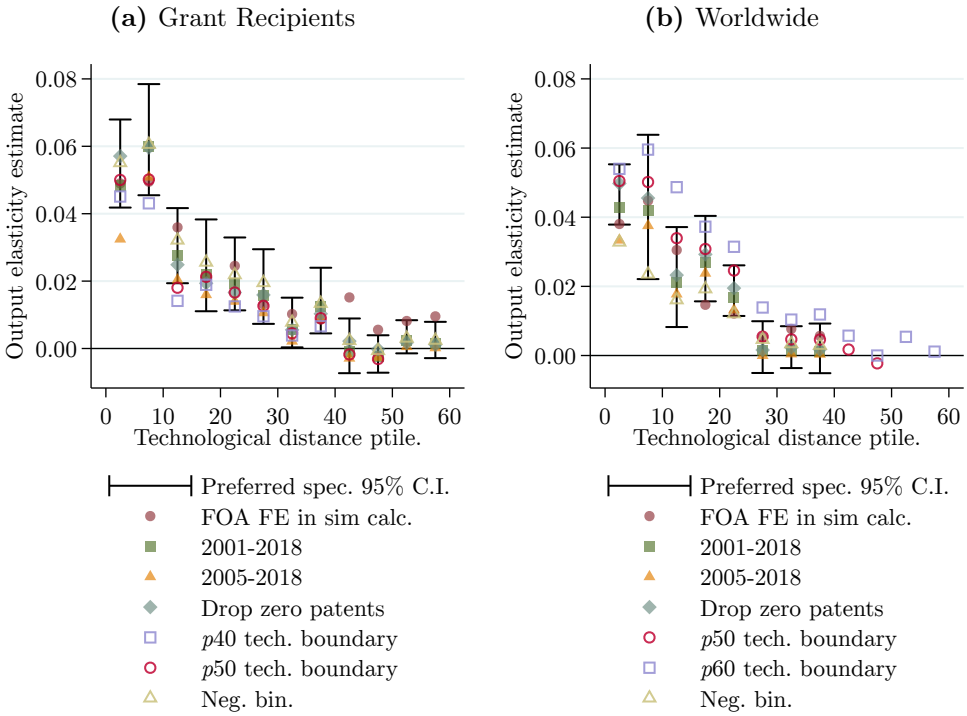
The results are very similar when we constrain the sample to exclude early years, which suggests that it is not important that the state matches were much rarer in the early years, or that we use some of these early years of data to construct the patent-FOA text similarity. Likewise, dropping observations with zero patent flows, making minor adjustments to the boundary of technological spillovers, or estimating the regression as a negative binomial model all yield very similar results.

Alternative Investment Discounting

The stock-flow knowledge production function model assumes that current investments have some (possibly discounted) ability to generate output in all future periods. We are limited in our ability to understand the specific lag structure of production, but we explore this timing issue partly by altering our assumption about the discount value of the R&D stock.

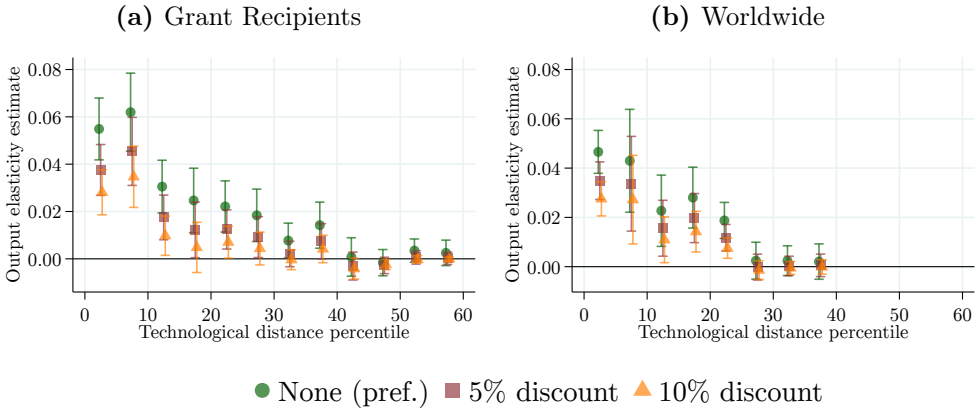
Figure E.4 plots the estimated coefficients from two regressions where we use a non-zero discount rate on the stock of R&D investments in our production function: 5% and 10%. If production is relatively short-term, this discounting should not alter our estimates since the meaningful variation in R&D stocks will not be compressed by the discounting. But we do see the estimates from these models tend to move closer to zero by a reasonable amount – roughly 30–50% for some point estimates. With discount rates of 5–10%, it would take roughly five to seven years to reduce this much variation in a stock, which we take as

Figure E.3: Patent Output Elasticity Estimates, Alternative Specifications



Notes: See text for descriptions of alternative specifications.

Figure E.4: Alternative Investment Discounting, Coefficients



Notes: Plots the point estimates and standard errors (clustered at the CPC group level) under alternative assumptions for the discount rate of the funding stock.

suggestive evidence that the most common production lags – the time from DoE investment to the appearance of new patents – are on this scale.

Alternative CPC Group Division

Our preferred approach to accounting for the fact that patents receive multiple CPC codes is to divide each code proportionally based on how many times it appears on the patent (e.g., if a patent has codes C01P2004/61, C01P2004/80, and C10L1/16, then we would assign 2/3 to CPC group C01P2004 and 1/3 to C10L1). This approach ensures that when we sum up these fractions over CPC groups, the resulting number (which serves as our dependent variable) corresponds to a count of "patent's worth" of CPC codes.

An alternative approach would be to give all CPC groups that appear on each patent a value of one (e.g., if a patent has codes C01P2004/61, C01P2004/80, and C10L1/16, then we could assign 1 to CPC group C01P2004 and 1 to C10L1). While simple, this approach yields a number that is much more difficult to interpret and benchmark. However, for the purposes of estimating spillovers, we cannot hypothesize why it would not be the case that both approaches yield similar magnitudes of spillovers.

Table E.1 reports the results from this alternative approach of assigning a value of one for all CPC groups listed on each patent (Panel b) compared to our preferred approach (Panel a). It is difficult to compare the elasticities across these two approaches within the same set of patents (i.e., within columns in the Table), because of the differences in the dependent variables. But most importantly, the magnitude of spillovers (here across geographic distances) is very similar in both cases. For example, grant recipients appear to account for roughly 18% (=0.54/2.97) of the net marginal product per our preferred approach, and in the alternative approach they account for roughly 21% (=2.06/9.63). Thus, it does not appear that the spillover magnitudes we are estimating are driven by our approach to handling CPC group assignment.

Table E.1: Alternative CPC Group Approach

	Grant recipients (1)	Recipients' counties (2)	US-wide (3)	Worldwide (4)
Panel (a): CPC Group Share Division				
Windfall $	0.134	0.125	0.123	0.130
	(0.021)	(0.016)	(0.015)	(0.014)
$\frac{\partial \text{ patent}}{\partial \text{ \$1M}}$	0.54	1.40	1.73	2.97
	[0.5,0.6]	[1.2,1.6]	[1.5,1.9]	[2.5,3.3]
Panel (b): Any CPC Group Flag = 1				
Windfall $	0.063	0.080	0.077	0.081
	(0.016)	(0.011)	(0.011)	(0.010)
$\frac{\partial \text{ CPC flag}}{\partial \text{ \$1M}}$	2.06	5.12	6.10	9.63
	[1.9,2.3]	[4.4,6.0]	[5.2,7.2]	[8.0,11.4]
N obs.	235,406	235,384	235,384	235,384
Tech. boundary	$p60$	$p40$	$p40$	$p40$
Year F.E.	Y	Y	Y	Y

Notes: Reports the output elasticity estimates from regressions using the "simple" model that aggregates all technological spillovers into a single bin per either the preferred approach of dividing patents amongst CPC groups equally (Panel a, which replicates Table 2 in the main text) versus assigning a count of one for all CPC groups listed on each patent (Panel b). Standard errors clustered at the CPC group level are reported in parentheses.

E.5 Additional Results: Paper Trails and Conduits

Comparison to Citation-based Approaches

The vast majority of empirical economic research on R&D spillovers, much of which stems from early influential papers such as Jaffe et al. (1993), has used front-page patent-to-patent citations as a proxy or evidence of a spillover. As noted in the main text, while this data has proven useful in many regards, it is not without serious limitations as noted by, at least, Alcácer et al. (2009), Arora et al. (2018) and Bryan et al. (2020).

Our empirical approach does not rely on these citations to identify spillovers. But we can impose this assumption to explore the extent to which the R&D spillovers we identify may be reflected in the paper trail. We do this by estimating another regression using our main specification (Eq. 4), but only include patents in the dependent variable that are connected to patents from DoE SBIR grant recipients through a direct citation (which we term the "1^o" approach) or through any possibly combination of citation links (which we term the "All^o" approach).

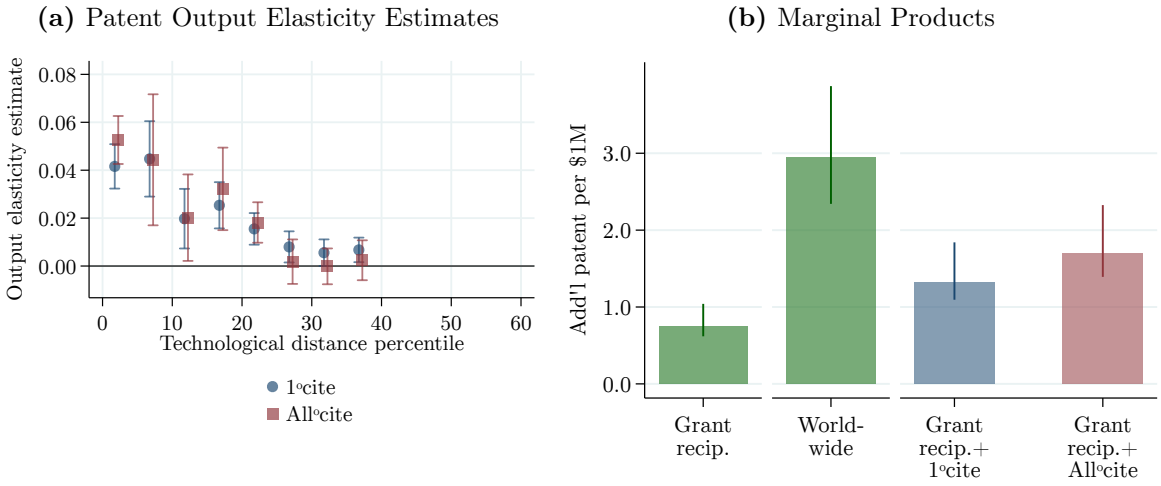
Figure E.5a plots the coefficients from these regressions, which behave very similarly to what we estimate when we use the universe of USPTO patents as the dependent variable. But as seen in Figure E.5b, relying on the 1^o or even the All^o approach captures at most only about 50% of the net output we observe from using the universe of USPTO patents. In other words, it appears that roughly half of the spillovers we identify are not reflected in citation linkages. Importantly, we cannot test what share of citation linkages reflect spillovers. Still, these results suggest that our approach to capturing R&D spillovers in the patent record could continue to prove useful as it may be much more flexible.

Individuals versus Firms as Conduits of R&D Spillovers

In our preferred specifications, we assign the geographic location of a patent to be equally representative of the locations of the individual inventors and firm assignees on the patent – if a patent has one inventor in one location and one firm assignee in another location, each receives one half of a patent in terms of the dependent variable. At the boundary cases of grant recipients or the entire universe of USPTO patents, this distinction is irrelevant. But for all interior cases, this choice of how to divide patents across locations may be relevant for understanding how R&D spillovers permeate geographic space.

To explore this further, we estimated the main regression specification again, assigning the geographic location to be entirely based on either the inventors' locations or the firm assignees' locations. Figure E.6 plots the resulting distribution of marginal products across

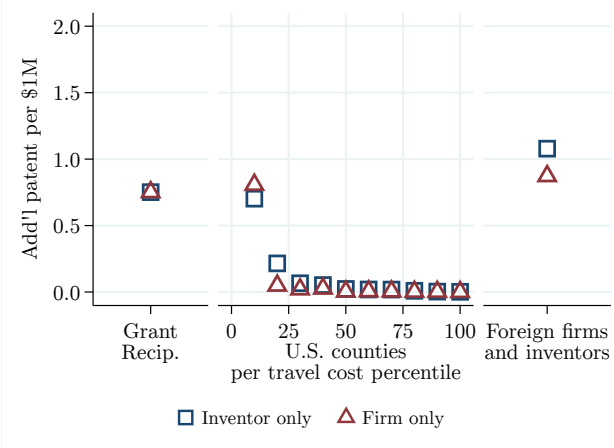
Figure E.5: Comparison to Citation Linkages



Notes: Plots the coefficient estimates from using the two alternative citation-based approaches to capture spillovers (Panel E.5a) and the implied marginal products from our preferred approaches (shown for grant recipients and the worldwide inventor sets in green) compared to these citation-based approaches (Panel E.5b).

geographic space (incorporating all technological spillovers) and shows that the specifics of how we assign geographic location does not make a large difference in our results.

Figure E.6: Alternative Inventor/Firm Attributions



Notes: Plots the average marginal product when attributing patents' geographic location entirely to either inventors or firm-assignees (where the preferred specification makes an equal attribution).

This consistency is likely driven in part by the fact that geographic locations of firms and inventors are relatively correlated: 83% of all inventor-assignee pairs are from the same country, and amongst pairs where one is located in the US, 50% are located in the same state and 31% are located in the same county. But it may also reflect the fact that both

firms and inventors are equally important conduits of R&D spillovers across geographic space. There are some minor differences in the shapes of the within-US spillovers and the level of international spillovers. It appears that firms may be slightly more important in facilitating spillovers over very short geographic distances and individuals may be slightly more important for facilitating international spillovers.

F The Explore-Exploit Index and other Regional Correlates of R&D Spillovers

Studies have explored the role of forces such as trade patterns and market sizes (Eaton and Kortum 2002), foreign direct investment (Branstetter 2006), and multinationals (Griffith et al. 2006) in facilitating R&D spillovers. Still, inference in these settings is often either indirect or based on aggregate patterns. Here, we report the results of a purely descriptive search for the correlates of R&D spillovers. We cannot make any causal statements, but our setting and data provide a unique opportunity to (1) estimate how specific regions (e.g., US counties, foreign countries) are more or less likely to benefit from spillovers, and then (2) regress these estimates on features of each region to explore what is more or less correlated with the level of spillovers into that region.

In addition to features motivated by prior work, we focus on the degree to which a region appears to *exploit* knowledge produced by SBIR firms and focus on technologically similar inventions versus using that knowledge to *explore* technological space and focus on technologically distant inventions. We term this feature simply the "exploit-explore index" of a region. The notions of exploitation and exploration are pervasive in innovation economics and typically posed as a tradeoff (e.g., Manso 2011). But the importance of these alternative strategies, and whether such a tradeoff exists, has not received much attention in the context of R&D spillovers.

There are good reasons that exploit-oriented regions may be responsible for a large fraction of spillovers: the R&D performed by SBIR firms may de-risk ideas surrounding only the particular technologies they pursue (Howell 2017); the new patents obtained by SBIR firms may be signals that draw investors attention to those particular technologies (Conti et al. 2013); the nature of absorptive capacity (Cohen and Levinthal 1990) may lead only firms who focus on the particular technologies funded by the DoE to benefit from these advances (Aghion and Jaravel 2015); and the large adjustment costs of modern science (Myers 2020) might constrain inventors from taking these new ideas and applying them to more distant technologies.

Conversely, there are also good reasons why explore-oriented regions may be more likely to generate the spillovers we observe. First, the theoretical motivation for these subsidies should steer the DoE to target technologies where appropriating value – obtaining a patent – is difficult. So, perhaps once the SBIR firms are successful in these technological areas, they can be used in other areas where patenting incentives are relatively higher. Furthermore, recent work by Acemoglu et al. (2020) provides evidence that when firms are more "creative"

or "open to disruption" they are more likely to pursue new lines of research and develop more radical patents. This could suggest that when regions are more willing to take an idea developed by an SBIR firm and use it in a novel way, they will ultimately be more productive and produce more patents.

Region-specific Spillover Levels

First, to obtain region-specific estimates of the relative spillovers into that region, we estimate a series of regressions that use our preferred specification:

$$\mathbb{E}[Y_{jt}^r \,|\, W_{jtb}] = \exp\Big(\sum_{b\in\mathcal{B}} \frac{W_{jtb}}{\overline{W}}\theta_b^r + \tau_t^r\Big)\ , \tag{F.1}$$

which we estimate for each region r – either BEA-defined economic areas or foreign countries – which yields many estimates of the $\theta^{b,r}$ parameters. When focusing on domestic spillovers, Y_{jt}^r is based on the flow of patents from only firms or inventors in economic area r except for any patents from DoE SBIR grant recipients who might also be located in r. When focusing on international spillovers, Y_{jt}^r is based on the flow of patents from only firms or inventors in foreign country r, which by construction excludes patents from DoE SBIR grant recipients.

For each of these regressions, we then calculate the relative amount of spillovers into that region by simply summing up the $\theta^{b,r}$ estimates – larger elasticities means larger spillovers, and for this purely vertical metric, we are not concerned whether these spillovers are at low or high technological distances. To account for the uncertainty in these estimates (since we use these numbers as the outcome variable in the prediction exercise), we use the popular empirical Bayes procedure to shrink our estimates. Thus, our final estimate of the relative level of spillovers into each region r is:

$$\text{Relative Spillover Level}_r \equiv \underbrace{\sum_b \widetilde{\theta^{b,r}}}_{\substack{\text{sum of post-}\\ \text{shrinkage estimates}}}\ ,$$

where $\widetilde{\theta^{b,r}}$ is the post-shrinkage estimate of the parameter. We focus on the elasticities (and not the absolute levels of patent spillovers) because we do not want to introduce a mechanical connection between the baseline patenting levels in each region and the degree to which we think that region is benefitting from spillovers.

The "Exploit-Explore" Index

As motivated in the main text, one of the features we think might be important for influencing spillover levels is each region's propensity to use the knowledge created by SBIR firms to either exploit or explore technology space. An exploitation tendency involves the development of patents that tend to be more similar to the R&D performed by SBIR firms, and vice versa for regions with an exploration tendency. In other words, we can proxy for a regions' stance on this exploit-explore index by measuring the share of patents produced by spillovers in that region that are more similar (low tech. distance) or less similar (high tech. distance) from the DoE's objectives.

Figure F.1a illustrates how we use our $\theta^{b,r}$ estimates to calculate this proxy. We fit a separate line through each of the r-specific set of $\theta^{b,r}$ estimates (using the post-shrinkage estimates), the slope of which describes how the relative flow of patents changes with each increase in the technology distance percentile:

$$\text{Exploit-Explore}_r \equiv \underbrace{\frac{\partial \widetilde{\theta^{b,r}}}{\partial b}}_{\substack{\text{slope of linear fit of} \\ \text{post-shrinkage estimates}}},$$

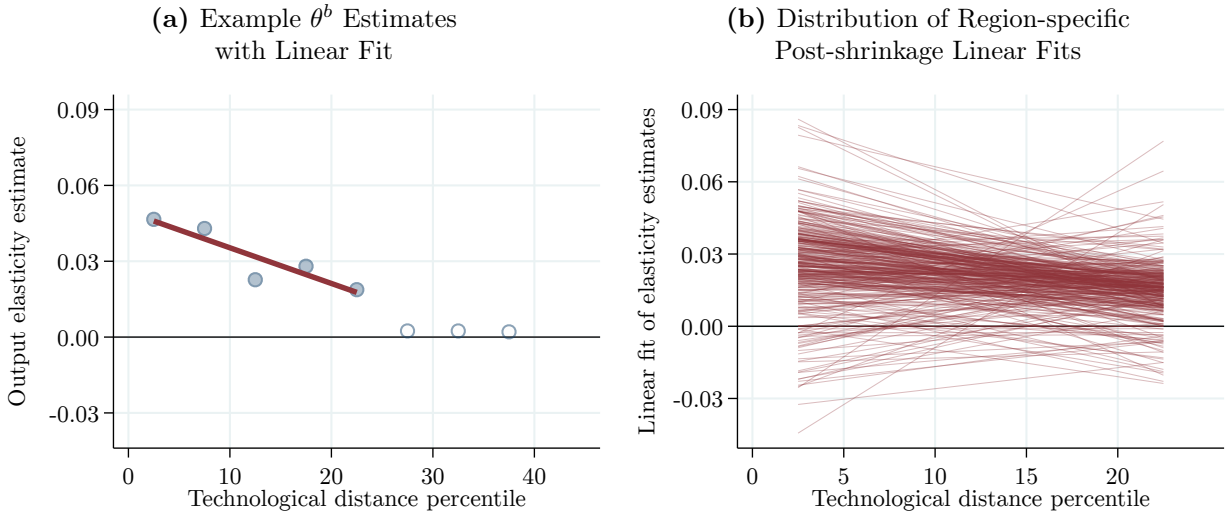
where more positive values indicate a more explorative orientation.

To see how this captures the notion of exploitation versus exploration, consider a region where we estimate equal $\theta^{b,r}$ parameters for all b – the slope of the fitted line is zero. This implies that the spillovers into that region led to an increase in patents evenly across all feasible areas of technology space. This is likely to happen only if the firms and inventors in that region are willing and able to explore technology space. Conversely, consider a region where we estimate a non-zero $\theta^{b,r}$ at the closest technological distance – the slope of the fitted line is very negative. This implies that the spillovers into that region led to an increase in patents only in the same areas that the DoE targeted grants towards. This is likely to happen only if the firms and inventors in that region avoid (or are ineffective at) exploring technology space and instead prefer to exploit.

It was apparent that, in virtually all cases, we estimate very small elasticities for distances beyond the 25th percentile of technological distance. Thus, we use only the estimates from within this technological distance boundary for all of the analyses in this section. This helps minimize a purely mechanical relationship that would arise between our measure of relative spillovers and the exploit-explore proxy.[7]

[7]We also tested other versions of this measure based on the share of patent spillovers that are within the

Figure F.1: Estimates of Region-specific Spillovers and the Explore-Exploit Index



Notes: Figure F.1a plots an example set of output elasticity (θ^b) estimates, one for each of the eight values of b (here, recreating Figure 2d) along with a linear fit of these point estimates. Figure F.1b plots the actual distribution of these linear fit aligns based on the region-specific estimates using within-US economic areas (domestic spillovers) and non-US countries (international spillovers).

Figure F.1b plots the distribution of the fitted lines across all regions, the slope of which is our Exploit-Explore index. There is clearly variation across regions in terms of their tendency to produce more exploitative or more explorative patents via spillovers.

Other Regional Features

For the set of possible correlates in addition to the exploit-explore index, we collect a large amount of data on regions from various sources (see Appendix B.5 for more). We focus on features that have been suggested to be relevant by prior work (e.g., firm sizes, trade volumes and compositions, venture capital levels, foreign direct investment, proximity to universities and federal labs, etc.).

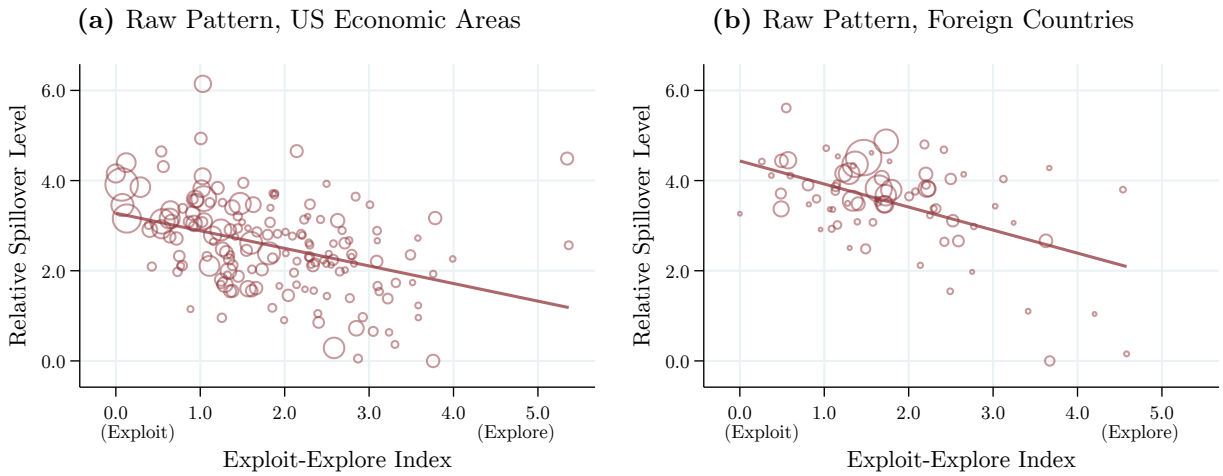
F.1 The Exploit-Explore Index is Unique and Important

To begin, Figure F.2 plots regions' exploit-explore index (x-axis) and how it covaries with the relative spillover levels in that region (y-axis). For comparability, the spillover levels and exploit-explore metrics are normalized within either set of regions so that their minimum is zero and a one unit change equals one standard deviation. Our focus is on relative differences, not absolute values. In both cases of within-US (Fig. F.2a) and international (Fig. F.2b)

5th or 10th percentile of technological distances – more discrete approaches to estimating the slope of the coefficients – and obtained very similar results.

spillovers, there is a clear correlation – regions with more of an exploit orientation tend to have relatively larger amounts of spillovers. At both domestic and international levels, regions that are one standard deviation closer to the exploit end of the spectrum have relative spillover levels that are roughly one half of a standard deviation larger on average.

Figure F.2: Regions that Tend to Exploit also Tend to Produce More Spillovers



Notes: Panels F.2a–F.2b are scatterplots and linear fits of the total amount of spillovers a region is responsible for as a function of their exploit-explore index (defined in text), with the markers weighted by the annual average total USPTO patent output of the region. Excludes regions with extremely small patent output. Both metrics are normalized within either set of regions (US economic areas or foreign countries) so that the minimum is zero and a one unit change equals one standard deviation.

Next, we explore other correlates of R&D spillovers and test whether the tendency for more exploitative regions to produce relatively more spillovers is a unique pattern or simply reflects other underlying correlations. We use the stability selection procedure proposed by Meinshausen and Bühlmann (2010). This amounts to performing a series of bootstrap sampling (we use 100 iterations), where a random half of the sample data is retained and a Lasso is performed on the subsample. The "Importance" score is then the share of subsamples in which a variable is selected as relevant by the Lasso. We supplement this process by also calculating the partial R^2 of each variable selected within an iteration (via OLS) and then calculate what we term the "effective" partial R^2 by taking the average of the partial R^2 values for each variable across all iterations, where the partial R^2 is zero if the variable is not selected by the Lasso.

The left-hand columns of Tables F.1 and F.2 report the results from this exercise. In short, we find many of the features related to the supply of and demand for energy technologies to be relevant, but few features consistently explain more variation in a region's ability to capitalize on spillovers than their exploit-explore index. Even after conditioning on large vectors of

other relevant controls, regions that appear more willing to focus on the same technologies that the SBIR firms pursued are more likely to create more patents. In the case of domestic spillovers, the index is the fourth most important feature (out of 50 possible), with only industry clusters in IT, production technologies, and oil and gas appearing more important. At the international level, the index is found to be the most important feature.[8]

It does not appear to be the case that this exploration orientation is simply a proxy for other economic fundamentals of these regions. First, the variable itself is selected as important by the stability selection routine, which includes the large vector of features that should provide direct or proxy measures of many fundamentals. Second, we also perform a series of stability selection routines where we treat the exploit-explore index as the outcome, and we find that it is more difficult to predict this feature (using all other features) than it is to predict relative spillover levels. The right-hand columns of Tables F.1 and F.2 report the results from the same stability selection exercise, this time treating the index as the outcome to predict. In both domestic and international cases, it appears more difficult to predict this index than it is to predict the relative spillover levels. We take this as evidence that this index is not simply reflecting other fundamentals and may in fact capture a unique way in which firms and inventors differ across regions.

We emphasize again that the results here need not reflect causal effects – some features highly correlated with spillovers may simply reflect other unobservable differences across regions. Still, the apparently large importance a region's exploitation orientation highlights something that, to our knowledge, has received little attention to date.[9]

This finding is very much in line with Aghion and Jaravel (2015), who discuss the implications of absorptive capacity – the exploitation of knowledge created by others (Cohen and Levinthal 1990) – for growth models and other dimensions related to R&D spillovers. In the context of absorptive capacity, the term "exploit" has often been used generally to refer to the process of extracting value from other's advances. And the complementarities between firms' R&D investments are often considered only in the sense of the rate of R&D. Our results in this section indicate that this exploitation may also involve an important notion of

[8]Other important correlates of domestic spillovers include the number of workers with advanced scientific degrees, distance from SBIR grantees, and GDP per capita. Other important correlates of international spillovers include features related to natural resource rents, pollution, and clean energy, which all suggest that the supply of and demand for specific types of energy in different countries might be playing an important role in incentivizing foreign inventors to make use of this DoE-funded R&D.

[9]Our results are not at odds with Acemoglu et al. (2020), who focus more on the probability of observing radical patents (whereas we focus on total patent counts) and whose results might reflect factors important in de novo R&D (whereas our focus on spillovers likely ties us to R&D that is more cumulative or sequential in nature).

the direction of R&D, since we find regions more likely to stay focused on the technologies where initial advances are made (exploiters) tend to create more new patents than those who appear to use those advances elsewhere in technology space (explorers). Altogether, our results suggest a need continued work on why different groups of innovators are more likely to exploit or explore technology space, especially after they observe others generating new knowledge, and how this influences the rate and direction of invention writ large.

Table F.1: Predicting Spillovers & Exploration Orientation across 174 US Economic Areas

Panel (a): Full Sample Selection

Net Spillovers		Exploration Orientation	
N Features Selected	26 / 50	N Features Selected	13 / 49
R^2	0.52	R^2	0.41

Panel (b): Stability Selection

Net Spillovers			Exploration Orientation		
Feature	Importance $[0,1]$	Effective partial R^2	Feature	Importance $[0,1]$	Effective partial R^2
Cluster: IT	1.00	0.19	Cluster: Oil & Gas	0.98	0.09
Cluster: Prod. Tech.	0.94	0.07	Cluster: IT	0.98	0.06
Cluster: Oil & Gas	0.80	0.03	Wages	0.91	0.02
Exploit-Explore index	0.77	0.05	Travel cost	0.88	0.02
Scientific workforce	0.71	0.03	Research Nuc. Reactor	0.86	0.02
Dist. from SBIR grantees	0.70	0.01	Cluster: Prod. Tech.	0.84	0.02
GDP per capita	0.64	0.05	Internat. migration	0.77	0.02
Cluster: Construction	0.60	0.02	FFRDC, any	0.72	0.02
Cluster: Metal mining	0.58	0.04	Power Nuc. Reactor	0.65	0.02
Cluster: Coal mining	0.50	0.02	Cluster: Lighting	0.64	0.03
Cluster: Auto.	0.47	0.01	Cluster: Enviro.	0.58	0.02
Firm sizes	0.39	0.01	University count	0.54	0.01
Cluster: eComm. & Distr.	0.34	0.01	Young adult share	0.47	0.00
Labor force productivity	0.33	0.01	Labor mobilization	0.46	0.01
Exports	0.29	0.01	Cluster: Elec. power	0.45	0.01

Notes: Panel (a) reports the number of features selected and the R^2 of the resulting regression when the Lasso is applied to the full sample. Panel (b) reports the importance of each feature as the share of 100 bootstrap samples where the feature is selected by the Lasso, as well as the effective partial R^2 which is the the average partial R^2 across all bootstrap subsamples, by construction set to zero when the feature is not selected. All results based on 174 observations of US economic areas.

Table F.2: Predicting Spillovers & Exploration Orientation across 207 Foreign Countries

Panel (**a**): Full Sample Selection

Net Spillovers		Exploration Orientation	
N Features Selected	27 / 83	N Features Selected	3 / 82
R^2	0.97	R^2	0.27

Panel (**b**): Stability Selection

Net Spillovers			Exploration Orientation		
Feature	Importance $[0,1]$	Effective partial R^2	Feature	Importance $[0,1]$	Effective partial R^2
Exploit-Explore index	0.52	0.26	Nat. reso. rents, %GDP	0.48	0.24
Nat. reso. rents, %GDP	0.43	0.27	Renew. elec. output	0.38	0.09
Labor force partic.	0.37	0.20	Export to US: Nuclear	0.34	0.14
NO2 emiss., energy	0.27	0.12	GDP per capita, growth	0.33	0.09
Renew. elec. output	0.24	0.07	Internet use rate	0.32	0.10
Renew. energy consump.	0.23	0.06	Pop. density	0.27	0.12
Methane emiss., energy	0.14	0.06	Internet serv. per capita	0.25	0.09
English speaking	0.14	0.05	GDP growth	0.20	0.04
GDP per capita	0.11	0.03	English speaking	0.18	0.04
In-US migrant stock	0.10	0.02	NO2 emiss., energy	0.16	0.08
Inflation rate	0.09	0.02	Pop. growth	0.16	0.04
Pop. growth	0.09	0.02	Unemployment rate	0.14	0.04
Unemployment rate	0.08	0.03	Export to US: Biotech.	0.13	0.05
Export to US: Biotech.	0.07	0.03	Labor force partic.	0.13	0.04
Avg. USPTO patent rate	0.07	0.02	FDI: Chemical Sector	0.12	0.03

Notes: Panel (a) reports the number of features selected and the R^2 of the resulting regression when the Lasso is applied to the full sample. Panel (b) reports the importance of each feature as the share of 100 bootstrap samples where the feature is selected by the Lasso, as well as the effective partial R^2 which is the the average partial R^2 across all bootstrap subsamples, by construction set to zero when the feature is not selected. Imports and exports are based on bilateral trade with the US.

G Quality Effects and Patent Value Capture

G.1 Isolating Quality-margin Effects

The following analyses and findings center on determining (1) whether SBIR funding effect patent quality, and then (2) whether we can make any reasonable inference about the share of patent-based value that accrues to different sets of firms and inventors. Answering the first question is necessary to answer the second, and the answer to the second can start to shed light on the size of any gap between the private and social returns to R&D. We use forward citations as our focal proxy for patent quality because of its clear association with the private value of patents (Kogan et al. 2017).

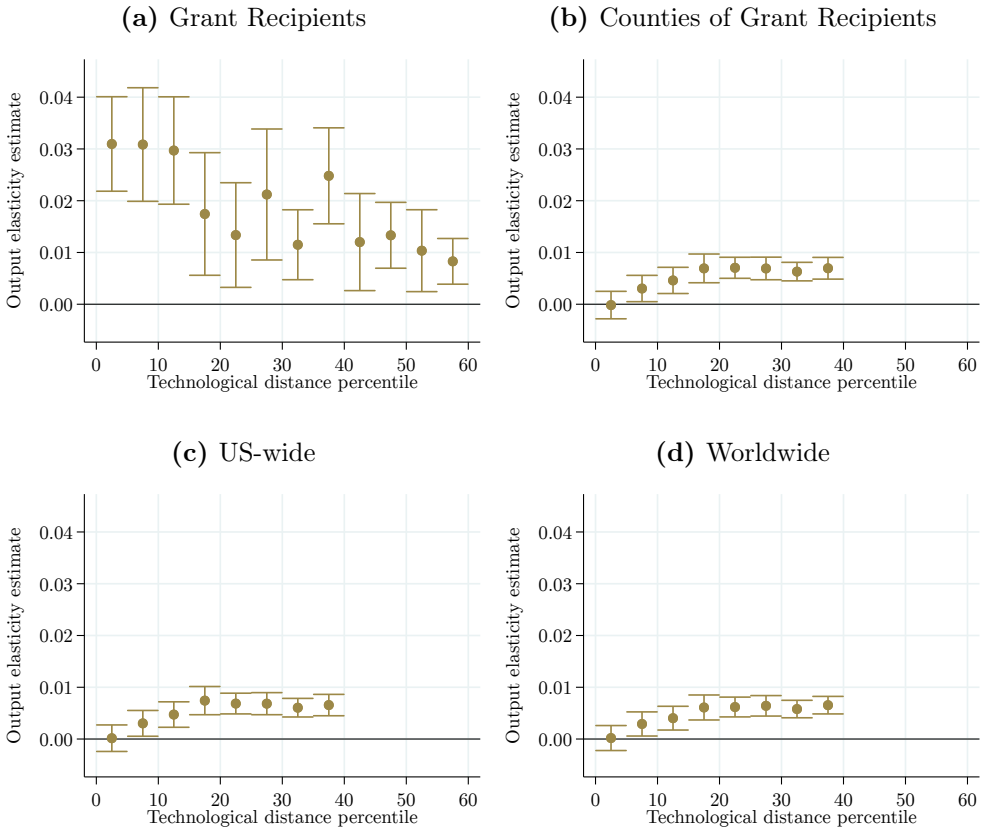
First, we estimate our main regressions again, but replace the dependent variable to be the number of citations-per-patent (instead of raw patent counts) within each CPC level observation. Figure G.1 plots the point estimates from these regressions, which clearly indicate that SBIR funding increases the citations per patent, but only in a meaningful way when we focus only on the patents of grant recipients.

Next, we want to approximate what share of the total patent-based value generated by the SBIR grants is captured by (patents obtained by) the grant recipients. The results reported in the main text indicate that grant recipients are responsible for only about 25% of the net patent output they stimulate. Thus, if we ignore any potential differences in the quality of these new patents and make the most generous assumptions possible about other important dimensions, this can imply that, as a lower bound, grant recipients also capture only about 25% of the net patent value the create following the SBIR grant.[10] This order of magnitude is in line with the few firm-level (Bloom et al. 2013) and macroeconomic studies (Jones and Williams 1998) that estimate this fraction to be on the scale of 25–50%.

However, that we observe increases on the quality margin only for grant recipients suggests that this estimate may be too low. Depending on the private value associated with forward citations, it may be the case that the private value of the grant recipients' patents is larger than the spillover-based patents that other firms and inventors obtain. As shown in Figure G.2, these quality effects are markedly different, so it may be that grant recipients capture much more than 25% of the net patent value. In the main text, we show how incorporating these quality effects alters the implied share of patent value captured by different sets of

[10] This lower bound is also based on the following assumptions: (1) spillovers are entirely driven by pure productivity shocks such that non-SBIR firms are *not* increasing private investments (which would reduce the true social returns); (2) product market rivalry spillovers after the patenting stage are negligible; and (3) none of the spillovers are due to market transactions where grant recipients are compensated (e.g., via patent licenses).

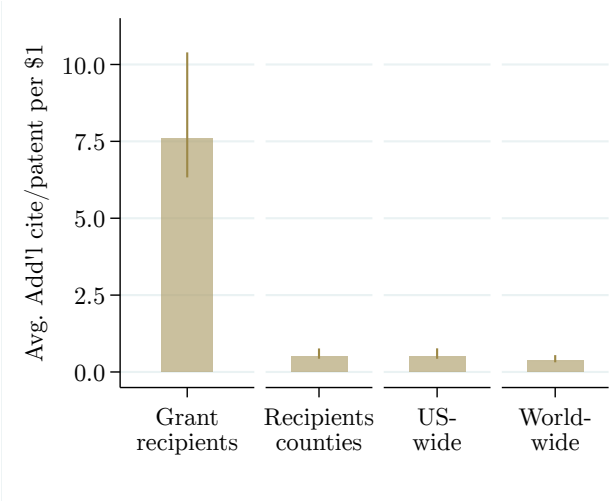
Figure G.1: Cite-per-Patent Output Elasticity Estimates



Notes: Plots the point estimates and 95% C.I. (standard errors clustered at the CPC group level) from estimating the main regression using forward citations-per-patent as the dependent variable.

firms and inventors.

Figure G.2: Average Level Changes in Cite-per-Patent



Notes: Plots the average marginal increase in patent quality associated with $1 million when examining four different aggregations of patents, with error bars indicating the 5th-95th percentiles across FOA topics.

G.2 Combined Quality-and-quantity-margin Effects

In the preceding analyses, we isolated effects on the quality of patents (as proxied with forward citations). Another popular approach in similar studies is to jointly estimate the effect of R&D investments on both the quality and quantity margins by using citation-weighted patent counts. Typically, a patent is weighted by the count of forward citations it has received to date, also possibly adding a value of one to that count. The former approach implies that the marginal citations (holding patent counts fixed) are associated with infinitely more value compared to marginal patents (holding citation counts fixed), while the latter implies that marginal patents and citations are associated with equal value. Recall, the empirical evidence to date suggests that marginal patents are associated with anywhere from three to twenty times as much value (to the recipient firm) than marginal citations.

While we prefer our approach of independently investigating the quality margin (by focusing on the change in the citations-per-patent), we also investigated the implied magnitude of spillovers using these other popular approaches of citation weighting.

Table G.1 reports the results from estimating our "simple" model that focuses only on geographic spillovers and uses a single bin of investments that spans the entire boundary of technological spillovers. Panel (a) recreates the four rightmost columns of Table 2 in the main text for comparison.

Overall, and in line with our results in the previous sub-section, we find that spillovers are smaller if we put a larger implicit weight on citations. Panel (b), which only values citations, indicates that grant recipients are responsible for about 75% (=6.77/8.97) of net citation output from these investments. Panel (c), which values citations and patents equally, indicates that grant recipients are responsible for about 38% (=3.78/10.05) of net patent-plus-citation output from these investments. These magnitudes largely overlap with the range of magnitudes we report in Figure 4 in the main text. To summarize, it appears that unless one places an extremely large value on citations relative to patents, the spillovers from these R&D grants are economically meaningful.

Table G.1: Alternative Approaches to Citations

	Grant recipients (1)	Recipients' counties (2)	US-wide (3)	Worldwide (4)
Panel (a): No citation weights				
Windfall $	0.134	0.125	0.123	0.130
	(0.021)	(0.016)	(0.015)	(0.014)
$\frac{\partial \text{ patent}}{\partial \text{ \$1M}}$	0.54	1.40	1.73	2.97
	[0.5,0.6]	[1.2,1.6]	[1.5,1.9]	[2.5,3.3]
Panel (b): Citation weight = forward citations				
Windfall $	0.552	0.320	0.316	0.322
	(0.049)	(0.027)	(0.026)	(0.023)
$\frac{\partial \text{ citation}}{\partial \text{ \$1M}}$	6.77	5.60	6.78	8.97
	[6.6,7.2]	[4.9,6.4]	[5.9,7.6]	[7.8,10.1]
Panel (c): Citation weight = 1 + forward citations				
Windfall $	0.469	0.284	0.280	0.274
	(0.041)	(0.023)	(0.022)	(0.018)
$\frac{\partial \text{ patent+citation}}{\partial \text{ \$1M}}$	3.78	5.58	6.83	10.05
	[3.7,4.1]	[4.8,6.3]	[5.8,7.7]	[8.5,11.4]
N obs.	235,406	235,384	235,384	235,384
Tech. boundary	$p60$	$p40$	$p40$	$p40$
Year F.E.	Y	Y	Y	Y

Notes: Reports the output elasticity estimates from regressions using the "simple" model that aggregates all technological spillovers into a single bin. Standard errors clustered at the CPC group level are reported in parentheses.

References for Appendices

Acemoglu, D., Akcigit, U., and Celik, M. (2020). Radical and incremental innovation: The roles of firms, managers and innovators. *American Economic Journal: Macroeconomics*, forthcoming.

Aghion, P. and Jaravel, X. (2015). Knowledge spillovers, innovation and growth. *The Economic Journal*, 125(583):533–573.

Agrawal, A., Galasso, A., and Oettl, A. (2017). Roads and innovation. *Review of Economics and Statistics*, 99(3):417–434.

Alcácer, J., Gittelman, M., and Sampat, B. (2009). Applicant and examiner citations in us patents: An overview and analysis. *Research Policy*, 38(2):415–427.

Angrist, J., Imbens, G., and Krueger, A. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Arora, A., Belenzon, S., and Lee, H. (2018). Reversed citations and the localization of knowledge spillovers. *Journal of Economic Geography*, 18(3):495–521.

Azoulay, P., Graff Zivin, J., Li, D., and Sampat, B. (2019). Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *Review of Economic Studies*, 86(1):117–152.

Branstetter, L. (2006). Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States. *Journal of International economics*, 68(2):325–344.

Bryan, K., Ozcan, Y., and Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy*, 49(4):103946.

Clarke, D. (2017). Estimating difference-in-differences in the presence of spillovers. *Mimeo*.

Cohen, W. and Levinthal, D. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, pages 128–152.

Conti, A., Thursby, J., and Thursby, M. (2013). Patents as signals for startup financing. *The Journal of Industrial Economics*, 61(3):592–622.

Delgado, M., Porter, M., and Stern, S. (2016). Defining clusters of related industries. *Journal of Economic Geography*, 16(1):1–38.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Griffith, R., Harrison, R., and Van Reenen, J. (2006). How special is the special relationship? Using the impact of US R&D spillovers on UK firms as a test of technology sourcing. *American Economic Review*, 96(5):1859–1875.

Griliches, Z. and Mairesse, J. (1995). Production functions: The search for identification. *Mimeo*.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Howell, S. (2017). Financing innovation: Evidence from R&D grants. *American Economic Review*, 107(4):1136–64.

Jaffe, A., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3):577–598.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 132(2):665–712.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542.

Manso, G. (2011). Motivating innovation. *Journal of Finance*, 66(5):1823–1860.

Mayer, T. and Zignago, S. (2011). Notes on CEPII's distances measures: The GeoDist database. *Mimeo*.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Myers, K. (2020). The elasticity of science. *American Economic Journal: Applied Economics*, 12(4):103–134.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, 95(1):450–460.