Are models getting harder to find?

Master of Philosophy in Economics
Faculty of Economics
University of Cambridge
Word count: 9,800
August 2020

The candidate acknowledges that this is their own work, and that they have read and understood the University's definition of Plagiarism.

Abstract

We estimate an R&D-based growth model using: (1) data on machine learning performance using a monthly panel dataset on the top performance across 93 machine learning benchmarks, and (2) data on research input derived from data on academic publications. Using a co-integrated error correction approach, we estimate the average research elasticity of performance to be around .02% to .13% across three sub- elds of machine learning considered. Our estimates indicate modest positive inter-temporal knowledge spillovers, but stark diminishing returns to research e ort (or large stepping-on-toes e ects in the context of the knowledge production function). We illustrate the upshot of our results by computing the paths of average researcher productivity and nd a sharp decline of between 4% and 26% per year. To further explain our results, we assess the contributions of di erent e ects to the decline in research productivity by computing counterfactual growth paths. We nd that while the decline in productivity is partially due to it becoming more di cult to improve on machine learning benchmarks for which performance is close to ideal, it is mostly due to other sources of diminishing returns to research e ort.

Contents

1 Introduction

R&D-based growth models are a core part of the toolkit in the study of economic growth. A notable example is endogenous growth theory—such as the models developed by Romer, 1990, Grossman and Helpman, 1991 and Aghion and Howitt, 1990—in which the process of knowledge accumulation is brought to the foreground of the analysis of economic growth. In these models, the R&D sector initiates a process of knowledge accumulation that contributes to productivity in economic production, and often determines growth at the frontier. In addition to the study of economic growth, R&D-based growth models have found application in environmental economics research (Acemoglu et al., 2012; Aghion and Jaravel, 2015), income inequality (Aghion, 2002; Lloyd-Ellis, 1999), and the topic of automation and employment (Prettner and Strulik, 2017; Acemoglu and Restrepo, 2018).

As it has been pointed out (e.g. in Kruse-Andersen, 2017), the conclusions of R&D-based growth models are often strongly affected by the underlying growth mechanisms. For example, consider the first-generation fully endogenous growth model (such as that found in Romer, 1990). These incorporate an idea production function of the form:

$$\frac{\dot{A}(t)}{A(t)} = \alpha X(t) \tag{1}$$

where $\dot{A}(t)/A(t)$ is total-factor productivity growth in the economy and $X(t)$ is some measure of research intensity, such as the portion of labour or the portion of total economic activity that is directed to R&D. In general, the equation implies that the growth rate in ideas is proportional to the number of researchers: scaling up the number of researchers results in a proportional increase in the growth in the number of new ideas. In turn, one of the implications is that a research subsidy that increases the number of researchers permanently will permanently raise the growth rate of the economy. Doubts have been cast on these conclusions (e.g. Bloom et al., 2020). Despite that, the fully endogenous variety remains widely used in economic growth research.[1] Our work tests this basic tenet using microeconomic empirical data.

We choose the field of machine learning to investigate microeconomic evidence on idea production functions, and test the hypothesis of constant research productivity. The field of machine learning is chosen for three reasons. Firstly, many series on the performance of machine learning models are available. This is in part because it is common practice for researchers to assess the performance of novel methods or models by means of benchmark experiments (Barredo et al., 2020; Hothorn et al., 2005)—empirical experiments whereby the researcher applies their methods to solve a fixed and well-defined task using some dataset or some data generating process of interest. Secondly, given that the set of tasks and dataset involved in benchmark experiments are generally unchanged over time, these experiments produce performance measures that enables one to directly compare machine learning models developed at different points in time.

Thirdly, the field of artificial intelligence (AI) is itself directly of interest in economics. This includes recent work on the effects of productivity and growth effects of the deployment of AI systems (Aghion et al., 2017; Agrawal et al., 2018; Nordhaus, 2015), and work on automation and employment (Acemoglu and Restrepo, 2018; Graetz et al., 2015). Since machine learning is the dominant recent approach to AI, the idea production function of the field of machine learning has implications for how one might model the development of AI systems and the evolution of its economic implications.

Our work is organised as follows. We first draw out the relationship of our work to the existing literature. Since our work is related to existing work in both economics and machine learning, we provide two brief distinct overviews of the relationship to the existing work in each domain. In our analytical framework section we develop a theoretical R&D-based growth model, based on Aghion and Howitt, 1990 and Jones, 1995, and develop an empirical model that incorporates the cointegrating relationship that is implied by the theory. Next, our two datasets—our performance dataset and our research input dataset—are presented, along with a description of our data collection methods. Using a cointegrated error correction approach, we estimate the coefficients that describe the intertemporal knowledge spillover effects and the magnitude of diminishing marginal returns. We subsequently illustrate the quantitative implications of our estimates by computing the implied paths of average productivity. Finally, we show that our results are robust to justifiable changes in the empirical approach. Our work concludes with a discussion that includes a description of the contributions of our work to the existing literature, as well as the important limitations of our approach.

[1] See Bloom et al., 2020 and Kruse-Andersen, 2017 for discussions on the widespread use of fully endogenous variety of growth models in economics.

2 Relationship to the existing literature

2.1 Research productivity literature

There are numerous existing studies that provide evidence that research productivity is slowing down. Notable amongst these is the comprehensive recent work by Bloom et al., 2020. Similarly to our work, they use microdata—theirs spanning the domains of semiconductors, agricultural production and medical innovations—to study idea production function at the micro level. By computing the flow of ideas per researcher, they find that research productivity is falling reliably in each domain at rates that range from 4 percent per year (for the discovery of new molecular entities), to 13 percent per year (for firm-level revenue productivity). In their words: "everywhere we look we find that ideas, and the exponential growth they imply, are getting harder to find" (Ibid., p. 1). Their work, like ours, focusses on drawing out the tight connections of research productivity to the theory of economic growth.

A substantial portion of related research productivity literature uses data on patents, and specifically the number of (quality-adjusted) patents per researcher or per dollar of research spending. Early works along these lines are documented in the survey by Griliches, 1998 which generally focus on the United States, and find near-monotonic decreases in the patents per R&D dollar ratio. Similar trends are found in Europe and elsewhere (Evenson, 1993; Lanjouw and Schankerman, 2004). Additionally, Kortum, 1993 use data from 20 U.S. manufacturing industries and find, consistent with earlier work, a decline by a factor of three of the R&D-to-patent ratio over the 30-year period ending in the late 1980's. Kortum, 1997 finds that the number of patents per researcher declined by roughly 5% per year from the 1950's to the 1980's.

There have been cases in which the mostly ubiquitous downward trend in research productivity has reversed. For example, Kortum and Lerner, 1998 documents a surge of U.S. patenting from the mid-1980's onward leading the number of patents per researcher to stabilize and even increase. Similarly, by studying high-tech patenting (including software, cloud computing, machine learning, semiconducters), Webb et al., 2018 finds implied improvements in researcher productivity in inventor productivity in software and semiconductors in the late 1990's. Nevertheless, consistent with the usual story, they find that for all other fields considered, the number of patents per inventor has declined near-monotonically.

Finally, there is a very lage and rich literature on research productivity in science and academia. In most cases assessment of scientific productivity has been based on publication and citation data (including bibliometric indicators such as the h-index, and impact factors) though some used other indicators, such as the number of patents, peer research assessments, and the awarding of prizes. (see Bar-Ilan, 2008 for a review of the literature). Much is known about the correlates of research productivity as measured by bibliometric indicators—such as research group size (von Tunzelmann et al., 2003; Carayol and Matt, 2004), institution prestige (Ramsden, 1994), participation in research collaborations (Braun et al., 2001), gender (Halevi, 2019; Amano-Patiño et al., 2020; Abramo et al., 2009). However, the application of bibliometric indicators for assessing scientific impact of quality is controversial and have various well-known limitations (Aksnes et al., 2019). Our work employs experimental performance data specific to the field of machine learning, that arguably more directly tracks scientific progress than many bibliometric indicators.

2.2 Progress in machine learning literature

There are a large number of surveys of artificial intelligence and machine learning that review the recent progress made these domains. These include reviews of natural language processing (Cambria and White, 2014; Li et al., 2020; Sun et al., 2017), computer vision (Rawat and Wang, 2017; Lu and Weng, 2007), games playing (Yannakakis and Togelius, 2018), medicine and healthcare applications (Deo, 2015; De Bruijne, 2016; Rong et al., 2020), as well as general reviews of the field as a whole (Jordan and Mitchell, 2015). These works describe the emerging research trends, showcase noteworthy demonstrations of utility in applications, and highlight areas of promise, amongst other things. There tends to be widespread agreement of there having been exciting progress in many parts of the field of machine learning, driven in part by the development of new learning algorithms, improvements in theory, and the availability of data and low-cost computation. Although these works are important to gain an 'inside view' of the extent of the progress made, the reviews generally do not provide a sufficiently rigorous data-driven treatment needed to warrant inferences about what the rate of progress or the level of research productivity has been at different points in time.

There is some existing work that take a data-driven approach to measuring progress in machine learning. Amongst these are the analyses the performance of state-of-the-art models on benchmark experiments presenting in Stanford University's Human-Centered Artificial Intelligence Institute's Annual AI Index Report (Perrault et al., 2019; Shoham et al., 2018). However, these works mostly offer descriptive statistics, and

offer little in the way of analyses that aim to uncover insights about the rate of progress. More similarly to our work, a review of algorithmic progress by Grace, 2013 examines performance data on a few machine learning benchmarks in natural language processing and computer vision tasks. They find that for some tasks, the rate of improvements decreased over time. However, little explanation of this finding is offered. Moreover, given when their work was done, the amount of performance data analysed is meagre compared to the amount of data that is currently available.

A notable recent work on algorithmic progress in machine learning is that by Hernandez and Brown, 2020. They consider the algorithmic efficiency of a variety of image classification models devised between between 2012 and 2019. As is common in the field of computer science (see e.g. Kozen, 2012), they assess algorithmic progress by comparing its run time or time complexity (i.e. the time or number of operations required to execute the algorithm). Specifically, they implement a variety of image classification models created between 2012 and 2019, and measure the amount of computation required to reach a fixed level of performance. They find that more recent classification models reliably required less computation to reach a given level of performance, thus being more efficient than those preceding it. In particular, they find that the latest (at the time) state-of-the-art approach required 44 times less computation than the earliest model to reach a chosen level of performance. Overall, their measure of the level of algorithmic efficiency, given in terms of computational cost, was found to double approximately every 16 months. Our work differs from theirs in that our primary focus is estimating the contributions of research effort to progress in machine learning.

3 Analytical framework

The identification strategy is as follows. A general version of an R&D-based growth model is developed along the lines of Romer, 1990 and Jones, 1995. The theoretical model describes the evolution of technology given a path of effective research effort, and predicts a cointegrating relationship between improvements in technology and research effort. We exploit the predicted cointegrating relationship using an error correction model that is estimated using: (1) a monthly panel dataset on the top performance across 93 machine learning benchmarks, and (2) data on research input derived from data on academic publications over the 2012 to 2020 period. The theoretical model is designed to capture the long-run dynamics of the system, which are the parameters of primary interest. The short-run dynamics of the empirical model are left unrestricted, permitting various adjustment processes to the cointegrating relationship.[2]

3.1 Theoretical model

Consider an economy in which final goods can be consumed, invested, or dedicated to research (time is continuous and indexed $t \quad 0$),

$$Y(t) = C(t) + I(t) + R(t). \tag{2}$$

where $Y(t)$ is aggregate output of final goods, $C(t)$ is aggregate consumption, $I(t)$ is aggregate capital investment, and $R(t)$ is aggregate investment in R&D.

As in Aghion et al., 1998, final goods are produced from labor specialized capital goods. There is a fixed measure of product variety normalized to unity,

$$Y(t) = L(t)^{1 \ \alpha}\left(\int_0^1 m(i,t)^{\alpha} A(i,t)di\right), \ \ \alpha \ 2 \ (0,1), \tag{3}$$

where $m(i,t)$ is the quantity of specialised capital good i, $A(i,t)$ is the productivity associated with i, and $L(t)$ is aggregate labor input. Again, following Aghion et al., 1998, it is assumed that an amount $A(i,t)$ of raw materials is required to produce specialised capital good i (i.e., more advanced specialised goods take more resources to build). The market-clearing condition for capital is:

$$K(t) = \int_0^1 m(i,t)A(i,t)di. \tag{4}$$

Let the technological level, $A(t)$, be defined as the average productivity associated with the specialized capital goods:

$$A(t) \quad \int_0^1 A(i,t)di. \tag{5}$$

[2]Our approach of exploiting the predicted cointegrating of an R&D-based growth model is similar to the approach taken by Kruse-Andersen 2017 and Ha and Howitt, 2007.

As in Jones, 1995, we assume that investments in research increase the technological level such that

$$\dot{A}(t) = \lambda A(t)^{\phi}\left(\frac{R(t)}{A(t)}\right)^{\sigma}, \quad \lambda > 0, \quad \sigma \in [0,1], \quad A(0) > 0. \tag{6}$$

As pointed out in Jones, 1995, the $A(t)^{\phi}$ term captures an intertemporal knowledge spillover for technological opportunities. Namely, it captures the effect of the current level of technology on scientists' ability to make progress. This might be positive ($\phi > 0$), what is called in Aghion et al., 1998 a 'standing-on-shoulders effect' (such that innovations today are bootstrapped by previous progress) or negative ($\phi < 0$): a 'fishing-out effect', (such that past discoveries make new ideas harder to find).[3] Finally, in the case where $\phi = 0$, productivity in the research sector would be independent of the level of technological development.

Here $R(t)/A(t)$ is the amount of final goods spent on research effort divided by the average productivity in creating final goods, and represents the measure of effective research input. The parameter σ captures the diminishing returns at a point in time. That is, doubling the number of researchers will produce fewer than double the number of unique ideas or discoveries because of the duplication of work, or some other source of diminishing returns. This parameter is often termed the duplication externality parameter, and $\sigma < 1$ is described as a 'stepping on toes effect' (see Jones, 1995; Gomez, 2011). Finally, note that the coefficient σ is also simply equivalent to the elasticity of technological progress with respect to research effort.

In addition, the stock of capital evolves in the usual way,

$$\dot{K}(t) = I(t) \quad \delta K(t), \quad \delta \in (0,1), \tag{7}$$

where δ denotes the capital depreciation rate. Specialized capital good varieties are produced under monopolistic competition, while final goods are produced under perfect competition. It follows that in equilibrium, all specialized capital good varieties are produced in the same quantity. Hence $m(i,t) = \bar{m}(t)$ for all i. Thus, it follows that $Y(t) = L(t)^{1 \ \alpha}\bar{m}(t)^{\alpha}A(t)$. This gives us the well-known production function,

$$Y(t) = K(t)^{\alpha}[A(t)L(t)]^{1 \ \alpha}. \tag{8}$$

We can rewrite (6) to yield an expression for technological progress,

$$g_A(t) \quad \frac{\dot{A}(t)}{A(t)} = \lambda A(t)^{\phi \ 1}\left(\frac{R(t)}{A(t)}\right)^{\sigma}. \tag{9}$$

Along the lines of Bloom et al., 2020 we construct a measure of the 'effective number' of scientists, by dividing total R&D expenditure by wage, giving the number of researchers that the economy's RD spending could purchase at that time. Letting $w(t) = \bar{\theta}Y(t)/L(t)$ be the wage for labour in this economy, with $\bar{\theta} \in (0,1)$. Then, we can define the effective number of scientists $\tilde{S}(t)$ as follows:

$$\tilde{S}(t) \quad \frac{R(t)}{w(t)} = \frac{R(t)L(t)}{\bar{\theta}Y(t)}. \tag{10}$$

To focus on the long-run equilibrium, the capital-output ratio is assumed constant: $K(t)/Y(t) = \kappa > 0$. This seems appropriate as this ratio has been mostly constant in many industrialised countries over long time frames, such as in the United States (Jones, 2016), as well as many others (D'Adda and Scorcu, 2003). Using (10), we can rewrite (9):

$$g_A(t) = \bar{\lambda}A(t)^{\phi \ 1}\tilde{S}(t)^{\sigma}, \quad \text{where } \bar{\lambda} = \lambda\bar{\theta}^{\sigma}\kappa^{\frac{\sigma\alpha}{1 \ \alpha}}. \tag{11}$$

In doing so, we have obtained a simple model that describes the evolution of $A(t)$ given the path of $\tilde{S}(t)$, i.e., it describes the evolution of technological progress with respect to the effective number of scientists.

Finally, dividing both sides by the effective number of scientists yields the path of average productivity, that is, productivity per scientist:

$$g_A(t)/\tilde{S}(t) = \bar{\lambda}A(t)^{\phi \ 1}\tilde{S}(t)^{\sigma \ 1}. \tag{12}$$

[3] It should be noted that the presence of a `standing-on-shoulders' e ect does not imply that the *rate* of technological progress is bootstrapped by the technological level. Equivalently, it need not mean that technological progress itself is accelerating with a constant amount of research input. This only occurs if the e ect is su ciently large; speci cally, it requires $\phi > 1$.

3.2 Empirical specification

Using logarithmic approximation of the growth rate $g_A(t)$,[4] the empirical log-discrete version of (10) becomes:

$$g_{i,t+1} = \bar{\lambda} \quad (1 \quad \phi) \ln A_{it} + \sigma \ln \tilde{S}_{it}, \tag{13}$$

where i indexes the relevant machine learning benchmark. Normalising in terms of $(1 \quad \phi)$, the co-integrating relationship amounts to:

$$\tilde{\lambda} \quad \ln A_{it} + \tilde{\sigma} \ln \tilde{S}_{it} \quad I(0), \text{ where } \tilde{\lambda} \quad \frac{\bar{\lambda}}{1 \quad \phi}, \ \tilde{\sigma} \quad \frac{\sigma}{1 \quad \phi}. \tag{14}$$

By Granger's representation theorem, if a group of variables is cointegrated, then they can be characterized as being generated by an error correction mechanism (Engle and Granger, 1987). Hence, we represent process (10) using an error correction model,

$$g_{i,t+1} = \delta_i^{0} \mathbf{d}_{it} + \tilde{\phi}(\ln A_{it} \quad \tilde{\sigma} \ln \tilde{S}_{it}) + \sum_{l=1}^{k_1} \lambda_{il} \Delta \ln A_{i,t \ \ l} + \sum_{l=1}^{k_2} \Lambda_{il} \Delta \ln \tilde{S}_{i,t \ \ l} + e_{it}. \tag{15}$$

The variable $\mathbf{d}_i$ contains the deterministic components, for which there are two cases. In the first case, $\mathbf{d}_{it} = (1, 0)^{0}$ so (15) contains a constant; in the second case, $\mathbf{d}_{it} = (1, t)^{0}$ so g_{t+1} is generated with a constant and a trend. The coefficients $(\delta_i^{0} \ \lambda_i \ \Lambda_i)$ are each heterogeneous (i.e. benchmark-specific), and $(\tilde{\sigma} \ \tilde{\psi} \ \tilde{\phi})$ are homogeneous long-run coefficients. Note that the coefficient $\tilde{\phi}$ is simply equivalent to $\phi \quad 1$. Lagged differences are included to allow for various adjustment processes to the cointegrating relationship, as well as mean-reverting dynamics in performance improvements.

Because of the risk of cross-sectional dependence across our panels, we take special care in modelling our error component e_{it}. Cross-sectional correlation of errors in panel data applications in economics are common, and may emerge from the presence of omitted common effects, spatial effects, or could arise as a result of interactions within socioeconomic networks (Chudik and Pesaran, 2013). In our series on machine learning benchmarks, common shocks may include the occurrence of conferences, the release of new frameworks or datasets, or the existence of synchronised funding cycles. In addition, different machine learning models may be capable of achieving state-of-the art performance on multiple benchmarks. These considerations suggest that we are likely to have common factors exerting influence on multiple series. In fact, as we shall see, for multiple groupings of benchmarks, we find evidence of cross-sectional dependence.

Hence, to accommodate for cross-sectional dependence across panels, the error e_{it} is assumed to have a multi-factor structure,

$$e_{it} = \lambda_i^{0} \mathbf{f}_t + \epsilon_{it}, \tag{16}$$

where $\mathbf{f}_t$ is a matrix of unobserved common factors and λ_i a vector of heterogeneous factor loadings. This specification permits cross-sectional correlations in the errors that are caused by common factors (including those correlated with our regressors of interest). This specification permits a general degree of error cross-sectional dependence by considering a multifactor structure with differential factor loadings over the cross-section units (Pesaran, 2006). The error term ϵ_{it} are the individual-specific (idiosyncratic) errors assumed to be serially uncorrelated and independently distributed of the regressors in (15).[5] Finally, the multifactor error e_{it} is assumed to be covariance-stationary.

To yield consistent estimates in the cases where we have error cross-sectional dependence, we estimate (13) using the Common Correlated Effects (CCE) approach developed by Pesaran, 2006. Including a sufficient number of lags of cross-sectional averages in individual equations (at least as many as the number of unobserved common factors), this approach enables the filtering out of differential effects of unobserved common factors. Hence, it yields consistent estimates of individual slope parameters and pooled long-run coefficients (Chudik and Pesaran, 2015). Thus, the model that is estimated is:

$$g_{i,t+1} = \delta_i^{0} \mathbf{d}_{it} + \tilde{\phi}(A_{it} \quad \tilde{\sigma} \tilde{S}_{i,t}) + \sum_{l=1}^{k_1} \lambda_{il} \Delta \ln A_{t \ \ l} + \sum_{l=1}^{k_2} \Lambda_{il} \Delta \ln \tilde{S}_{t \ \ l} + \sum_{k=0}^{p_k} \zeta_i^{0} \bar{\mathbf{Z}}_{\mathbf{t} \ \ \mathbf{k}} + e_{it}, \tag{17}$$

where the matrix $\bar{\mathbf{Z}}_t = (\overline{\ln \tilde{S}}_t \ \bar{g}_{t+1})$ contains cross-sectional averages of $\ln \tilde{S}_{i,t}$ and $g_{i,t+1}$. Throughout our empirical analysis, we will estimate variants of this equation.

[4] The discrete growth term $g_{i,t+1}$ is approximated using a standard log approximation $g_{i,t+1} \quad \ln A_{i,t+1} \quad \ln A_{it}$.

[5] Note that our assumption of a lack of serial correlation is not a particularly strong assumption and can be accommodated by including a su cient number of lagged di erences of our regressors.

4 Data

We use two separate datasets that we constructed or collated. The first is a performance dataset that covers top-performing machine learning or statistical models on 93 benchmarks over the 2012-01-01 to the 2020-06-01 period. These benchmarks cover the sub-fields of computer vision, natural language processing, and machine learning on graphs. The second is a research input dataset that includes a monthly series of publications, the number of authors, and a relative wage-weighted number of authors across the same sub-fields of computer vision, natural language processing, and machine learning on graphs, over the same period.

4.1 Performance dataset

The dataset includes performance series for 56 computer vision benchmarks, 16 natural language processing benchmarks, and 21 machine learning on graphs benchmarks. All performance measures are given in terms of accuracy or average precision (for additional details about the performance metrics see Measures of Performance in the appendix). We chose the aforementioned three sub-fields of machine learning because there are many benchmarks for tasks within the sub-fields, and because they span relatively long time-frames. In addition, benchmarks for tasks in the relevant sub-fields often employed the same performance metrics, such as accuracy or average precision.

Performance scores are bound between 0 and 1 (or 0 and 100), and correspond to accuracy or average precision. In case scores were between 0 and 100, we scaled these by a factor of 100. For both performance metrics, the higher the performance score, the better the model.

For each given benchmark, the performance series is a series of the best-performing models for each given month. Specifically, let A_{ti}^m denote model m's performance on benchmark i, that was submitted or made available at time t. Then, we construct our top-performance series A_{ti}^m as follows:

$$A_{it} \quad \max\{A_{ki}^m : k = 1, \ldots, t\}. \tag{18}$$

The series starts on the month when the first model or publication describing that model was submitted (whichever is sooner). Since new benchmarks are constantly being introduced, the panel dataset is unbalanced. Specifically, the length of a panel for a given benchmark is between 9 months to 139 months, with an average length of 56 months.

To ensure a minimally adequate amount of variation within each panel, we included only data on benchmarks that had over three top performing models reported over the 2012-01-01 to 2020-06-01 period.

The majority of the data is from Papers with Code, an open-source repository on machine learning code, publications and performance metrics. Furthermore, we harvested data from the KITTI Vision Benchmark Suite website, which hosts machine learning model performance data for a variety of computer vision benchmarks (Geiger et al., 2012). Finally, additional data on various natural language benchmarks were harvested from The Allen Institute for AI, a research institute. The data was harvested using our own Python scripts.

From Figure 1, it can be seen that different performance series start at different levels, have varying amounts of instances when progress is made, and progress is made at variable intervals. In order to ensure that our data covers the period over which the benchmark was relevant in the relevant sub-field, we deem a benchmark obsolete when it has been more than 90 days since the last model submission.[6]

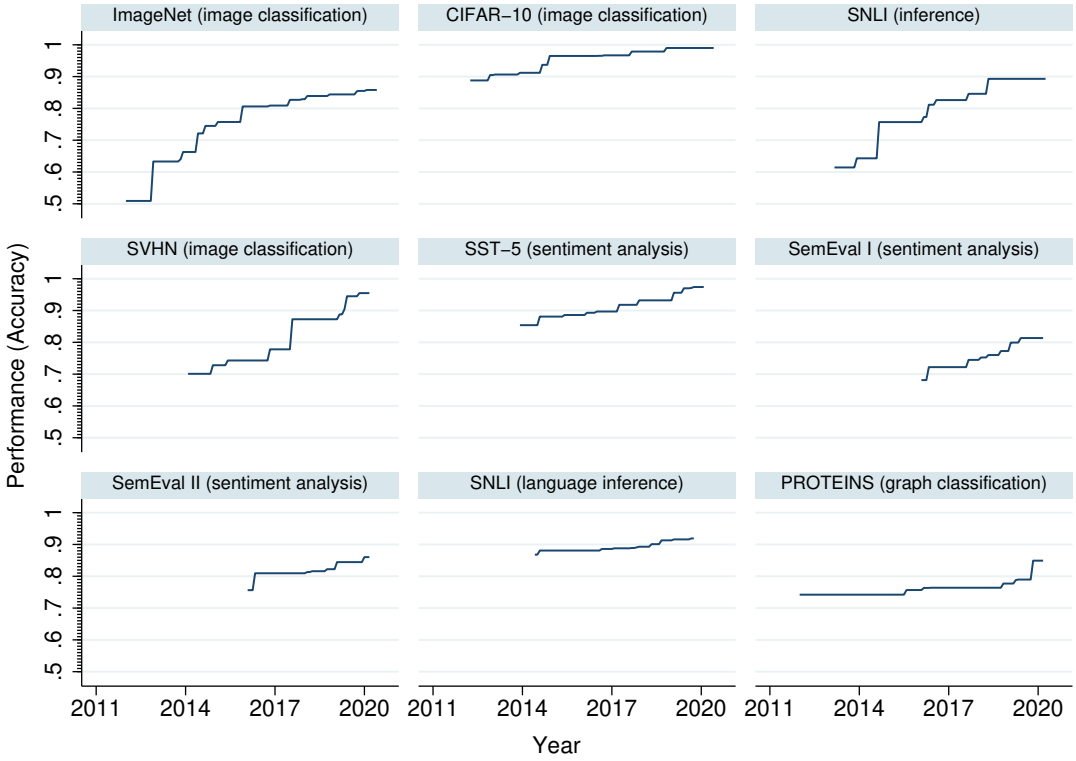
4.2 Research input dataset

The primary source of our data is from the Core Collection of the Web of Science (WOS) database. The WOS Core Collection is a large database of peer-reviewed, scholarly journals published worldwide (including Open Access journals) in disciplines ranging over the sciences, arts and humanities. The WOS Core Collection has been used in various previous bibliometric studies of machine learning and related disciplines (see e.g. Cioffi et al., 2020; El-Alfy and Mohammed, 2020).

In addition, we also use data on research input from arXiv, an online openly accessible repository for scholarly papers in various scientific disciplines; arXiv is commonly used to disseminate manuscripts and share their results with a wide community of researchers before peer review (Perrault et al., 2019). Hence, arXiv displays more recent work than the other sources. Therefore, we supplement WOS data with data harvested from arXiv for the final year of our series.

[6] Note that this specific operationalisation is not required for our main findings, as we will see in our Robustness Checks section.

Figure 1. A selection of performance series across various benchmarks



Performance time-series for a selection of benchmarks across the fields of image classification, natural language processing and machine learning on graphs.

For the sub-fields of computer vision and natural language processing, we identified the relevant publications by constructing keyword queries that are broadly in line with The Association for Computing Machinery (ACM) Computing Classification System (ACM, 1998). The ACM Computing Classification System (CCS) is a hierarchical classification system used to index and classify major areas and topics of the fields of computing and computer science (Lin et al., 2012). For the sub-field of machine learning on graphs, our queries were based on the taxonomy in Chami et al., 2020.[7]

As illustrated in figure 1, the WOS dataset is larger for all but the final 6-to-12 months of the sample-window, as it reports both more publications and a greater number of associated authors. The growth rate in the number of publications and authors is higher according to arXiv data—which likely reflects the growing popularity of arXiv amongst computer scientists (Sutton and Gong, 2017).

For the period starting in 2019-06-01, we supplement WOS data with arXiv data. Specifically, we add publications into the series that are found on arXiv but not in WOS.[8] Admittedly, this cut-off point is selected fairly arbitrarily, and by by performing a moving-window estimation exercise in our Robustness Check section, we confirm that our particular choices made here do not change our main results.

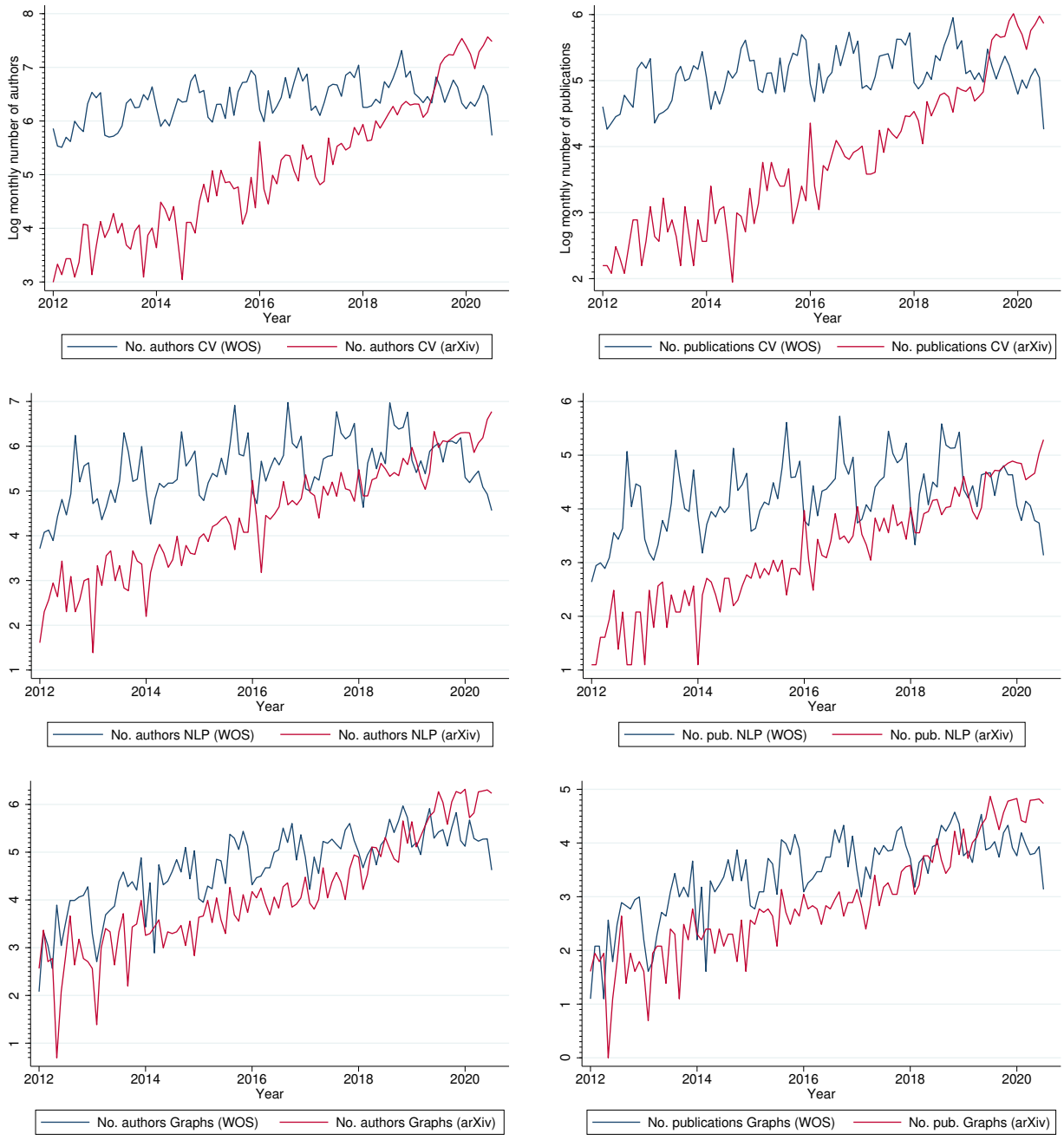
Using the data on the number of authors and publications, we construct a measure of the effective numbers of researchers separately for each sub-field. As we argued in our Theory section, standard R&D-based growth models predict that progress in technology is a function of research effort $\tilde{S}(t)$, defined as the total investment in research at time t divided by the wage rate at time t; this gives the number of researchers that the economy's R&D spending could purchase at that time.

However, since research involves more than the labour input of researchers, our number of authors metric will understate the true effective number of researchers. In the field of machine learning, other costs may include the cost of hardware and data (LeCun et al., 2015). To address this, we assume that for each scientific publication, there is an equipment cost that is paid on top of the wages to scientists. This equipment cost is assumed to be a constant fraction of the estimated mean labour-cost of a publication in that sub-field for

[7] A detailed description of all our search queries may be found in the Publication data collection Methodology section of the Appendix.

[8] Using simple functions on Microsoft Excel, we assessed the degree of similarity of titles and abstract.

Figure 2. Monthly number of authors (left column) and publications (right column) across three sub-fields of machine learning



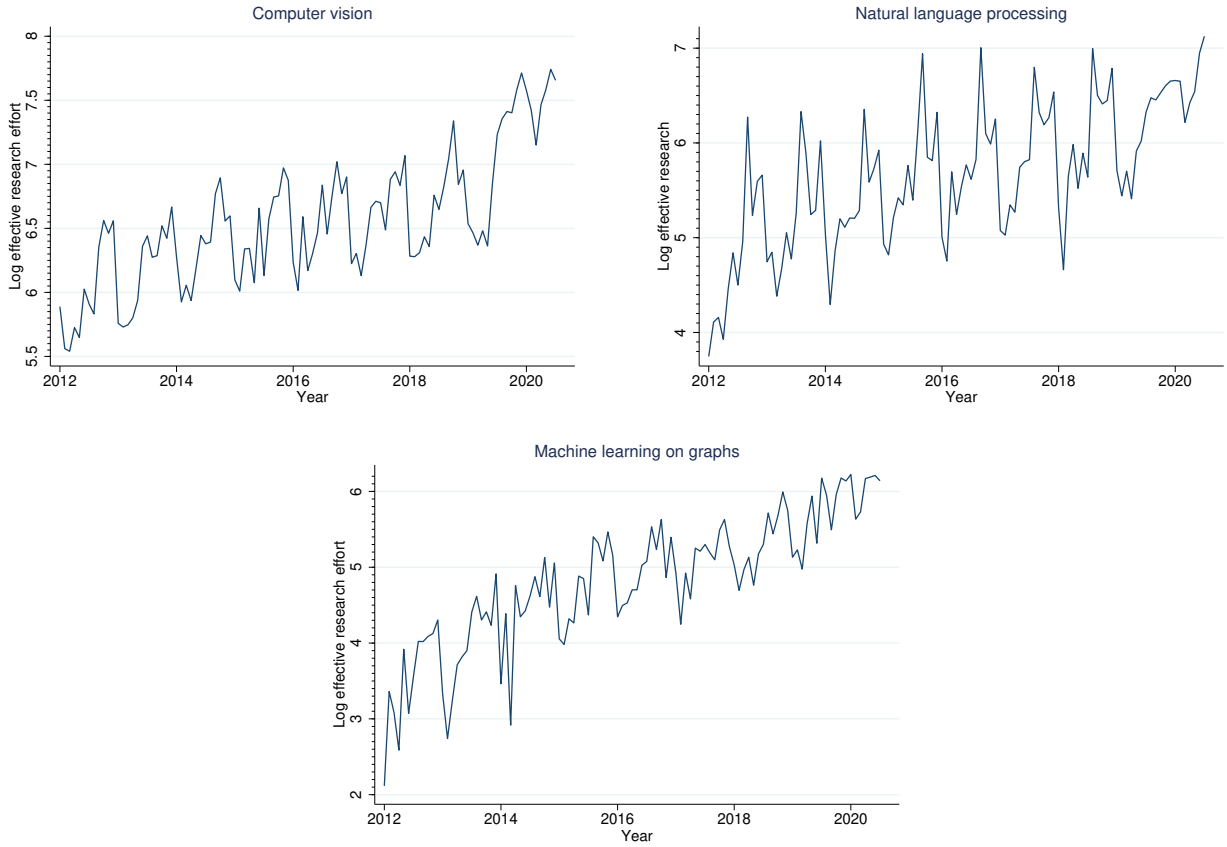
Number of authors and number of publications across categories over the 2012-01-01 to 2020-06-01 period for computer vision (CV), natural language processing (NLP) and machine learning on graphs (Graphs). Data is based on the WOS Core Collection and arXiv.

that month.[9]

A potential shortcoming of our measure of research effort is that there is no adjustment for changes in the quality of the researchers. Ideally, to know whether algorithms are getting harder to find, we would

[9] A shortcoming of using a constant-fraction of the labour-cost is that the `equipment' spending may trend upward over time. To the extent this is true, our measure of the e ective number of researchers may understate the rise in research e ort, and hence underestimate the fall in research productivity. To address this, we redid our results using a coarse adjustment whereby we assume that the fraction increases by 3% per year to accommodate any possible trends in per-project equipment expenditure.

Figure 3. Log monthly effective numbers of researchers across sub-fields of machine learning



Effective numbers of researchers for the sub-fields of computer vision, natural language processing and machine learning on graphs over the 2012-01-01 to 2020-06-01 period. We assume a constant fixed cost of 10% the mean monthly labour cost per publication as a baseline.

want to control for a measure of the level of productivity of scientists when faced with a research problem of some fixed level of difficulty. This measure is of course not observed, nor is the level of difficulty of problems encountered by researchers at different points in time. After all, if such measures were available, the answer of whether algorithms are getting harder to find would also essentially be known. Nevertheless, failure to control for the quality of researchers could result in biased estimates. For example, if the average quality of researchers increased systematically, failure to control for this may mistakenly lead us to conclude that researcher productivity was constant or decreasing.

Hence, along the lines of Bloom et al., 2020, we further construct a quality-adjusted measure of the effective numbers of researchers for each sub-field by weighting the number of researchers by a measure of relative wages. This may be a more desirable measure in that it weights the various research inputs according to their relative prices: if expanding research involves employing researchers of lower relative productivity, this will be properly measured by the number of wage-adjusted number of researchers.

Since research is performed globally, constructing this requires us to identify the machine learning research effort shares by country, and weighting these by the relative wages. We use data from de Kleijn et al., 2017 on research output shares by country, together with our dataset to identify the number of researchers for a list of major countries that collectively generate over 80% of yearly research in various fields of machine learning.[10] Subsequently, we weight the number of researchers for each country using data on wages for those engaged in scientific and technical activities across the relevant countries using data from the International Labour Organization—specifically ILOSTAT's mean nominal monthly earnings of employees by economic activity for ISIC-REV.4: M. *Professional, scienti c and technical activities* (ILO, 2020).

Throughout our analysis, we prioritise estimating the relevant coefficients using our data on quality-

[10] These are China; United States; India; United Kingdom; Germany; Japan; Spain; France and Italy | see de Kleijn et al., 2017.

adjusted effective numbers of researchers. Our baseline results are produced on the assumption that equipment costs amount to 10% of the total monthly labour costs. In our Robustness Check section we further consider a wide range of fractions that make up the equipment costs and show that our headline results are robust to alternative equipment cost specifications. Moreover, we check the robustness of our results by using numbers that are not adjusted for relative wages, along with other justifiable changes in the empirical approach.

5 Empirical analysis

Overview

To validate our approach, we first find evidence that our variables of interest are non-stationary and, as predicted by our R&D-based growth model, cointegrated. We subsequently estimate the parameters of our error correction model. We find evidence of positive intertemporal knowledge spillovers but also very pronounced diminishing marginal returns to research effort, or stepping-on-toes effects. We consider alternate model-specifications which seem broadly consistent with our baseline results. To illustrate the upshot of our estimates, we compute the implied paths of average productivity for each sub-field in machine learning over time, and find that the predicted level of average productivity declined by between 4% to 26% per year over the 2012-01-01 to 2020-06-01 period.

5.1 Stationarity and cointegration tests

The performance series for each benchmark is non-stationary by construction. This may be seen from the definition of A_{it},

$$A_{it} \quad \max\{A^m_{ki} : k = 1, \ldots, t\}, \tag{19}$$

for all i and t. The process is assumed to be stochastic, and can in turn be represented as follows

$$A_{it} = A_{i,t\ \ 1} + \epsilon_{ti}, \ \text{ where } \epsilon_{ti} = |u_{ti}| \text{ for } u_{ti} \quad D(\mu, \sigma^2). \tag{20}$$

which satisfies the definition of a unit-root process. Indeed, we confirm that there is little-to-no evidence of stationarity in any of our performance variables using Pesaran, 2007's test for unit roots in heterogenous panels with cross-section dependence (see Table 7 in the appendix).

In much of our empirical analysis, we treat the research effort series as a time-series variable. Since we group benchmarks according by sub-fields, research effort only varies across time, but not across benchmarks. Hence, in contrast to our performance series, we treat these series as time series and test their stationarity by implementing a standard augmented Dickey-Fuller (DF) test (Dickey and Fuller, 1979).

Table 1 presents the results of two variants of augmented DF tests. We include a test based on the Augmented DF auxiliary regression that includes both a constant and a trend. A trend is likely appropriate since it is visible from Figure 3 that research effort trends upward over time (albeit at different rates). We find no evidence of stationarity in research effort. Moreover, repeating the procedure the first-differenced counterparts of each series, we find strong evidence for stationarity in differences—indicating that the series are integrated of an order of 1.

Given the non-stationarity of the individual variables, a cointegration relationship is necessary for the stationarity for our linear combination of the variables of interest. We test the null hypothesis of no cointegration by carrying out different types of tests for whether e_{it} is non-stationary.

First, we use the residual-based cointegration test procedure for dynamic panels by Pedroni, 2004, which involved Phillip-Perron (PP) and DF tests, as well as a variance ratio test. We further corroborate these results using a test developed by Westerlund, 2005, which is a test that involves an initial estimation of the associated nuisance parameters, i.e. parameters that are unrelated to the cointegration relationship. Both tests are robust to serially correlated errors and accommodate individual specific short-run dynamics, individual specific fixed effects and deterministic trends, and individual specific slope coefficients.

The cointegrating relationship derived from R&D-based growth model is given by,

$$A_{i,t} = \delta_i^{\emptyset} \mathbf{d}_{it} + \beta_i \tilde{S}_{it} + e_{it} \tag{21}$$

where $\delta_i^{\emptyset} \mathbf{d}_{it}$ allows for panel-specific means, and panel-specific time trends, and the residuals e_{it} represent deviations from the long-run relationship.

Table 1
Test-statistics for the Augmented Dickey-Fuller test for each research effort series in levels and first-differences

	Research effort (levels)		
	Computer vision	NLP	Graphs
Constant	0.46 (0.90)	3.00 (0.13)	2.27 (0.99)
Constant and trend	1.667 (0.76)	1.26 (0.65)	0.74 (1.00)
	Research effort (first-difference)		
	Computer vision	NLP	Graphs
Constant	8.89 (< .001)	9.01 (< .001)	8.33 (< .001)
Constant and trend	9.01 (< .001)	9.01 (< .001)	8.71 (< .001)

The test's p-values are reported in brackets. The number of lags of di erences were selected by considering the Bayesian Information Criterion for each series. We ran our Augmented DF auxiliary regression with 1 lag for computer vision and 3 lags for NLP and Graphs. The number of observations for each variable was 152 k (where k denotes the number of lags of di erences) for the series in levels, and 151 k for the series in rst-di erences.

The DF and PP tests are implemented by fitting the model in (21) using ordinary least squares, obtaining the predicted residuals $\hat{e}_{it}$, and then fitting the following regression model,

$$\hat{e}_{it} = \rho_{1i}\hat{e}_{i,t\ \ 1} + \sum_{j=1}^{p} \rho_{j+1,i}\Delta\hat{e}_{i,t\ \ j} + u_{it} \tag{22}$$

where $\Delta\hat{e}_{i,t\ \ j}$ is the jth lag of the first difference of $\hat{e}_{it}$, $j = 1, ..., p$ is where p is the number of lag differences, and u_{it} is a stationary error term. In case of the PP tests, we have a panel-specific AR parameter ρ_i, whereas in the DF tests, $\rho_i = \rho$, for all i.

Table 2. Test-statistics of a battery of tests for the cointegration of research effort and performance

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Variance ratio	3.58*** (2)	1.14 (3)	3.43*** (0)	11.29*** (3)	40.72*** (2)
Modified Phillips-Perron t	6.18*** (2)	3.74*** (3)	12.87*** (0)	2.06** (3)	14.14*** (2)
Phillips-Perron t	1.52* (2)	2.79*** (3)	7.77*** (0)	2.01** (3)	1.93** (2)
Augmented Dickey-Fuller t	1.96** (2)	2.99*** (3)	6.54*** (0)	2.19** (3)	0.51 (2)
Variance ratio (West)	5.42***	2.42***	4.08***	2.03**	9.49***
Number of panels	25	16	20	30	62
Average number of periods	45.46	52.81	49.1	48.03	48.53

All tests have a common null hypothesis of no cointegration. Testing for cointegration simply amounts to testing the stationarity of (22). The VR tests are based on Breitung, 2002 with a test statistic constructed as a ratio of variances. Lags are reported in brackets beside the reported test statistic. The lags were selected using the lag-selection algorithm by Newey and West, 1994. The Aggregate benchmark group is an amalgam of all three series that are measured in accuracy (AC). *, **, *** indicate signi cance at the 10%, 5% and 1% signi cance level respectively.

Table 2 presents the results of the tests in the case in which our auxiliary regression (21) is estimated without a trend. We reject the null hypothesis of no cointegration at the 5% significance level for most benchmark groups in the case where the cointegrating relationship does not contain a time trend. For the panels for natural language processing and the Aggregate series, we fail to reject the null hypothesis of

no cointegration using the variance ratio statistic, and the augmented DF test. Yet, for these benchmark groups, the other test statistics do suggest that there is evidence for cointegration. As it has been argued that Westerlund's variance ratio test performs better in small samples (Westerlund, 2005), we prioritise the variance ratio (West) results. Finally, we re-run these tests after including a time trend in our cointegrating relationship, and similarly fail to reject the null hypothesis of no cointegration (see Additional cointegration tests in the Appendix).

5.2 Error correction model analysis

We adjudicate between different model-specifications in the following ways. Firstly, the number of lags are chosen by performing likelihood tests after estimating our models by maximum likelihood.[11] In addition, we further consider the model's adjusted R^2 and evidence of autocorrelation in the errors using the Cumby-Huizinga test for autocorrelation (Cumby and Huizinga, 1990).[12] The number of lags of cross-sectional averages is chosen by considering evidence of cross-sectional dependence by Pesaran, 2015.[13] In the case when adding additional lags of cross-sectional averages has an ambiguous effect on cross-sectional dependence, we defer to a heuristic suggested by Chudik and Pesaran, 2015 of introducing $\lfloor \bar{T}^{\frac{1}{3}} \rfloor$ lags of the dependent variables and the strictly exogenous variables are added (where $\lfloor . \rfloor$ is the floor operator), based on their finding consistency improvements in Monte Carlo experiments.

Table 3 presents our estimates of our baseline model—the empirical model derived in the analytical framework section. We find that σ, the long-run research elasticity of performance (hereafter elasticity of performance), is positive and significant for each group of benchmarks considered (with the exception of machine learning on graphs, which were not significant). Long-run elasticity point estimates are small: these range from 0.13 to 0.023. This implies that a one-percent increase in research effort improves performance by between 0.13% to 0.023% in the relevant performance metric on a given benchmark. In our baseline estimations, the returns to research effort in computer vision and natural language processing were found to be roughly similar (between 0.10 and 0.13), and our composite series of all benchmarks measured on accuracy had the lowest elasticity of performance, at 0.023.

These estimates of small elasticities suggest the presence of very substantial stepping-on-toes effects. To illustrate the magnitude of these effects, our estimate for computer vision benchmarks measured in average precision (for which we found the largest elasticity of performance) imply that it would roughly take an additional 2060 researchers to double the impact of the first 10 researchers, holding all constant.[14] Our estimate of σ indicates a larger stepping-on-toes effect than is usually found in the literature on R&D. For example, a meta-analysis by Sequeira and Neves, 2020 finds an average σ coefficient of around 0.2—roughly double the values we found for computer vision and natural language processing. For reference, values of σ below 0.2 are generally considered to be cases of large duplication externalities in work on optimal R&D spending (see e.g. Shiell and Lyssenko, 2014; Chu, 2010). Furthermore, theoretical work by Jones, 1995 finds that values of σ below 0.25 can generate over-investment in R&D in a decentralised economy, whereas Strulik, 2007 suggests the market may even allocate too much R&D effort at σ values as large as 0.48.

As expected given the cointegrating relationship, the coefficient on the level of performance $\tilde{\phi}$ is negative (ranging from -0.11 to -0.22) and significant at the 1% significance level for all benchmark groups. Since $\tilde{\phi} = \phi \quad 1$, our baseline estimation results place $\hat{\phi}$ within the 0.78 to 0.89 range. Hence, these results indicate standing-on-shoulders effects; it gets easier to generate new ideas (or models and algorithms in this case) as the level of performance improves. One the other hand, the spillover effects are lower than the standard assumption of $\phi = 1$ in fully endogenous variety of R&D growth models (Kruse-Andersen, 2017). A value of ϕ less than unity means that the standing-on-shoulders effects are modest in the sense that, although the flow of ideas (in the $\dot{A}$ sense) is increasing in the level of performance, the overall rate of progress (in the proportional $\dot{A}(t)/A(t)$ sense) is actually slowed down. In short, our estimates indicate that whilst it becomes easier to generate new innovations with increases in the level of performance, it gets harder to make progress on machine learning benchmarks once a model get closer to ideal performance.[15]

Moreover, since in our error-correction model, the coefficient $\tilde{\phi}$ is equivalent to our error-correction term, our estimates suggest that the process error corrects relatively quickly. Specifically, it suggests that most of

[11] To do this, we estimate our models using the `xtpmg` command by Blackburne III and Frank, 2007.

[12] The Cumby-Huizinga test for autocorrelation is implemented using the `actest` by Baum and Scha er, 2015.

[13] These tests are implemented using the `xtdcce2` command by Ditzen, 2018.

[14] That is, $10^{0.13} \quad \frac{1}{2}2070^{0.13}$.

[15] To further illustrate this point, consider the case in which performance was not bounded between 0 and 1 (or 0% and 100%), but measured on the positive real numbers. In this case, a value of ϕ on the unit interval implies that the improvements in the absolute terms on the performance metrics would increase in the level of performance, yet these improvements in relative (proportional) terms would decrease in the level of performance.

the additional insights generated by way of research, insofar as these enable performance improvements, are absorbed in the order of a few months after publication. Concretely, we compute the half-life of the effect of a shock, which indicates the length of time after a shock in research effort before the deviation in performance shrinks to half of its long-run impact. The median estimates of the half-lives for individual subcategories are found to be between 2.8 and 6 months. This may provide some indication of the relative extent to which the relevant research is dedicated with performance improvements (in contrast to, for example, developing theoretical insights that may take longer to have an impact on the field).

Table 3
Baseline estimation results from a cointegrated error correction model using machine learning performance and (quality adjusted) research effort data over the 2012-01-01 to 2020-06-01 period

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Log research effort	0.10 *** (0.029)	0.11 *** (0.004)	0.044 (0.036)	0.13 *** (0.047)	0.023** (0.009)
Performance (level)	0.22*** (0.057)	0.18*** (0.049)	0.11*** (0.025)	0.21*** (0.031)	0.15*** (0.02)
Disequilibrium half-life (months)	2.79*** (0.82)	3.49*** (1.05)	5.95*** (1.43)	2.94*** (0.49)	4.27*** (0.62)
R^2 adjusted	0.80	0.82	0.79	0.82	0.63
Number of lags	$(1,1)$	$(1,1)$	$(1,3)$	$(1,1)$	$(1,1)$
Number cr. lags	0	0	3	3	3
Test for C-D dependence (p-value)	0.70	0.21	< 0.01	0.01	< 0.01
Observations	1133	819	1072	1367	2793
Benchmarks	26	16	21	30	61

The disequilibrium half-life term is calculated by hand and the standard errors are approximated using a Taylor approximation as described in Appendix: section B.3.2. The number of lags are selected based on LR tests, together with considerations regarding cross-sectional dependence and autocorrelation in the error term. Number of lags are presented in brackets, where the first number denotes the number of lags on the differences in log performance and the second the number of lags on the differences in log research effort. In addition to lags of the differences of variables, we include lags of cross-sectional averages of the dependent variable and the research effort variable. Here the number of cr. lags is chosen to be the smallest integer in $[0, \hat{p}^{3}\overline{T}c]$ such that the null hypothesis for no cross-sectional dependence cannot be rejected at a 5% significance level, and exactly $\hat{p}^{3}\overline{T}c$ otherwise.[16] Small sample time series bias is corrected by recursive mean adjustment correction methods proposed by Chudik and Pesaran, 2015.

[9] Note that T is the arithmetic mean of T, and also that, in general, $T \neq T$ given that our panels are unbalanced.

When a trend is included in the error correction model, to accommodate for the possibility that $\bar{\lambda}$ was time-varying, the estimated coefficients on log research effort are significant and positive for all (with again the exception of machine learning on graphs), and slightly smaller, ranging from 0.089 to 0.016 (see Table 9 in the appendix). In addition, our estimates imply that ϕ remains within the unit interval but decreases—thus giving more muted standing-on-shoulders effects. Finally, the time-trend coefficients are generally positive and significant, but small. The largest year-trend coefficient was 0.003, for computer vision in average precision—suggesting that the constant term trended slowly, and that the relevant process adjusted only rather gradually over time.

Table 4 provides the results from our co-integrated error correction model with time-varying performance elasticities. The interaction between research effort and time is positive and significant for a majority of the benchmark groups (with the only exception being machine learning on graphs). Our results indicate a minor upward trend in the elasticities of performance. The fastest trending elasticity of performance is for computer vision benchmarks measured in average precision. However, even for this benchmark group, our point estimate indicates that the trend only amounts to .0004 per month, which amounts to a change of 0.0048 per year. For reference, 0.0048 represents just 0.3% of our point estimate for the σ for the same benchmark group from our baseline estimation results. As such, the stepping-on-toes effect do seem not to have weakened substantially over the sample period.

Table 4
Estimation results from cointegrated error correction model that includes time-varying performance elasticities

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Log research effort	0.040 (0.030)	0.010 (0.007)	0.065 (0.0488)	0.051 (0.050)	0.011 (0.010)
Log research effort month	0.0002 *** (0.00005)	0.00003*** (0.00001)	0.00002 (0.0001)	0.0004*** (0.0001)	0.00008** (0.00003)
Performance (level)	0.30*** (0.053)	0.28*** (0.058)	0.17*** (0.036)	0.48*** (0.082)	0.21*** (0.02)
Disequilibrium half-life (months)	1.91 *** (0.041)	2.11 *** (0.52)	3.72 *** (0.87)	1.06 *** (0.26)	2.94 *** (0.32)
R^2 adjusted	0.76	0.79	0.61	0.76	0.57
Number of lags	$(1, 1)$	$(1, 1)$	$(1, 1)$	$(1, 1)$	$(1, 1)$
Number cr. lags	0	0	3	3	3
Test for C-D dependence (p-value)	0.77	0.28	< 0.01	0.02	< 0.01
Observations	1133	819	1072	1367	2745
Benchmarks	26	16	21	30	60

Months are values between 0 and 149. Number of lags in brackets are given as (k_1, k_2) where k_1 denotes the number of lags on the di erences in log performance and k_2 the number of lags on the di erences in log research e ort.

5.3 Computing average productivity

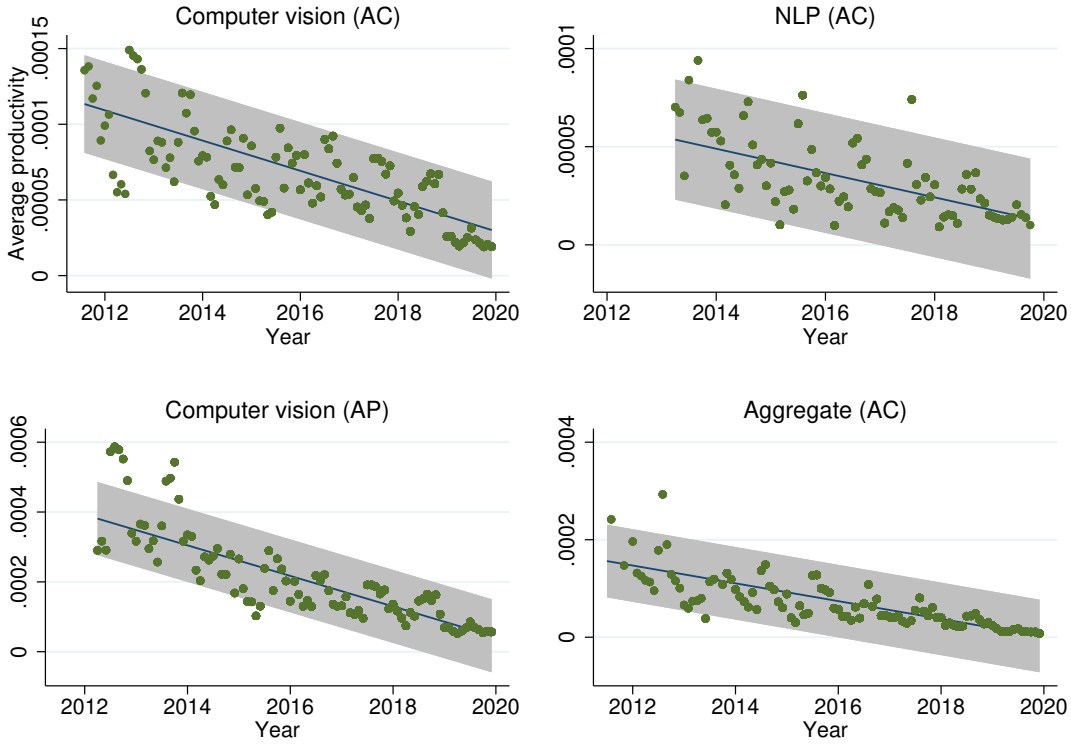
We now turn our attention to what our estimates imply about how researcher productivity has been faring over time. We compute the level of average research productivity in each sub-field of machine learning by fitting the paths of average productivity using our estimates,

$$g_{i,t+1}/\tilde{S}_{it} = \bar{\lambda} A_{it}^{\phi \ 1} \tilde{S}_{it}^{\sigma \ 1}. \tag{23}$$

That is, we fit parameters $\bar{\lambda}$, σ, and ϕ using their estimated counterparts. To compute the average productivity at the level of each sub-field (rather than at the level of the benchmark), we compute the arithmetic mean of (23) for each benchmark i for each benchmark group.

Computing these paths reveals a robust implication of our estimates: the average level of productivity per research effort is declining. Figure 4 presents our estimated average level of productivity for four benchmark groups, as predicted by baseline estimation results in Table 3. The predicted decline in average productivity in computer vision is especially pronounced; the mean year-on-year decline averages at 21% and 23% for the benchmark group measured in accuracy and average precision, respectively. For natural language processing the average decline is smallest yet still substantial: 13% year-on-year. We further fit (23) using the point

Figure 4
Monthly means of average research productivity across sub-fields of machine learning as predicted by baseline estimation results in Table 3



Individual markers represent the mean of the predicted monthly level of productivity for all benchmarks. Monthly mean levels of productivity are predicted by tting equation (23) with point-estimates from Table 3. The solid lines are linear regressions, and the shaded region represents a 95% con dence interval. The con dence interval represents that for an individual forecast, which includes both the uncertainty of the mean prediction and the residual.

estimates from the two other model specifications: one that includes a trend, and another that includes time-varying elasticity of performance. Modulo some quantitative differences (see Table 5), the picture remains broadly the same: the predicted level of average productivity seems to be declining rapidly.

Finally, we assess the contributions on the average productivity decline of our two effects: the intertemporal spillovers (whereby past advances on benchmarks results in future improvements being harder to make), and the effect of diminishing returns to research at points in time (or what is also called the duplication externality). To do this, we impose restrictions on the relevant parameters and compute counterfactual productivity paths in the cases in which these effects were turned off, as it were. Concretely, for the intertemporal spillover effects on the rate of progress to be omitted, we impose the restriction $\phi = 1$, so that the rate of progress becomes independent of the level of performance. Likewise, for the effect due to diminishing returns at points in time to be omitted, we impose constant marginal returns and restrict σ to 1. Table 5 summarises the mean year-on-year average productivity changes by considering the point estimates of our three model-specifications.

Table 5
Computed year-on-year proportional change in average productivity across five benchmark groups based on estimation results of three models, including computed counterfactual paths

		Benchmark group				
Model type	Restrictions	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
(1) baseline	None	0.21 (.008)	0.13 (0.12)	0.23 (0.011)	0.23 (0.006)	0.26 (0.006)
	$\phi = 1$	0.17 (0.007)	0.095 (0.011)	0.17 (0.01)	0.17 (0.007)	0.21 (0.006)
	$\sigma = 1$	0.05 (0.003)	0.03 (0.001)	0.07 (0.007)	0.06 (0.003)	0.06 (0.001)
(2) baseline with time trend	None	0.10 (0.009)	0.038 (0.010)	–	0.14 (0.007)	0.22 (0.007)
	$\phi = 1$	0.06 (0.009)	0.008 (0.01)	–	0.071 (0.01)	0.098 (0.009)
	$\sigma = 1$	0.09 (0.004)	0.14 (0.001)	–	0.046 (0.003)	0.001 (0.001)
(3) baseline with time-varying elasticity	None	0.12 (0.006)	0.13 (0.12)	0.23 (0.011)	0.19 (0.005)	0.26 (0.006)
	$\phi = 1$	0.070 (0.005)	0.097 (0.01)	0.17 (0.01)	.11 (.006)	.20 (.006)
	$\sigma = 1$	0.05 (0.003)	0.04 (0.001)	0.071 (0.007)	0.077 (0.003)	0.06 (0.001)

Figures are based on the paths computed using point estimates of our error correction models. Model (1), (2), and (3) refer to the model speci cations presented in tables 3, 4, and 8 respectively. The standard errors are generated by bootstrapping with 1000 replications. The estimates for graphs using our estimates for σ were negative, and were therefore omitted.

For all benchmark groups and model specifications, we find that imposing the restriction that $\phi = 1$ results in a much smaller reduction in the predicted rate of productivity decline, than when imposing the restriction that $\sigma = 1$. Thus, the presence of diminishing research at points in time (i.e. our low estimates for σ) accounts for most of the implied productivity decline over the 2012-01-01 to 2020-06-01 period. For our baseline model and the model with time-varying elasticity of performance, the restriction of $\sigma = 1$ still produces declining productivity path, indicating that the intertemportal spillovers do contribute to the decline in productivity. However, these effects were less robust or smaller in size. For instance, when a trend was included in the relevant model, there are some cases with gains in average productivity indicating that the increase in $\bar{\lambda}$ dominates the negative intertemportal spillovers.

6 Robustness checks

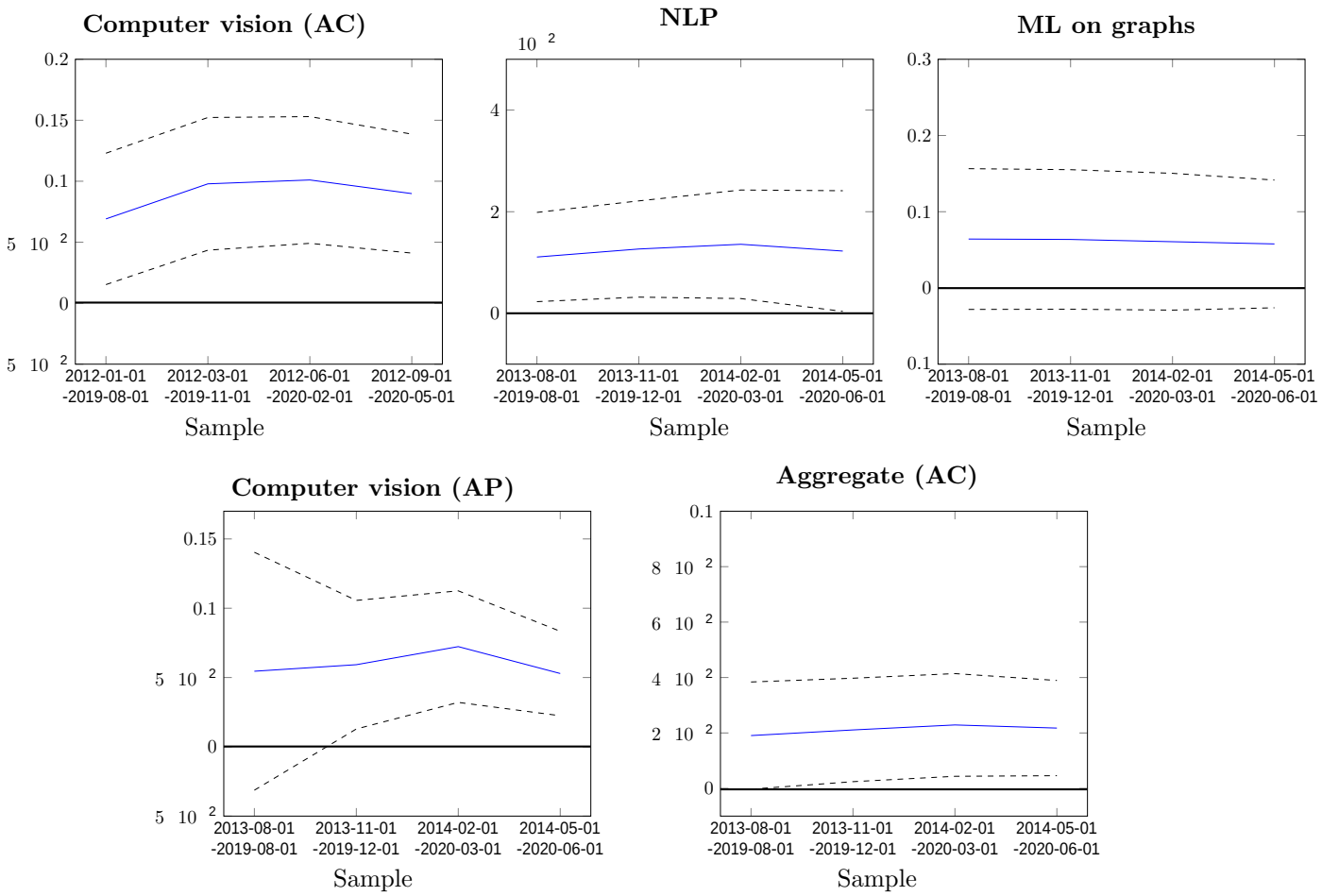
In this section, we show that our earlier results are robust to justifiable changes in the empirical approach. The patterns from our earlier work repeat: (1) our estimates reliably place elasticities of performance in the 0.1 to 0.01 range, and (2) the intertemporal spillover is estimated at between 0.7 and 0.8 (i.e. $\tilde{\phi}$ at between -0.3 and -0.2).

6.1 Moving window estimation

It is important to verify that the start and end years of the sample are not essential for the results. In this section we perform a moving-window estimation exercise by estimating our model on different partitions of our samples. Specifically, we estimate our earlier models over a rolling-window chosen so that the start-and end-periods depart by one year from the sample's actual start- and end-periods. Since the last year of our research input dataset was supplemented with arXiv data, this simultaneously enables us to confirm that our results do not crucially depend on the inclusion of that segment of data.

Figure 4 displays the results from estimating our baseline model from table 3 on a variety of time-windows. This broadly confirms our earlier baseline estimation results that our model quite robustly places the elasticity of performance at a level between 0.1 to 0.02, depending on the benchmark group involved. With the exception of ML on graphs (for which the coefficient had previously not shown statistical significance) and one sample partition for computer vision (AP), the 95% confidence bands are within the 0 to 0.20 range.

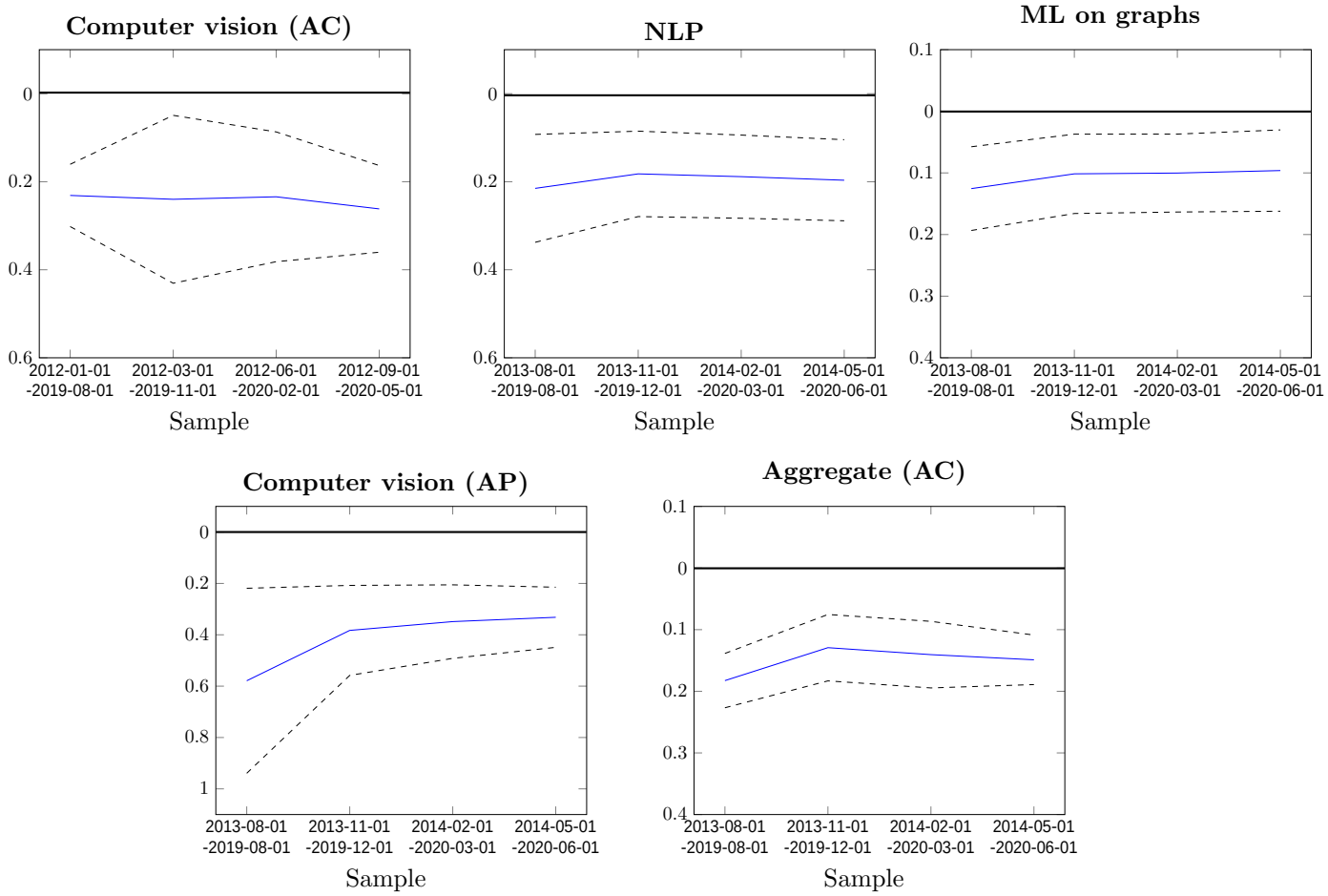
Figure 5: Estimates of σ over various sub-sample periods for the baseline regression model of Table 3.



Estimates of elasticity of performance based on various sample window selections. The solid blue lines represent point estimates, and dashed lines represent 95 pct con dence bounds.

Similarly, figure 6 shows the moving estimates for the coefficient $\tilde{\phi}$. The point estimate of $\tilde{\phi}$ for computer vision (AP) for the sub-sample starting in 2013-08-01 seems much lower than our earlier estimates—indicating a stronger negative intertemporal spillover effect. Besides this, neither our point estimates of $\tilde{\phi}$ nor the confidence bounds show much surprising variation relative to our earlier results.

Figure 6
Moving window estimation results for $\tilde{\phi}$ over various sub-sample periods in model presented in Table 3.



Estimates of $\tilde{\phi}$ for our baseline model (Table 3) based on various sample window selections. The solid blue lines represent point estimates, and dashed lines represent 95 pct con dence bounds.

6.2 Changing variable definitions

6.2.1 Robustness to adjustments for relative wages

Our earlier results were estimated using research input data which was adjusted for relative wages. Table 6 presents the estimation results of our error correction model estimates using non-wage adjusted data. The results are almost identical to our baseline results presented in table 3.

Table 6
Baseline estimation results from a cointegrated error correction model using machine learning performance and (non-quality adjusted) effective numbers of researchers over the 2012-01-01 to 2020-06-01 period

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Research effort	0.099*** (0.030)	0.011*** (0.004)	0.043 (0.033)	0.13 *** (0.047)	0.023*** (0.003)
Error correction term	0.23*** (0.06)	0.18*** (0.049)	0.12*** (0.022)	0.21*** (0.030)	0.15*** (0.020)
Disequilibrium half-life (months)	2.65 (0.79)	3.49 (1.05)	5.42 (1.06)	2.94 (0.47)	4.27 (0.62)
R^2 (adjusted)	0.78	0.83	0.74	0.81	0.61
Number of lags	$(1, 1)$	$(1, 1)$	$(1, 3)$	$(1, 1)$	$(1, 1)$
Number of cr. lags	0	0	3	3	3
Test for C-D dependence (p-value)	0.81	0.21	< 0.001	0.01	< 0.001
Observations	1133	819	1072	1367	2783
Benchmarks	26	16	21	30	61

6.2.2 Robustness to additional equipment costs

The data on the effective numbers of researchers used in our baseline estimations involve adjustments for additional equipment costs. Specifically, we assumed the cost was 10% percent of total labour input cost per publication. We replicate our baseline results presented in Table 3, with three changes in the approach to equipment costs (1) no adjustments, (2) a constant 50% of-labour-cost adjustment, and to account for a possible upward trend in equipment expenditure, (3) a 50% of-labour-cost adjustment that appreciates at 3% per year. We find that the estimates of each coefficient and their standard errors are almost identical across each of the three approaches to adjusting for equipment costs (see table 11 in the appendix).

7 Discussion

The empirical estimates obtained above reflect the presence of stark diminishing returns to research effort in each of the sub-fields of machine learning considered. An implication of this is a sharp decline in average research productivity—which we place between 4% to 26% per year. This result coincides with recent micro-level evidence of Bloom et al., 2020 that shows that research productivity seems to be declining in various industries, products, and firms. Consistent with their work, we show that it requires increasingly more research input to ensure constant exponential growth.

Nevertheless, the results obtained in the present work contrasts with those obtained Bloom et al., 2020, and it is primarily in the differences wherein this work's main contributions lie. Starting with the theoretical model, their work considers the standard fully endogenous variety of R&D-based growth model with the restriction that $\phi = 1$, whereas in our work, we consider a more general model that nests the fully endogenous type. Their work therefore arguably only provides evidence against a particular subset of growth models, namely the fully endogenous type. Our second contribution to the existing literature is that of applying the cointegrated error correction approach—an approach that has shown promise in the analysis of growth models using macro-data (such as in Ha and Howitt, 2007 and Kruse-Andersen 2017)—to the identification idea-production functions using panel micro-data. To do this, we have modified the approach to ensure robustness to cross-sectional dependence across our panels by employing Pesaran, 2006's dynamic common correlated effects approach for heterogeneous panel data models.

Our work does have various important limitations. Amongst these are mismeasurement issues, on both the output and input sides. On the research input side, the number of authors on research publications is likely to be imperfect measure of the number of researchers involved in R&D. When building our dataset, we try to be as careful to address this issue, by including a variety of publication types (such as peer-

reviewed articles, electronic pre-prints, conference proceedings, and more). However, insofar as the true number of researchers is systematically decoupled from the number of researchers that produce the types of output that we tracked, our estimates will be biased. In terms of our performance dataset, our dataset may be incomplete in some relevant way despite our best efforts at cross-validation. One plausible issue is that the data is incomplete: that there are machine learning models that are better than the reported state-of-the-art that are not reported, and that there are models that are noncompetitive that are not made public. The former would result in us understating the amount of progress in machine learning, and, in turn, overstating the extent to which there are diminishing marginal returns to research. In contrast, the latter mismeasurement issue may result in us deeming a benchmark obsolete despite there being active research effort dedicated to solving a task—which would result in an overstating of the amount of progress.

Relatedly, whilst there are many advantages to using improvements of performance on benchmark experiments as a measure of performance in machine learning (such as data availability, and ease of comparison), there are also some key disadvantages. For instance, a compelling argument is made by Blagec et al., 2020 that benchmark performance metrics often produce inadequate reflection of a model performance, and the level of performance may vary with arbitrary properties of the benchmark dataset. More generally, it should not be surprising that a single metric, or a small number of such metrics, may fail to capture some important aspects of the capacities of machine learning models. In light of such considerations, our work should be understood as an attempt to estimate one idea-production functions for the field of machine learning, amongst many others that could be explored.

One obvious source of bias in our estimates is the concurrent progress in the performance of hardware used to train machine learning models. We attempted to include relevant regressors for this—such as the number of parameters of a machine learning model—but failed to collect the requisite data to meaningfully control for hardware performance. We instead used time trends that may capture the effect of progress in hardware on progress in performance—which is an imperfect approach that may fail to eliminate bias. However, it should be noted that failure to partial out the effect of hardware progress on benchmark performance improvements will result in an overestimate of the growth of research productivity, as a part of the contributions of hardware improvements to machine learning progress are actually being attributed to researchers. Hence, it is likely that our estimates understate the extent of diminishing returns to research effort, and the rate of decline in average productivity.

Moreover, approach may fail to address an additional types of endogeneity in the modelled process; namely the potential simultaneity of progress and research effort. That is, the research effort allocation decisions may crucially depend on researchers' expectations of their levels of productivity—which in turn may be correlated with future levels of progress. Hence, our estimates might again understate the extent to which there are diminishing returns to research effort, as researchers will actively avoid areas they anticipate will not bear fruit. On the other hand, the fact that we observed strong consistent growth in research effort despite stark diminishing marginal returns suggests that this feedback effect is likely to be muted.

A Appendix: Supporting results

A.1 Unit root tests for performance variables

As is well known, correlation across errors in panels has serious drawbacks on commonly used panel unit root tests, since several of the existing tests assume that series are cross-sectionally independently distributed (Pesaran, 2007).

We therefore apply the panel unit root test procedure proposed in Pesaran, 2007 for testing unit roots in dynamic panels subject to (possibly) cross-sectionally dependent as well as serially correlated errors. The procedure involves augmenting the standard Augmented Dickey-Fuller regressions for the individual series with current and lagged cross-section averages of all the series in the panel (Pesaran, 2007).

Table 7. *p*-values for Pesaran, 2007 test for unit roots our performance series

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Number of lags	Constant				
1	0.97	0.32	1	1	0.99
2	0.81	0.62	1	1	0.92
3	0.85	0.64	1	1	0.98
4	0.99	0.21	1	0.99	1
	Constant and trend				
1	0.94	0.01	0.99	0.80	0.94
2	0.91	0.44	1	0.86	0.99
3	0.51	0.42	1	0.58	0.99
4	1	0.01	1	0.72	1
Observations	1146	819	1072	1367	2783
Benchmarks	26	16	21	30	61

Null hypothesis assumes that all series are non-stationary (large p-values indicates lack of evidence for stationarity).

A.2 Additional cointegration tests

We test for cointegration of the variables of interest using the residual-based cointegration test for dynamic panels advanced by Pedroni, 2004 and Westerlund, 2005. In our main results, we presented the test results for the cointegrating relationship,

$$A_{i,t} = \delta_i^{\emptyset} \mathbf{d}_{it} + \beta_i \tilde{S}_{it} + e_{it} \tag{24}$$

that includes a constant only. Table 8 presents the test statistics of the same tests that are run when the auxiliary regression (21) includes a time-trend. We find that our conclusion are broadly unchanged. In some cases we fail to reject the null of no cointegration using the variance ratio statistic, and the augmented Dickey-Fuller test. Yet, for these benchmark groups, the other test statistics do suggest that there is evidence for cointegration. For each benchmark group we reject the null with the Westerlund, 2005 test, which we prioritise due to better small-sample performance.

Table 8.
Test-statistics of a battery of tests for the cointegration of research effort and performance (with time trend in cointegrating relationship

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Variance ratio	22.8*** (2)	4.95*** (3)	3.43*** (0)	9.80*** (3)	80.62*** (2)
Modified Phillips-Perron t	5.44*** (2)	1.30* (3)	12.87*** (0)	1.42* (3)	18.22*** (2)
Phillips-Perron t	1.93 (2)	0.34 (3)	7.77*** (0)	2.57*** (3)	0.84 (2)
Augmented Dickey-Fuller t	1.29* (2)	0.22 (3)	6.54*** (0)	2.60*** (3)	2.99*** (2)
Variance ratio (West)	5.20***	2.34***	4.08***	2.75**	10.39***
Number of panels	25	16	20	30	62
Average number of periods	45.46	52.81	49.1	48.03	48.53

A.3 Additional error correction model estimates

Table 9. baseline error correction model with time-trend

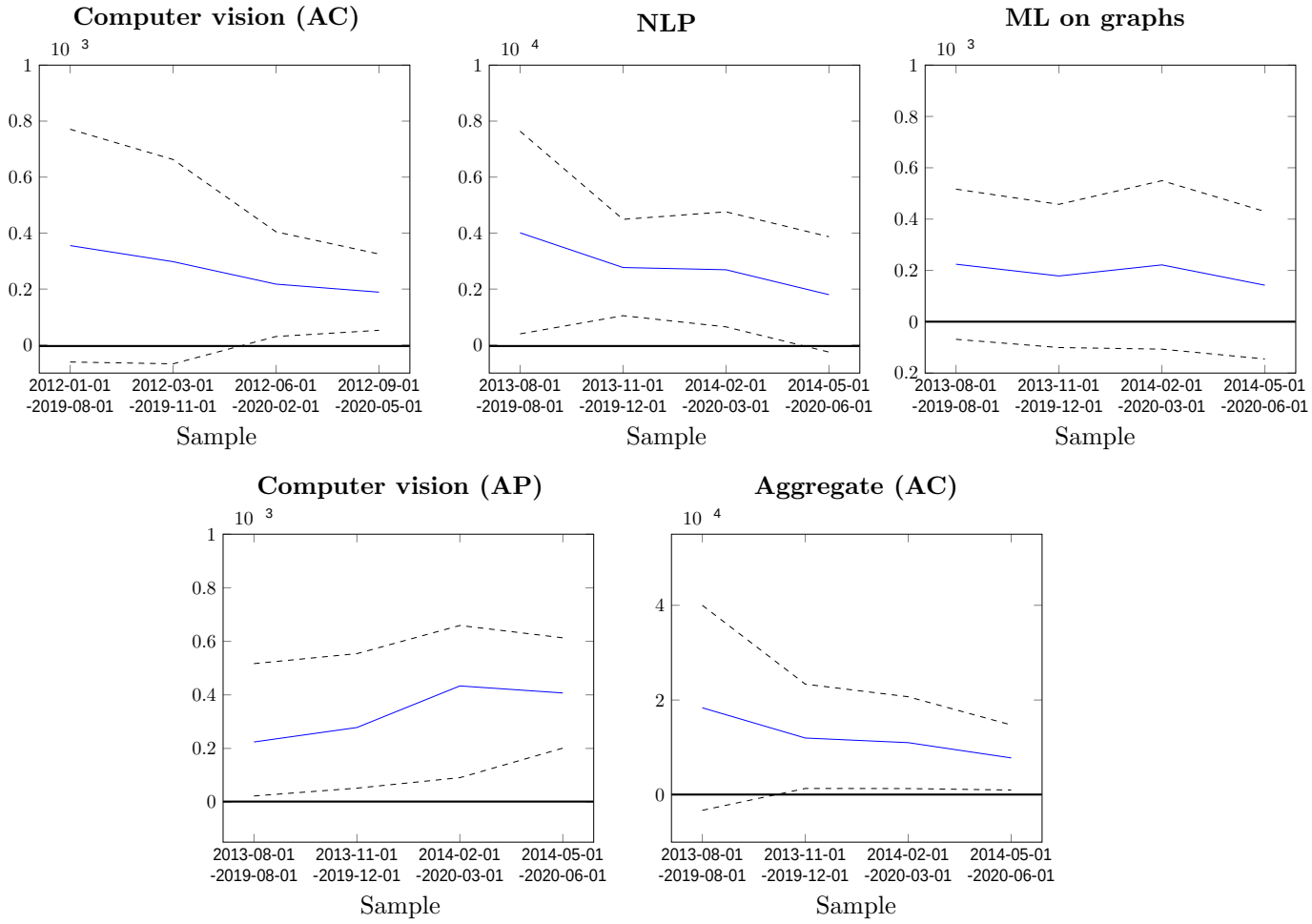
	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Effective research effort	.068*** (.027)	.016** (.008)	.021** (.011)	.089** (.043)	.019*** (.007)
Yearly trend	.001 *** (.0004)	.0002*** (.0001)	.0001 (.0001)	.003*** (.001)	.0006** (.0003)
Performance (level)	.31*** (0.06)	0.30*** (.058)	.070*** (.44)	.48*** (.078)	.21*** (.023)
Disequilibrium half-life (months)	1.87*** (.43)	1.94*** (.45)	62.3*** (9.55)	1.05*** (.024)	2.94*** (0.36)
R^2 (adjusted)	0.78	0.78	0.78	0.77	0.60
Number of lags	$(1, 1)$	$(1, 1)$	$(1, 3)$	$(1, 1)$	$(1, 1)$
Number of cr. lags	0	0	3	3	3
Test for C-D dependence (p-value)	0.84	0.24	< 0.001	0.01	< 0.001
Observations	1146	819	1072	1367	2783
Benchmarks	26	16	21	30	61

Table 10.
Estimation results of model with time-varying performance elasticity to research effort (no quality adjusted research input data)

	Benchmark group				
	Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
Log effective research effort month	.00016*** (.00005)	.00003 *** (.000008)	.00002 (.0001)	.0004*** (.0001)	.0009*** (.0003)
Log research effort	.051 (.046)	.006 (.009)	.076 (.089)	.20* (.11)	.040 (.025)
Performance (level)	31*** (.05)	.28*** (.06)	.18 (.036)	.48 (.082)	.21*** (.022)
R^2 (adjusted)	0.76	0.80	0.63	0.76	0.57
Number of lags	$(1, 1)$	$(1, 1)$	$(1, 3)$	$(1, 1)$	$(1, 1)$
Number of cr. lags	0	0	3	3	3
Test for C-D dependence (p-value)	0.75	0.26	< 0.001	0.02	< 0.001
Observations	1133	819	1072	1367	2783
Benchmarks	26	16	21	30	61

A.4 Moving window estimation results: time-varying elasticities

Figure 7: Estimates of trend in performance elasticity to research effort over various sub-sample periods in model presented in Table 3.



Estimates of performance elasticity to e ective research e ort based on various sample window selections. The solid blue lines represent point estimates, and dashed lines represent 95 pct con dence bounds.

A.5 Equipment cost robustness check

Table 11.
Replication of baseline results presented in Table 3, with three changes in the approach to equipment costs.

		Benchmark group				
Equip. cost adjustments		Computer vision (AC)	NLP (AC)	Graphs (AC)	Computer vision (AP)	Aggregate (AC)
None	Log research effort	.10 *** (.029)	.01 *** (.004)	.02 (.09)	.13 *** (.047)	.02 ** (.001)
	Performance (level)	.22** (.057)	.18*** (.049)	.12*** (.023)	.22*** (.031)	.15** (.020)
	R^2 (adjusted)	0.80	0.82	0.75	0.82	0.61
	Test for CD dep. (p-value)	0.70	0.21	< 0.001	0.01	< 0.001
$f = 0.5$	Log research effort	.10 *** (.030)	.01 *** (.004)	.05 * (.03)	.12 *** (.05)	.02 ** (.01)
	Performance (level)	.22 (.057)	.18*** (.049)	.13*** (.02)	.21*** (.03)	.15*** (.01)
	R^2 (adjusted)	0.80	0.83	0.75	0.82	0.61
	Test for CD dep. (p-value)	0.71	0.21	< 0.001	0.01	< 0.001
$f = 0.5$ + 3% p.a.	Log research effort	.10 *** (.03)	.01 *** (.004)	.05 * (.03)	.13 *** (.046)	.02 ** (.01)
	Performance (level)	.22*** (.06)	.18*** (.049)	.12*** (.02)	.21*** (.03)	.15*** (.02)
	R^2 (adjusted)	0.80	0.83	0.75	0.61	0.61
	Test for CD dep. (p-value)	0.71	0.21	< 0.001	0.01	< 0.001
Model/ estimation details	Number of lags	$(1, 1)$	$(1, 1)$	$(1, 3)$	$(1, 1)$	$(1, 1)$
	Number of cr. lags	0	0	3	3	3
	Nr. of observations	1133	819	1072	1367	2783
	Benchmarks	26	16	21	30	61

Replication of baseline results presented in Table 3, with three changes in the approach to equipment costs (1) no adjustments (2) a constant 50% of-labour-cost adjustment, and to account for a possible upward trend in equipment expenditure, (3) a 50% of-labour-cost adjustment that appreciates at 3% per year.

B Appendix: Additional details

B.1 Publication data collection methodology

To identify the publications in the relevant sub-fields of machine learning, we performed a keyword search, as is common in the study of bibliometrics (see, e.g. Andrés, 2009).

For natural language processing and computer vision, the relevant keyword queries were constructed to be in line with The Association for Computing Machinery (ACM) Computing Classification System (ACM, 1998). The ACM Computing Classification System (CCS) is a hierarchical classification system used to index and classify major areas and topics of the fields of computing and computer science (Lin et al., 2012). For machine learning on graphs, the keyword queries were constructed to be in line with the taxonomy by Chami et al., 2020. Table 12 presents a detailed description of the queries.

Table 12
Keyword queries used to identify the relevant publications in the relevant sub-fields of machine learning on arXiv and WOS

Sub-field	Query based on	Keywords included	Fields covered (WOS)	Domains covered (arXiv)
Computer vision	ACM Subject Classes I.2.10, I.4, and I.5.	Image processing; computer vision; pattern recognition; scene understanding; image segmentation; scene analysis; image representation	All computer science categories; probability and statistics	Computer science; statistics
Natural language processing	ACM Subject Classes I.2.7, excludes work on formal languages (programming languages, logics, formal systems)	Natural-language processing; language generation; language models; language parsing; language understanding machine translation speech recognition; speech synthesis; text analysis	All computer science categories; probability and statistics	Computer science; statistics
Machine learning on graphs	Based on taxonomy by Chami et al., 2020	Node classification; graph auto-encoder; network embedding; graph embedding; graph regularization; graph neural networks; graph classification; graph representation learning; graph convolutional networks; community detection; graph attention networks; node classification; link prediction	All computer science categories; probability and statistics	Computer science; statistics

B.2 Measures of performance

B.2.1 Average precision

Consider a classifier that, for any given input (such as an image), generates a list of N attributes, ranked in descending order of predictive probability. Let x_k be a variable takes a 1 if attribute k applies and a 0 if attribute k fails to apply. For any k N precision is defined as follows (Kishida, 2005):

$$\text{precision@}k = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{25}$$

In turn, average precision is defined as:

$$\text{average precision} = \frac{1}{k}\sum_{k} \text{precision@}k \tag{26}$$

Finally, mean average precision is defined as the arithmetic mean of average precision over the total number of problems.

Accuracy and related classification evaluations

Accuracy and related measures are commonly used to evaluate classifiers. These measures are generally defined as some ratio of the total of instances that are correctly predicted by the trained classifier when tested with the unseen data (Hossin and Sulaiman, 2015).

B.2.2 Accuracy in classification

Consider a classifier that produces a 1 N vector of attributes, and a corresponding 1 N vector of predicted probabilities of the attribute matching. A prediction is then defined as a 1 K vector that is constructed by taking the k highest probability attributes, with $0 < k$ N. A correct prediction is said to have been made when the true attribute is an element of the prediction vector.

$$\text{Top-}k \text{ Accuracy} = \frac{\#\text{Correct predictions}}{\#\text{Total predictions}} \tag{27}$$

Error rate

Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated:

$$\text{Error Rate} = \frac{\#\text{Incorrect predictions}}{\#\text{Total predictions}} \tag{28}$$

B.3 Mathematical details

B.3.1 Disequilibrium half-life

The speed of adjustment in an error correction model is often measured by the half-life, that is, the time needed in order to eliminate 50% of the deviation. We report the half-life, rather than the error-correction coefficient, as, for most, the former has a more natural interpretation.

The half-life h is given as follows: $h = \frac{\ln 2}{\ln(1+\tilde{\phi})}$, for any $\tilde{\phi}$ 2 (1, 0). To see this, consider an error correction model of the sort:

$$\Delta Y_{t+1} = \tilde{\phi}(\underbrace{Y_t \quad \mathbf{X_t}\hat{\beta}}_{e_t \quad 1}) + \mathbf{Z_t}\alpha + \epsilon_t, \tag{29}$$

where $\hat{\beta}$ are estimated coefficients, Y_{t+1} is a weakly stationary process, and $\mathbf{X}_t$ and $\mathbf{Z}_t$ are matrices of random variables. Moreover, the process is error-correcting, i.e. the error correction term satisfies 1 $< \tilde{\phi} < 0$. Then, this can be represented in its autoregressive form:

$$Y_{t+1} = (1 + \tilde{\phi})Y_t \quad \tilde{\phi}\beta\mathbf{X}_t + \alpha\mathbf{Z_t} + \epsilon_t \tag{30}$$

Given the weak-stationarity of Y_t, it is true that $\mathbb{E}[Y_t] = \mu$, for some μ 2 $\mathbb{R}$. The half-life, in the context of a shock, is defined as the amount of time, denoted h, until the process is expected to halve its distance to the stationary mean.

$$\mathbb{E}[Y_{t+h}] = \frac{1}{2}Y_t \tag{31}$$

$$= (1 + \tilde{\phi})^h Y_t. \tag{32}$$

$$( ) \quad h = \frac{\ln 2}{\ln(1 + \tilde{\phi})}, \text{ for any } \tilde{\phi} \text{ 2 } (\quad 1, 0). \tag{33}$$

We compute the standard error as follows. First, we consider the variance of h

$$\text{Var}(h) = \text{Var}\left(\frac{\ln 2}{\ln(1 + \tilde{\phi})}\right) = \ln^2 2 \text{Var}\left(\frac{1}{\ln(1 + \tilde{\phi})}\right) \tag{34}$$

Using a first order Taylor approximation, this is approximately

$$\ln^2 2\left(\frac{1}{(\bar{\phi} + 1)\ln^2(\bar{\phi} + 1)}\right)^2 \sigma^2_{\tilde{\phi}} \tag{35}$$

where $\bar{\phi}$ is the mean-group coefficient, and $\sigma^2_{\tilde{\phi}}$ is the variance of $\tilde{\phi}$. Thus,

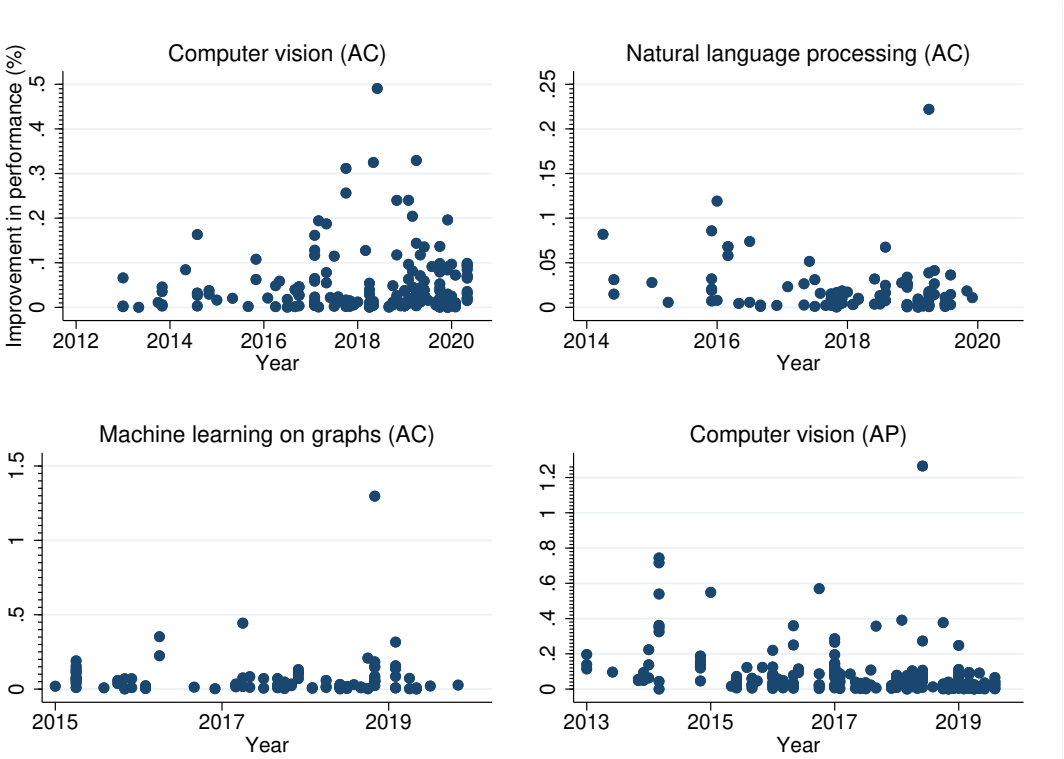
$$se(\phi) \quad \frac{\ln 2\left(\frac{1}{(\phi+1)\ln^2(\phi+1)}\right)\sigma_{\tilde{\phi}}}{p\overline{N}}. \tag{36}$$

Since $\tilde{\phi}$ is estimated using a mean group estimator on Stata using the `xtdcce2` package, we can compute the standard error (36) directly.

C Additional figures

C.1 Plot of performance improvements

Figure 8.
Performance improvements across the fields of computer vision, natural language processing and machine learning on graphs



Scatterplot of monthly improvements in top-performing performance over time across benchmarks for the sub-categories of computer vision, natural language processing and machine learning on graphs. Improvements of 0 are excluded from this figure. AC = Accuracy, AP = Average Precision.

Table 13
List of all machine learning benchmarks used

Computer vision	Natural Language Processing	Machine learning on graphs
ImageNet	SST-2 Binary classi cation	SNLI
CIFAR-10	SST-5 Fine-grained classi cation	Citeseer
CIFAR-100	IMDb	Pubmed
MNIST	Yelp Binary classi cation	CiteSeer with Public Split xed 20 nodes per class
SVHN	Yelp Fine-grained classi cation	Cora with Public Split: xed 20 nodes per class
STL-15	SemEval 2014 Laptop	PubMed with Public Split: xed 20 nodes per class
Mini-Imagenet 5-way (1-shot)	SemEval 2014 Rest	Cora (0.5%)
OMNIGLOT - 5-Shot, 20-way	MOSI	Cora (1%)
OMNIGLOT - 1-Shot, 5-way	DBpedia	Cora (3%)
Tiered ImageNet 5-way (5-shot)	TREC-6	CiteSeer (0.5%)
CUB 200 5-way 1-shot	Text Classi cation on IMDb	CiteSeer (1%)
CUB 200 5-way 5-shot	Yahoo! Answers	PubMed (0.03%)
CIFAR-FS 5-way (1-shot)	R8	PubMed (0.1%)
FC100 5-way (5-shot)	Cora	BlogCatalog
Mini-Imagenet 5-way (10-shot)	AG News	Reddit
Stanford Dogs 5-way (5-shot)		Cora Full-supervised
Stanford Cars 5-way (1-shot)		PubMed (0.5%)
ImageNet - 10% labeled data		NCI1
ImageNet - 1% labeled data		D&D
CIFAR-10, 4000 Labels		PTC
SVHN, 1000 labels		Citeseer
cifar-100, 10000 Labels		PROTEINS
STL-10, 1000 Labels		
Fine-grained Stanford Cars		
Fine-grained CUB-200-2011		
FGVC Aircraft		
COCO test-dev BOXAP (*)		
COCO test-dev AP50 (*)		
COCO test-dev AP75 (*)		
COCO minival AP50 (*)		
COCO minival APL (*)		
PASCAL VOC 2007 (*)		
SUN-RGBD val (*)		
Weakly Supervised PASCAL VOC 2007 (*)		
PASCAL VOC 2012 test (*)		
Car detection Easy (*)		
Car 3D detection Easy (*)		
Car 3D detection Moderate (*)		
CAR BEV Easy (*)		
CAR BEV Moderate (*)		
CAR BEV Hard (*)		
Pedestrian detection Moderate (*)		
Pedestrian detection Hard (*)		
Cyclist detection Easy (*)		
Cyclist detection Moderate (*)		
Pedestrian 3D Detection Moderate (*)		
Pedestrian 3D Detection Hard (*)		
Pedestrian BEV Easy (*)		
Pedestrian BEV Moderate (*)		
Pedestrian BEV Hard (*)		
Cyclist 3D detection Easy (*)		
Cyclist 3D detection Moderate (*)		
Cyclist 3D detection Hard (*)		
Cyclist BEV Easy (*)		
Cyclist BEV Moderate (*)		
Cyclist BEV Hard (*)		

(*) indicates that model performance is measured in average precision. For all remaining benchmarks, model performance is measured in accuracy or error.

References

G. Abramo, C. D'Angelo, and A. Caprasecca. Gender differences in research productivity: A bibliometric analysis of the italian academic system. *Scientometrics*, 79(3):517–539, 2009.

D. Acemoglu and P. Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542, 2018.

D. Acemoglu, P. Aghion, L. Bursztyn, and D. Hemous. The environment and directed technical change. *American economic review*, 102(1):131–66, 2012.

ACM. Acm computing classification system [1998 version]. *ACM*, 1998.

P. Aghion. Schumpeterian growth theory and the dynamics of income inequality. *Econometrica*, 70(3): 855–882, 2002.

P. Aghion and P. Howitt. A model of growth through creative destruction. Technical report, National Bureau of Economic Research, 1990.

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

P. Aghion, L. Ljungqvist, P. Howitt, P. W. Howitt, M. Brant-Collett, C. García-Peñalosa, et al. *Endogenous growth theory*. MIT press, 1998.

P. Aghion, B. F. Jones, and C. I. Jones. Artificial intelligence and economic growth. Technical report, National Bureau of Economic Research, 2017.

A. Agrawal, J. McHale, and A. Oettl. Finding needles in haystacks: Artificial intelligence and recombinant growth. Technical report, National Bureau of Economic Research, 2018.

D. W. Aksnes, L. Langfeldt, and P. Wouters. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1):2158244019829575, 2019.

N. Amano-Patiño, E. Faraglia, C. Giannitsarou, Z. Hasna, et al. The unequal effects of covid-19 on economists' research productivity. Technical report, Faculty of Economics, University of Cambridge, 2020.

A. Andrés. *Measuring academic research: How to undertake a bibliometric study*. Elsevier, 2009.

J. Bar-Ilan. Informetrics at the beginning of the 21st century—a review. *Journal of informetrics*, 2(1):1–52, 2008.

P. Barredo, J. Hernández-Orallo, F. Martınez-Plumed, and S. h Éigeartaigh. The scientometrics of ai benchmarks: Unveiling the underlying mechanics of ai research. *Evaluating Progress in Arti cial Intelligence (EPAI 2020). ECAI*, 2020.

C. Baum and M. Schaffer. Actest: Stata module to perform cumby-huizinga general test for autocorrelation in time series. 2015.

E. F. Blackburne III and M. W. Frank. Estimation of nonstationary heterogeneous panels. *The Stata Journal*, 7(2):197–208, 2007.

K. Blagec, G. Dorffner, M. Moradi, and M. Samwald. A critical analysis of metrics used for measuring progress in artificial intelligence. *arXiv preprint arXiv:2008.02577*, 2020.

N. Bloom, C. I. Jones, J. Van Reenen, and M. Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–44, 2020.

T. Braun, W. Glänzel, and A. Schubert. Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, 51(3):499–510, 2001.

J. Breitung. Nonparametric tests for unit roots and cointegration. *Journal of econometrics*, 108(2):343–363, 2002.

E. Cambria and B. White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

N. Carayol and M. Matt. Does research organization influence academic production?: Laboratory level evidence from a large european university. *Research Policy*, 33(8):1081–1102, 2004.

I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy. Machine learning on graphs: A model and comprehensive taxonomy. *arXiv preprint arXiv:2005.03675*, 2020.

A. C. Chu. Effects of patent length on r&d: a quantitative dge analysis. *Journal of Economics*, 99(2): 117–140, 2010.

A. Chudik and M. H. Pesaran. Large panel data models with cross-sectional dependence: a survey. *CAFE Research Paper*, (13.15), 2013.

A. Chudik and M. H. Pesaran. Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*, 188(2):393–420, 2015.

R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice. Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability*, 12(2):492, 2020.

R. E. Cumby and J. Huizinga. Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. Technical report, National Bureau of Economic Research, 1990.

C. D'Adda and A. E. Scorcu. On the time stability of the output–capital ratio. *Economic Modelling*, 20(6): 1175–1189, 2003.

M. De Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, 2016.

M. de Kleijn, M. Siebert, and S. Huggett. Artificial intelligence: how knowledge is created, transferred and used. 2017.

R. C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

J. Ditzen. Estimating dynamic common-correlated effects in stata. *The Stata Journal*, 18(3):585–617, 2018.

E.-S. M. El-Alfy and S. A. Mohammed. A review of machine learning for big data analytics: bibliometric approach. *Technology Analysis & Strategic Management*, pages 1–22, 2020.

Engle and Granger. Engle, rf granger, cwj (1987). *Co-integration and error correction: Representation, estimation, and testing. Econometrica: Journal of the Econometric Society*, 55:251–276, 1987.

R. E. Evenson. Patents, r&d, and invention potential: International evidence. *The American Economic Review*, 83(2):463–468, 1993.

A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

M. A. Gomez. Duplication externalities in an endogenous growth model with physical capital, human capital, and r&d. *Economic Modelling*, 28(1-2):181–187, 2011.

K. Grace. Algorithmic progress in six domains. Technical report, Technical report, Machine Intelligence Research Institute, 2013.

G. Graetz, G. Michaels, et al. Robots at work: the impact on productivity and jobs. Technical report, Centre for Economic Performance, LSE, 2015.

Z. Griliches. Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press, 1998.

G. M. Grossman and E. Helpman. Quality ladders in the theory of growth. *The review of economic studies*, 58(1):43–61, 1991.

J. Ha and P. Howitt. Accounting for trends in productivity and r&d: A schumpeterian critique of semi-endogenous growth theory. *Journal of Money, Credit and Banking*, 39(4):733–774, 2007.

G. Halevi. Bibliometric studies on gender disparities in science. In *Springer handbook of science and technology indicators*, pages 563–580. Springer, 2019.

D. Hernandez and T. B. Brown. Measuring the algorithmic efficiency of neural networks. *arXiv preprint arXiv:2005.04305*, 2020.

M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

T. Hothorn, F. Leisch, A. Zeileis, and K. Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

ILO. Ilostat, Aug 2020. URL `https://ilostat.ilo.org/`.

Jones. R & d-based models of economic growth. *Journal of political Economy*, 103(4):759–784, 1995.

C. I. Jones. The facts of economic growth. In *Handbook of macroeconomics*, volume 2, pages 3–69. Elsevier, 2016.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255–260, 2015.

K. Kishida. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005.

S. Kortum. Equilibrium r&d and the patent–r&d ratio: Us evidence. *The American Economic Review*, 83 (2):450–457, 1993.

S. Kortum and J. Lerner. Stronger protection or technological revolution: what is behind the recent surge in patenting? In *Carnegie-Rochester Conference Series on Public Policy*, volume 48, pages 247–304. Elsevier, 1998.

S. S. Kortum. Research, patenting, and technological change. *Econometrica: Journal of the Econometric Society*, pages 1389–1419, 1997.

D. C. Kozen. *The design and analysis of algorithms*. Springer Science & Business Media, 2012.

P. Kruse-Andersen. Testing r&d-based endogenous growth models. *Available at SSRN 2947528*, 2017.

J. O. Lanjouw and M. Schankerman. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465, 2004.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He. A text classification survey: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*, 2020.

X. Lin, M. Zhang, H. Zhao, and J. Buzydlowski. Multi-view of the acm classification system. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 397–398, 2012.

H. Lloyd-Ellis. Endogenous technological change and wage inequality. *American Economic Review*, 89(1): 47–77, 1999.

D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

W. K. Newey and K. D. West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994.

W. D. Nordhaus. Are we approaching an economic singularity? information technology and the future of economic growth. Technical report, National Bureau of Economic Research, 2015.

P. Pedroni. Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the ppp hypothesis. *Econometric theory*, pages 597–625, 2004.

R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, and J. C. Niebles. The ai index 2019 annual report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*, 2019.

M. H. Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006.

M. H. Pesaran. A simple panel unit root test in the presence of cross-section dependence. *Journal of applied econometrics*, 22(2):265–312, 2007.

M. H. Pesaran. Testing weak cross-sectional dependence in large panels. *Econometric reviews*, 34(6-10): 1089–1117, 2015.

K. Prettner and H. Strulik. The lost race against the machine: Automation, education, and inequality in an r&d-based growth model. *Education, and Inequality in an R&D-Based Growth Model (December 1, 2017). cege Discussion Papers*, (329), 2017.

P. Ramsden. Describing and explaining research productivity. *Higher education*, 28(2):207–226, 1994.

W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

P. M. Romer. Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102, 1990.

G. Rong, A. Mendez, E. B. Assi, B. Zhao, and M. Sawan. Artificial intelligence in healthcare: Review and prediction case studies. *Engineering*, 6(3):291–301, 2020.

T. N. Sequeira and P. C. Neves. Stepping on toes in the production of knowledge: a meta-regression analysis. *Applied Economics*, 52(3):260–274, 2020.

L. Shiell and N. Lyssenko. Climate policy and induced r&d: How great is the effect? *Energy economics*, 46: 279–294, 2014.

Y. Shoham, R. Perrault, E. Brynjolfsson, J. Clark, J. Manyika, J. C. Niebles, T. Lyons, J. Etchemendy, B. Grosz, and Z. Bauer. The ai index 2018 annual report. *AI Index Steering Committee, Human-Centered AI Initiative*, 2018.

H. Strulik. Too much of a good thing? the quantitative economics of r&d-driven growth revisited. *Scandinavian Journal of Economics*, 109(2):369–386, 2007.

S. Sun, C. Luo, and J. Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017.

C. Sutton and L. Gong. Popularity of arxiv. org within computer science. *arXiv preprint arXiv:1710.05225*, 2017.

N. von Tunzelmann, M. Ranga, B. Martin, and A. Geuna. The effects of size on research performance: A spru review. *Report prepared for the O ce of Science and Technology, Department of Trade and Industry*, 2003.

M. Webb, N. Short, N. Bloom, and J. Lerner. Some facts of high-tech patenting. Technical report, National Bureau of Economic Research, 2018.

J. Westerlund. New simple tests for panel cointegration. *Econometric Reviews*, 24(3):297–316, 2005.

G. N. Yannakakis and J. Togelius. *Arti cial Intelligence and Games*. Springer, 2018. `http://gameaibook.org`.