Using Google Search Trends to Estimate Global Patterns in Learning

Serhat Arslan, Mo Tiwari, Chris Piech
Stanford University, Computer Science
sarslan@stanford.edu, motiwari@stanford.edu, piech@cs.stanford.edu

ABSTRACT
The use of the Internet for learning provides a unique and growing opportunity to revisit the task of quantifying what people learn about a given subject in different regions around the world. Google alone receives over 5 billion searches a day, and its publicly available data provides insight into the learning process that is otherwise unobservable on a global scale. In this paper, we introduce the Computer Science Literacy-proxy Index via Search (CSLI-s), a measure that utilizes online search data to estimate trends in computer science education. This measure uses statistical signal processing techniques to aggregate search volumes from a spectrum of topics into a coherent score. We intentionally explore and mitigate the biases of search data and, in the process, develop CSLI-s scores that correlate with traditional, more expensive metrics of learning. Furthermore, we use search trend data to measure patterns in subject literacy across countries and over time. To the best of our knowledge, this is the first measure of topical literacy via Internet search trends. The Internet is becoming a growing tool for learners and, as such, we anticipate search trend data will have growing relevance to the learning science community.

Author Keywords
Google Search Trends, measuring quality education, Informal education, curricula patterns

1. INTRODUCTION

Improving education is an implicit objective of scientific communities like Learning at Scale, as well as other international institutions such as the United Nations. The United Nations Sustainable Development Goal (SDG) 4 is to "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all." Quantifying progress towards this goal, however, has remained a difficult task – especially with regards to the *quality* of learning.

The problem of measuring education quality becomes especially challenging when we consider rapidly changing topics such as computer science and climate change. Computer science is more popular every year and is reaching an increasingly diverse set of students [11]. Similarly, the rapidly-evolving climate crisis demands frequent revision of climate change curricula.

Progress on SDG 4 has proven to so difficult to quantify that the United Nations has revisited whether it was an appropriate SDG at all [17]. Systematic reviews to understand how curricula are taught, and to whom, have been conducted [27, 4, 20]. However, these reviews are expensive to administer around the world, slow to respond to new curricula, and only include classroom education (thereby neglecting informal learning). Though often overlooked, informal learning is a crucial part of the overall fabric of education [8]. Furthermore, the number of informal learners and their progress are not decisively available, in contrast to class occupancy and examination scores in formal institutions. As a result, global exams such as PISA are unsatisfactory measures of learning. Such exams are administered in a small handful of countries, only test core concepts like literacy and numeracy in high school students, and are unable to answer questions about educational topics in the population as a whole.

In contrast with the methods described above, Internet search data provides a unique opportunity to understand global education. Google receives over 5 billion search queries a day, which is roughly 1 search per person per day globally. The overall search trend data is public and free. As the most popular search engine, it presents an opportunity to learn about tendencies and quantity of questions people have in different regions. In the field of disease modelling, Google searches are successfully used to track the spread of influenza, which is a loose indication that search data might prove useful for tracking the spread of education [12].

As a motivating example, consider a user searching "How to determine k in k-means clustering?" This user, by conducting this search, signals curiosity which is largely unique to when one is *learning*, *teaching*, or *practicing* artificial intelligence (AI). When this search is executed, the search data then contains a signal that this particular user is “AI-literate" or gaining AI-literacy. This single search tells us even more: we can observe if the search was during school session and its temporal relationship to other AI-related queries from the user.

One search from an individual, taken alone, may not paint a very convincing picture of CS literacy worldwide. However,

L@S '20, August 12–14, 2020, Virtual Event, USA.

ACM ISBN 978-1-4503-7951-9/20/08 ...$15.00.
http://dx.doi.org/10.1145/3386527.3405913

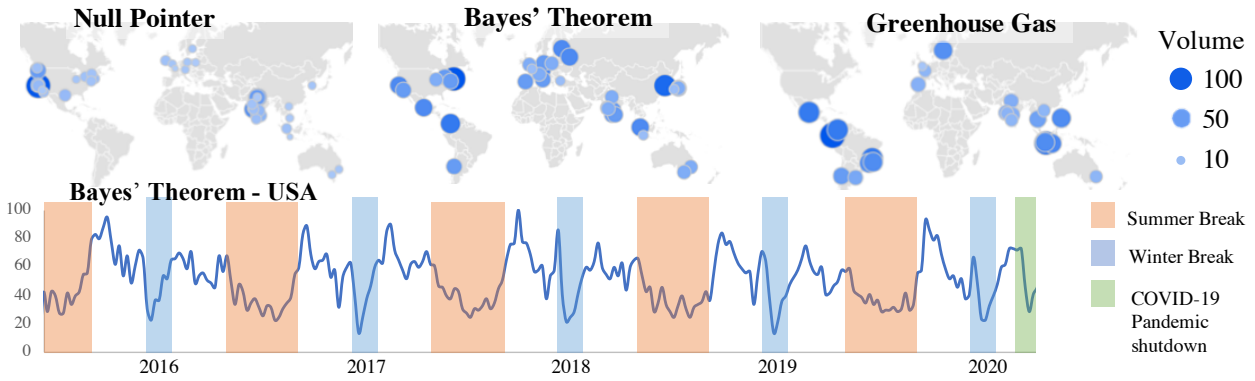

Figure 1. Visualization of data from Google search trends. Top Row: Cities with highest relative search volume for different topics, normalized. Greenhouse Gas, Bayes' Theorem, and Null Pointers are keywords related to climate change, artificial intelligence, and computer science, respectively. Bottom Row: Temporal patterns for a search topic volume over 5 years in the USA, overlaid with school holidays and breaks.

when aggregated over millions of searches relating to artificial intelligence, consistent trends emerge that tell a broader story of learning in a given population. Figure 1 visualizes the potential; there are clear patterns in the geography and timing of Google searches. For example, there are seasonal patterns for the keyword "Bayes' Theorem": there are substantial dips in search frequency when school is not in session (including in the summer, when work in industry continues). We further observe that the dip in searches during the COVID-19 pandemic corresponds to the shutdown of schools that took place across the United States. Though this dip is uncharacteristically large for the season, interest in Bayes' Theorem seems to quickly resurge as students begin to learn from home.

Additional examples of search term frequencies are demonstrated in Figure 2. Each of the three plots shows the relative popularity of a computer science topic over five years in the United States, overlaid with school holidays (orange). The first topic, "Random Variables", is largely learned only during the school year. As such, it demonstrates a large decrease in search frequency during the summer, winter, and Thanksgiving breaks. The second example, "Markov Chain", demonstrates a similar seasonal pattern, but the ratio of searches in the school year to searches during the summer is smaller. The last example, "Machine Learning", demonstrates dips only during the winter (i.e. Christmas) holiday. This suggests that "Machine Learning" is a concept which is searched for largely outside of school, potentially in the workforce.

This observation of seasonal patterns in learning-related search data suggests the existence of signal in search trend data that captures the utility of the search term both in formal school environments and informal environments outside of school [8]. Indeed, this observation provided significant motivation for our work. We believe that, individually, no single search term measures literacy of computer science. As such, we investigate the *composition* of signals for a variety of search terms to understand subject literacy.

More concretely, we propose the following research challenge: **how can we use Internet search trends to measure the literacy of a population with respect to a certain subject?** In this paper, our main contributions are to:

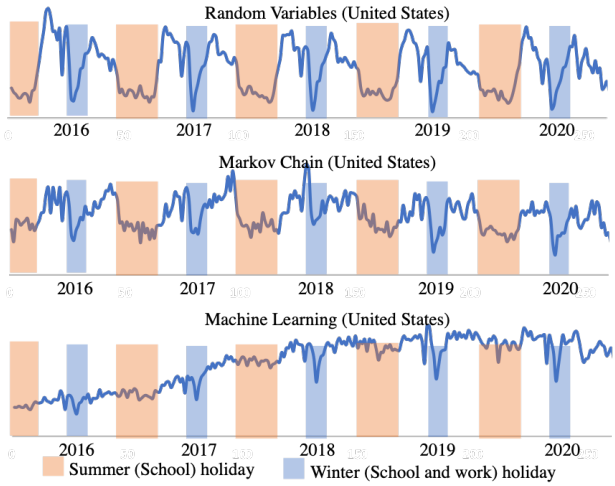

Figure 2. Search terms exhibit strong, consistent seasonal patterns but to different degrees. Each graph is relative search volume over time in the United States.

1. Pose the challenge of measuring depth of literacy via Internet search data and articulate the relevant biases of such data,
2. Identify strong seasonal patterns in education related search trend data that suggest the data is correlated with literacy,
3. Introduce a new measure, Computer Science Literacy-proxy Index bu Search (CSLI-s), for extracting a measure of computer science literacy from raw search statistics, which shows notable correlation with more expensive proxies,
4. Show corollary results including (i) uncovering trends across time and between countries and (ii) uncovering curricula patterns amongst countries,
5. Open-source our code at: `compedu.stanford.edu/googleTrends`.

Our results indicate that understanding subject literacy from Google search data is a fast and free method to gain insight into global education, especially for domains that require a computer like computer science, and for geographical regions where there are no other known measures.

We add an important cautionary note: though this data is large in quantity, it is not the case that "big data" is a panacea for challenges in estimating literacy [25]. In our analysis in Section 2, we provide a theoretical framework for articulating and addressing the several confounding variables including, but not limited to, the sampling bias in search data. It is hard, but not impossible, to understand and work around these challenges. We discuss the limitations of measuring subject literacy from search, but observe that despite the known challenges, it allows insight into previously unseen trends.

The analysis we provide is a proof of concept which could be used to analyze more fine-grained educational topics (such as human-computer interaction learning or AI learning) and evaluating learning for other curricula, especially for domains that require a computer. We also hope that insights in this work provide inspiration measures of learning that distinguish formal and informal learning. Above all, we believe this research is a first step in new research directions for quantifying global learning on a per-topic and per-region basis over time.

1.1. Related Work

1.1.1. Measuring Learning

Measuring learning has been commonly studied in formal education [9, 31, 41] and for informal learning [19, 45]. In the domain of informal learning, the resources an individual student uses is referred to as their Personal Learning Environments (PLE) [7]. While PLEs promote learning, they make informal learning harder to measure due to limitations on access to private user data. Noting the dominance of online platforms over PLEs, [1] provided detailed theory on online learning. Several studies developed in response to the rise of massive open-access online courses (MOOC). The first MOOCs were given by the CS departments of research universities such as Stanford, MIT, Harvard, and Berkeley. In 2012, millions of students signed up for online courses on Coursera, EdX, and Udacity; with relatively little cost to entry or exit, MOOCs attracted learners with a wide range of backgrounds, intentions, or personal constraints to participation [35]. The New York Times declared 2012 the "Year of the MOOC" [35] and researchers sought to classify and quantify the informal learners who were choosing these platforms [24, 22]. Since 2012, specialized platforms in CS which are open-access to formal and informal learners have developed such as Code.org, CodeAcademy, and KhanAcademy. The ways in which informal learners use these platforms has become a large area of research [36, 30, 38, 23].

1.1.2. Search Data for Science

Data from search engines has inspired researchers mostly to estimate or forecast economic indicators [5, 29, 10, 3, 18, 14], e.g. unemployment or inflation rates. [39] uses real time search data to estimate political tendencies of voters in a region, and show correlations with election statistics. Similarly, [37] and [12] predict the spread of influenza by using search data from Yahoo and Google, respectively. Also, [13] suggest using online user data to monitor public health. These studies commonly use the search data in regression models where data from official sources are also used as target variables. The model is then trained to fit the official sources and analyzed for the significance of the search data. This requires a labeled training set, e.g. actual unemployment rates. Labeled data, however, is not available for informal learning. Consequently, the method proposed in our work does not train a supervised model but rather uses statistical estimation tools to consolidate different search trends and directly compare regions[1] for a given time interval.

1.2. Organization

The rest of this paper is structured as follows: In Section 2, we mathematically describe the challenge of estimating subject literacy from search data. In Section 3, we describe the data available from Google Trends and, in Section 4, we present our methodology for measuring subject literacy from this data. In Section 5, we show that our measure has a strong correlation with other traditional and more expensive measures. Furthermore, in Section 6, we find that our metric reveals additional information about the geographical patterns in computer science curricula. In Section 7, we discuss best practices for our methods and how they could be applied to subjects other than computer science. We conclude in Section 8 with implications of our work for future research.

2. THEORY OF ESTIMATING LITERACY FROM SEARCH

In this section, we provide a mathematical formalization of our goal to estimate literacy from publicly available search data. Our formalization will provide a precise language for understanding our claims, assumptions, and the limitations of assessing subject literacy from search data.

Literacy Index: A meaningful subject literacy index for a region r should be a number, Θ_r, which captures the depth of how much the average person in a region knows about a given subject at a given point in time. The computation of Θ_r could assume that each individual i has their own literacy score, θ_i, where the literacy index for a region r is the average of these individual literacies, $\Theta_r = \mathbb{E}_{i \in r}[\theta_i]$. Ideally, the measurement for such a literacy index would require (1) a consortium of well-represented world experts design a test on a given subject and (2) have the test administered to a representative subset of people from each region. The value of Θ_r would be the average of such test scores. This test would be repeated for each region r at regular time intervals, e.g. annually. Indeed, this is the methodology used to compute Financial Literacy Index [33] maintained by the Organisation for Economic Co-operation and Development (OECD) in the Group of Twenty (G20) countries. Unfortunately, it would be a prohibitively expensive to run this ideal measurement for all topics and all countries. As such, we attempt to find a proxy measure to approximate Θ_r using readily available Internet search data. In order for a literacy index to be useful, we require that the measure correlates with proficiency in the topic, as measured by other methods such as standardized exams.

In the development of our proposed index, we assume that values θ_i are non-negative and that $\Theta_r = 0$ indicates that region

[1] Google Trends defines regions which usually correspond to countries. Some districts, however, are presented separately from the political entity on which they depend. We have also calculated CSLIs for regions instead of countries.

r has no literacy of a subject. Note that a subject literacy index, taken alone, would be a single summary statistic of a region but does not describe the distribution of θ_i completely.

Search-Based Proxy: In this work, we propose a subject literacy index computed from Internet search data. However, Internet data has several confounds, including a sampling bias. Instead of measuring learning from a random sample of individuals, we propose measuring learning from those who are using a search engine while learning. Based on a few assumptions, described in detail below, we claim that Θ_r can be measured as:

$$\Theta_r \approx \Pr(a_r) \cdot \hat{\mathbb{E}}[\theta_i | a_r] \quad (1)$$

and that this measurement is unbiased for the subject of computer science. In Equation 1, Θ_r is the literacy index of region r, $\Pr(a_r)$ is the probability that a user in region r has access to Internet search, and $\hat{\mathbb{E}}[\theta_i | a_r]$ is the average amount of search frequency and keyword diversity in a subject for users in r who have access to search. In the following subsections, we present motivation for Equation 1. Understanding the motivation illuminates both why measuring literacy from search is a promising opportunity and the ways in which we mitigate the effects of biases present in Internet search data.

Motivation for Equation 1: Using the law of total expectation, we can decompose the literacy score calculation into two terms, one for those who have access to search while learning and a term for those who do not:

$$\Theta_r = \mathbb{E}[\theta_r] = \mathbb{E}[\theta_i | a_r] \cdot \Pr(a_r) + \mathbb{E}[\theta_i | \text{not } a_r] \cdot \Pr(\text{not } a_r) \quad (2)$$

Assumption 1 (Literacy without access to search): Internet search is increasingly becoming part of learning, especially for disciplines that require a computer such as computer science and graphic design [32, 6]. As such, we assume that $\mathbb{E}[\theta_i | \text{not } a_r] \cdot \Pr(\text{not } a_r)$, the term that represents the subject literacy of people who do not have access to search, is close to zero. Mathematically, this is true if either or both of the components is near zero and the other is not unreasonably large. In particular, if access to Internet search is universal then $\Pr(\text{not } a_i)$ is zero. On the other hand, if it is unlikely that a user who does not have access to search is literate in the given subject then $\mathrm{E}[\theta_i | \text{not } s_i]$ is also close to zero. At the time of the writing of this work, the latter claim is especially believable for domains that *require* computers, such as computer science. Furthermore, while access to Internet search may not be universal, it has only increased in the last 20 years, especially in education [32, 6]. Given these assumptions, we take the second term in Equation 2 as negligible. This assumption should be re-evaluated before being applied to non-computer-based disciplines.

Assumption 2: Search Depth as a Proxy for Literacy: The second assumption we make is that search frequency and topical breadth is a reasonable proxy for subject literacy. Previous research [40] suggests that this assumption is reasonable for engineering disciplines, as information-seeking has been shown to be a substantial part of problem-solving in those fields [40]. Mathematically, we therefore claim that $\mathbb{E}[\theta_i | a_r] \approx \hat{\mathbb{E}}[\theta_i | a_r]$, where $\hat{\mathbb{E}}[\theta_i | a_r]$ is the estimate of literacy based on Internet search data. Whether this holds, depends largely on the methodology used to compose raw search data into a measure of literacy, i.e. the exact computation of $\hat{\mathbb{E}}[\theta_i | a_r]$ (see Section 5). We note that, in many other fields, information-seeking is *not* an indication of literacy, but rather an indication that a user is learning for the first time. To acknowledge this assumption, we consider our index a "literacy-proxy" and not a standard subject literacy measure.

The result of these assumptions is a simple formula:

$$\begin{aligned}
\Theta_r &= \mathbb{E}[\theta_r | a_r] \cdot \Pr(a_r) + \mathbb{E}[\theta_r | \text{not } a_r] \cdot \Pr(\text{not } a_r) && \text{(Equation 2)} \\
&\approx \mathbb{E}[\theta_r | a_r] \cdot \Pr(a_r) && \text{(Assump. 1)} \\
&\approx \hat{\mathbb{E}}[\theta_r | a_r] \cdot \Pr(a_r) && \text{(Assump. 2)}
\end{aligned}$$

This theory provides a groundwork for future researchers to argue for better proxy measurements of literacy. While these assumptions are significant, we believe that they are appropriate for measuring computer science literacy. Furthermore, we suggest that proposed measures and their assumptions can be validated by measuring their correlation with standard, more expensive tests of subject literacy. Moreover, we note that the assumptions presented herein are testable and the extent to which they are violated may be quantifiable. We leave methods to use such knowledge to mitigate known biases to future work.

3. GOOGLE TRENDS DATA

In this work, we use data from the Google search engine, which releases its search data publicly via Google Trends [15]. Google Trends presents a time series of Google search statistics for countries and sub-regions around the world. Trends automatically categorizes searches by topics and combines searches across different languages (e.g. "Artificial Intelligence" and "Kecerdasan Buatan", the Malay translation, are grouped together in the same topic). Google Trends does not expose the precise number of searches for a topic over a time frame in a given region. Instead it provides several secondary statistics, described below.

3.1. Interest by Region

For a single keyword k, Google Trends provides the ratio of queries for k to the number of total queries in each region, normalized to 100 divided by the maximum of this ratio over all regions. This data allows *comparing different regions' relative interest in the same keyword*. More concretely, if there are $s_k(r,t)$ queries for keyword k in region r in timeframe t and $S(r,t)$ total queries in region r for timeframe t, Google Trends exposes the following value for region r in time t:

$$V_k(r,t) = \frac{s_k(r,t)}{S(r,t)} \times M \quad \text{where} \quad M = \frac{100}{\max_i \frac{s_k(i,t)}{S(i,t)}}$$

For example, we can observe that in 2019:

$$V_{\text{Bayes Theorem}}(\text{Seoul}, 2019) = 100$$
$$V_{\text{Bayes Theorem}}(\text{Mexico City}, 2019) = 59$$
$$V_{\text{Bayes Theorem}}(\text{Bengaluru}, 2019) = 52$$

This means that the percentage of Google searches from Mexico City for "Bayes Theorem" is 59% of the corresponding percentage of searches in Seoul for the same topic. Google Trends also marks regions with low volume of search, ie. Madagascar. Such regions have either limited access to the Internet or an unusually low market share for Google. Google Trends data from these regions does not precisely represent the population in general and, as such, we omit the data from low search volume regions in all of our analyses.

3.2. Comparative Interest by Region

Additionally, Google Trends exposes the relative popularity between two topics. This value is the fraction of queries for a topic A over the sum of queries for topic A and B. More precisely, region r over time t is assigned a value of

$$V_{(A,B)}(r,t) = \frac{s_A(r,t)}{s_A(r,t) + s_B(r,t)}$$

scaled to 100 in total for comparison of keywords A and B. This data allows *comparing different keywords relative frequencies' in a given region.*

3.3. Comparative Interest by Time:

Finally, Google Trends also exposes keywords' relative interest over time. For a given timeframe and region, Google Trends exposes a value which reflects how the popularity of that term has changed, on a weekly basis, over the timeframe. We denote this value by $T_A(r,t_i)$, which represents the comparative interest in topic A in region r over week t_i:

$$T_A(r,t_i) = \frac{s_A(r,t_i)}{S(r,t)} \cdot k \quad \text{where} \quad k = \frac{100}{\max_j \frac{s_A(r,t_j)}{S(r,t)}}$$

We also note certain caveats for the different values described in Sections 3.1, 3.2, and 3.3. Only 80.1% of Internet search traffic is on Google and there are notable differences between countries. For example, market share for Google Search is 97% in India, 95% in Brasil, 81% in the USA. Two notable exceptions to the high market share are Russia (49%) and China (6%) [42]. As such, we expect results from these countries to be less trustworthy.

4. MEASURING CS LITERACY-PROXY FROM GOOGLE SEARCH

While we hope to be able to measure literacy of a variety of subjects, computer science is a natural first subject to investigate since it (mostly) requires a computer to learn. There are many traces of online behavior that may correlate with informal CS learning (Google Search queries, Code.org participation, GitHub commit activity, Stack Exchange browsing, etc.), but we do not have a well-defined way to use these proxy correlations to paint a picture of CS exposure for different countries.

In this section we describe our proposed metric, Computer Science Literacy-proxy Index by search (CSLI-s), as a country level score which quantifies the per-capita quality of computer science education based on Google Trends data.

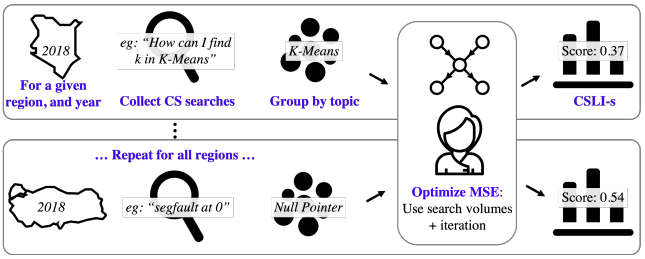

Figure 3. Overview of the CSLI-s metric, which computes CS learning scores for all geographic regions based on Internet search data.

As suggested in Section 2, we validate CLSI-s by measuring its correlation with (perhaps noisy) signals of GitHub usage, PISA scores, and online self-reported surveys (the last of which we conduct in limited number of countries due to feasibility and cost). Different measures of informal education should tell a coherent story of learning. Ideally, each measure would specify its limitations and potential biases. In future work, it would be useful to run a global survey of subject literacy to determine how to compose these different measures into a single score.

4.1. Alternative Metrics

To understand the utility of CSLI-s we compare it to several other metrics, described below.

PISA: PISA scores are composite scores of 15-year-old students' scholastic performance on mathematics, science, and reading for 70 nations. The most recent results, at the time of the writing of this paper, were published by OECD in 2016 [34].

Git: We propose a metric, "Git", that represents the count of GitHub users in each region, obtained via the GitHub Torrent [16] metadata for all 1.3 million public users. Similarly, **Git-Java** is the count of users in each region with at least one public Java repository. Java is a programming language commonly taught in introductory computer science classes.

Survey: We conducted a survey of 10,000 Internet users in 10 different countries asking "How much computer science education do you have (informal or formal)?" Respondents are selectively sampled to represent three demographic dimensions: age, gender, and geography. The survey was administered by Google Surveys during July 2019 in all countries with support for representative samples [28].

4.2. Computer Science Literacy-Proxy Index by Search (CSLI-s)

CSLI-s aims to evaluate the density of CS-related exposure on the Internet. In our approach, we utilize the minimum mean squared error (MMSE) linear estimator, the application of which we describe below. Intuitively, we use the MMSE linear estimator to develop the CLSI-s metric with two properties. Firstly, if the relative frequencies for queries on a CS-related topic (e.g. "Semaphore") are higher in a region, our metric should reflect a greater degree of computer science literacy. Secondly, the prediction error for the frequencies of various

topical search terms should be minimized, which would suggest that our metric is a useful proxy for CS literacy.

4.3. Background: The MMSE Linear Estimator

Intuitively, we model the underlying structure of our system with a single scalar latent variable for each region, representing its average subject literacy, and a single vector of observed variables, representing the keyword search frequencies. The input of the system is θ_r, which represents the extent of computer science learning for a random person in region r. The output of the system is the vector $W^r \in \mathbb{R}^M$, which represents the relative popularity for the keywords in region r.

A naive estimator would calculate $\mathbb{E}[\hat{\Theta}_r|W^r]$, the expected value of computer science literacy conditioned on the relative keyword popularity. However, the conditional distribution of computer science literacy, conditioned on the search data is unknown. Therefore, we use the *Minimum Mean Square Error Linear Estimator* to estimate Θ_r.

The MMSE linear estimator [26] assumes that the input-output relationship is affine, as shown in Equation 3[2] :

$$\hat{\Theta}_r = \Sigma_{W\Theta}^T \Sigma_W^{-1} W^r \tag{3}$$

where Σ_W is the covariance matrix for the keyword frequencies and $\Sigma_{W\Theta}^T = [\sigma_{W_1\Theta}, \ldots, \sigma_{W_M\Theta}]$ denotes the covariance vector for the relationship between keywords frequencies and computer science literacy. As its name suggests, the MMSE linear estimator minimizes the expected mean squared error between Θ_r and our estimate $\hat{\Theta}_r$ amongst all models that are linear between Θ_r and the observed W^r. The use of these values for the calculation of CSLI-s is provided in Section 4.4.

We can also estimate the average error for the MMSE. The mean square error of the MMSE Linear Estimator is:

$$MSE_{MMSE} = Var(\hat{\Theta}) - \Sigma_{W\Theta}^T \Sigma_W^{-1} \Sigma_{W\Theta} \tag{4}$$

where $Var(\hat{\Theta})$ is the variance of the calculated estimations. We use MSE_{MMSE} to iteratively determine the correlation between the keyword search frequencies, W_i, and CS literacy Θ.

4.4. Estimation Data

Our data matrix of Google Trends values has $N = 65$ rows, each corresponding to a region of high search volume[3], and $M = 26$ columns, each corresponding to the popularity of a specific keywords selected for the calculation of CSLI-s columns. We originally generated a list of 67 keywords relevant to CS via a survey administered to faculty of Stanford University Computer Science department in 2019. Faculty listed all terms that they thought were indicative of literacy in computer science. Terms without sufficient data for every region are excluded. Figure 4 shows the standard deviation in the count of queries for each keyword, sorted by keyword popularity[4]. As the popularity of the keywords decreases, the standard deviation of their frequencies (and hence the amount of information provided by the frequency of corresponding queries) across regions decreases as well. The distribution of frequencies appears Zipfian and suggests that the most popular keywords are the most informative. The data becomes highly sparse after the shaded border and the standard deviations drop sharply; thus, keywords to the right of the shaded border are excluded in the CSLI-s calculations.

Creation of the data matrix begins with the *interest by region* data (as defined in Section 3.1) for the most popular keyword in our list, $V_{Java}(r,t)$, in the first column. The following columns are filled with the *comparative interest by region* data (as defined in Section 3.2), normalized by the previous columns' entries. More precisely, our data matrix $D \in \mathbb{R}^{NxM}$ is defined by:

$$D_{ij}(t) = \begin{cases} V_{Java}(i,t) & j = 1 \\ V_{j-1}(i,t) \times \frac{V_{(j,j-1)}(i,t)}{V_{(j-1,j)}(i,t)} & j > 1 \end{cases}$$

where we have abused notation and simultaneously used j to respond to the jth most popular keyword ($j = 1$ corresponds to "Java") and its index.

4.5. Calculation of CSLI-s

The computation of CSLI-s requires generating covariance vectors and matrices for each region and substituting their values into Equation 3. The steps for generating those vectors and matrices are described below.

$\Sigma_W(t)$ is the covariance amongst the columns of the data matrix $D(t)$. It intentionally depends explicitly on year, because the explanatory power of a given keyword frequency, given the other keywords' frequencies, may change over time.

Each entry of $\Sigma_{W\Theta}^T = [\sigma_{W_1\Theta}, \ldots, \sigma_{W_M\Theta}]$ is the covariance between a keyword and computer science learning. Note that the following relationship holds: $\sigma_{W_i\Theta} = \rho_{W_i\Theta}\sigma_{W_i}\sigma_\Theta$, where σ_{W_i} and σ_Θ are the standard deviations of their respective random variables and $\rho_{W_i\Theta}\sigma_{W_i}$ is the correlation between W_i, the popularity of searches for i, and Θ. σ_{W_i} is calculated similarly to Σ_W above. In our iterative procedure, we calculate σ_Θ as the standard deviation among the estimated $\hat{\Theta}$ values and iteratively adjust $\rho_{W_i\Theta}$ values to ensure the observed standard deviations are consistent, where the MSE is minimized.

Each $\rho_{W_i\Theta}$ determines the significance of keyword i to the measurement of computer science literacy. A higher $\rho_{W_i\Theta}$ implies greater weight for the corresponding keyword in the

[2]For simplicity, we can assume $E[W^r] = E[\hat{\Theta}_r] = 0$ without loss of generality, since an additive shift of these values would not affect the rankings of different regions.

[3]The following regions that lack data on CS-related keywords, although not officially listed as low search volume regions by Google, are excluded in the CSLI-s calculations: Costa Rica, Venezuela, Lithuania, Guatemala, Ethiopia, Uzbekistan, and Cameroon. China is also excluded because of the ban on Google and limited use of Google in the country.

[4]Google Trends allows analyzing keywords as a search term or as a topic. A search term shows the data for all queries that had the exact term in the query text. A topic includes all queries related to the keyword in all languages with characters in the Latin alphabet. We restrict our analysis to topics to mitigate language issues in different regions.

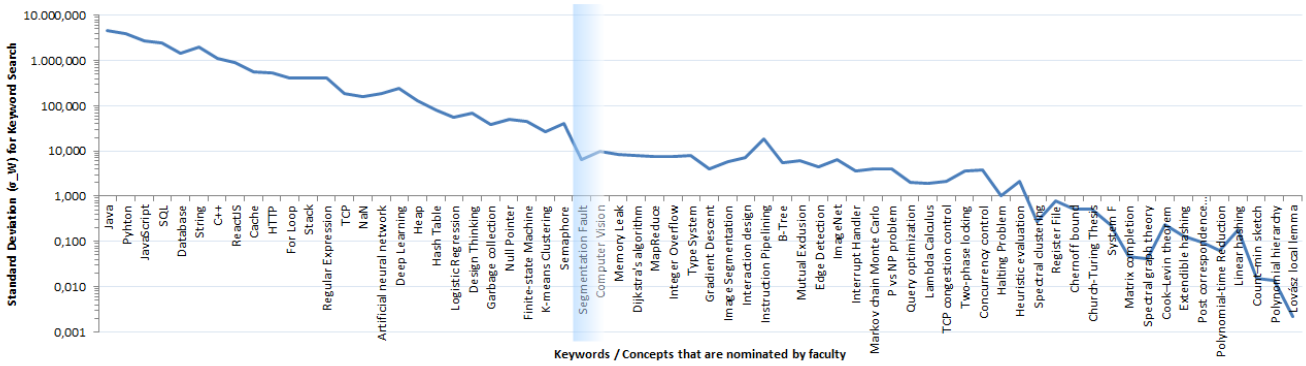

Figure 4. Standard deviations of keyword search data of 2018. The shaded border marks the beginning of excluded keywords.

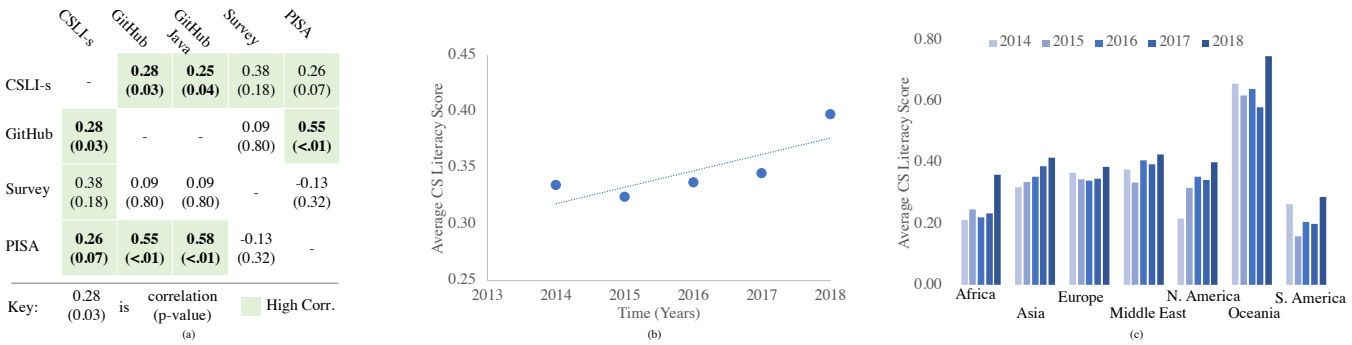

Figure 5. (a) CSLI-s over time, (b) CSLI-s by continental regions, (c) correlation between measures. Crucially, CLSI-s demonstrates correlation with other measures of CS literacy in (c).

computation of CSLI-s. Our iterative process converges to values of ρ that minimize the absolute value of the error in the model in estimating Θ, where the weights of keywords best fit the data. The convergence of our algorithm, and the uniquness of the values that minimize the error, is guaranteed by the strict convexity of Equation 4 in $\rho_{W_i\Theta}$, for each i. We note that, when comparing different time periods of data, the same values of ρ should be used for comparison so that keywords have consistant significance in the computation of CSLI-s.

$\hat{\Theta}^r$ is then calculated using the values derived above and Equation 3. $\hat{\Theta}^r$ reflects how much Google users in region r are learning about computer science. Because $\hat{\Theta}^r$ is only computed using the data from users with Internet access, we also scale $\hat{\Theta}^r$ by the percentage of individuals (as shown in Equation 5) using Internet ($U^r(t)$), published by International Telecommunication Union (ITU) [21]. As described in Section 2, we assume that members of each region that do not have Internet access have limited CS literacy.

$$CSLI\text{-}s^r(t) = (\hat{\Theta}^r(t) - min_i\hat{\Theta}^i(t)) \times U^r(t) \quad (5)$$

In the results presented below, we normalize the resulting values to the maximum score for easier comparison among years. A CLSI-s score of 1 in a given time period corresponds to the most literate region and 0 corresponds to the least literate region.

4.6. CSLI-s scores

Table 1 shows the normalized CSLI-s scores of the most and least literate regions for the last 5 years. The complete list of CSLI-s scores for all regions can be found in [2].

Region	2014	2015	2016	2017	2018
Australia	1.00	1.00	1.00	1.00	1.00
South Korea	0.88	0.94	0.88	0.86	0.88
Israel	0.88	0.81	0.96	0.81	0.78
⋮					
Pakistan	0.09	0.10	0.09	0.09	0.14
Bangladesh	0.05	0.07	0.06	0.05	0.21
Indonesia	0.08	0.01	0.01	0.04	0.07

Table 1. CSLI-s scores for selected regions.

The average CSLI-s scores from 2014 to 2018 are 0.33, 0.32, 0.34, 0.34, and 0.40, respectively. The increasing CSLI-s scores suggest that the global quality and quantity of computer science learning is increasing over time.

Figure 5 shows general trends of CSLI-s scores both globally and per continent. The positive slope of the trend line in (a) indicates a 1.5% increase per year in overall computer science learning around the world. The continental breakdown of the scores in (b) indicates that the increase in CSLI-s scores is largely due to Asian regions. These statistics may inform the computer science education community in ways that enable

policymakers to strengthen global upward trends, and where in the world best to do so.

5. EVALUATION OF RESULTS

It is difficult to understand the accuracy of CSLI-s without a "ground truth" measure of regions' per-capita computer science literacy. However, we compare CSLI-s to several alternative metrics, as suggested in Section 2. Our results, in general, indicate that CSLI-s generates scores consistent with those metrics. In particular, CSLI-s and the Human Development Index (HDI) [43] are strongly positively correlated ($r = 0.54$) on average. This is not surprising, as we might expect more developed countries to have higher levels of computer science education.

We also note that CSLI-s has a large (>0.25) positive Spearman correlation with all other measures tested: GitHub, Git-Java, and PISA, where the aggregate claim has p-value < 0.001. See Figure 5-(a) for the correlations and p-values for each individual measure. CSLI-s scores also demonstrate a high correlation with our user survey, though we note that due to a small sample size tested in only 10 countries, these results may have been due to chance (p = 0.18).

PISA was highly correlated with Github and Git-Java, but none of these correlated with the user survey. Finally, we note that both PISA, GitHub and CSLI-s are negatively correlated with population (correlations of -0.26, -0.29 and -0.33 respectively), suggesting relatively better computer science literacy for less populated nations.

6. PATTERNS IN CURRICULA

In addition to the regional, aggregate measures of literacy introduced in Section 4 and the temporal patterns we observed in Section 1, it is also possible to understand geographical trends in curricula using unsupervised machine learning techniques.

Figure 6 shows a clustering of countries by the popularity of computer science search terms between 2014 and 2018. Figure 6 is generated using t-SNE [44], whereby each region is represented as a point and distances between regions is inversely proportional to the similarity between their search terms' popularities. The figure suggests that geographically close regions are inclined to have similar types of computer science curricula; for example, the US is close to Canada and Austria is close to Germany. Perhaps the most striking cluster of countries in the embedding is the cluster containing Argentina, Brazil, Chile, Thailand, Indonesia, and Turkey (top right). Upon deeper inspection, we notice these countries tend to have a stronger emphasis on systems (such as Internet algorithms) and less emphasis on artificial intelligence than other countries. Another cluster contains Australia and New Zealand; this cluster is represents high search frequencies for terms related to data structures and algorithms, such as "Heap" and "Semaphore". This group also demonstrates low search frequencies for more theoretical topics like "Finite State Machines." The existence of clusters presents an interesting opportunity for future work, as Google Trends data may reveal novel insights into these subtle differences in countries' curricula.

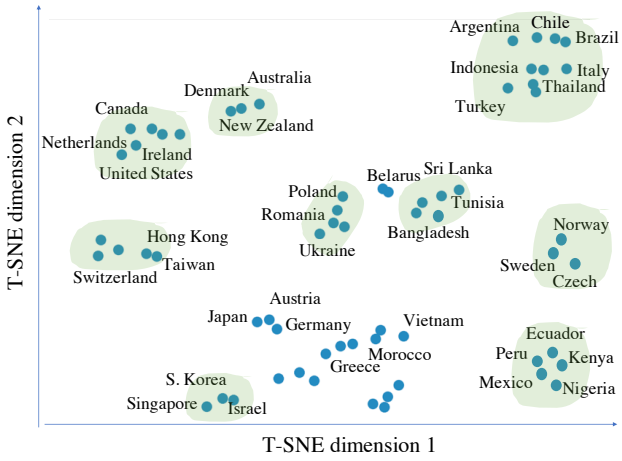

Figure 6. t-SNE embeddings of the search popularity between different clusters. Countries which are close to one another have similar computer science search patterns. The cluster with Thailand and Indonesia is the most distinct, and focuses more on systems for the Internet.

7. DISCUSSION

7.1. Best Practices

We note that while the process for calculating an index applies generally to many different fields, CSLI-s is made specific to computer science through the use of keywords related to computer science learning. To better understand how different keywords might affect our results, we performed the same analysis as in Section 4, but with the 40 most popular keywords from Figure 4. However, here is no significant difference in the results, which suggests that spanning specific subfields of computer science with the minimum amount of popular keywords is enough to calculate a meaningful CSLI-s score. We invite the community to develop a standard set of keywords and topics so that a consistent way to compute the CSLI-s for computer science subfields can be developed.

More generally, our computation of an index in computer science such as CSLI-s applies to other fields, e.g. history, where global literacy may be estimated with a selection of relevant keywords. We anticipate that an analogous index for other fields will not be heavily dependent on the exact set of keywords used, though we leave a quantitative analysis of this assertion to future work. We remind readers that our source code is available, and that the assumptions described in Section 2 should be studied carefully before using this index for decision-making extending our analysis to other fields. We present an example adaptation to the subject of climate change below, in Section 7.2. We encourage others to propose alternative measures of subject literacy from search data and to surface, examine, and mitigate sources of bias. We consider this a new research direction and encourage substantial caution before using the measure for large policy decisions.

We further note that CSLI-s provides a measure of computer science literacy *on average* for the entire population in a region. Regions with strongly developed computer science literacy may still get relatively low scores if computer science knowledge is unequally distributed amongst its populace. In light of

this observation, one strategy for increasing the CSLI-s score would be to provide equitable computer science education for the entire population. This is consistent with the United Nations' Sustainable Development Goal of education *for all*. Nonetheless, measuring formal and informal computer science literacy globally is an important problem. CSLI-s is a first step in what we hope is a rich and useful research direction.

7.2. Application to Climate Change

In the preceding sections, we built a case study around the use of Internet search data to measure a proxy of computer science literacy. In this subsection, we use the same methodology on a different topic: climate change. As for computer science, we noticed that there were substantial seasonal effects on search terms' frequencies; see Figure 7.

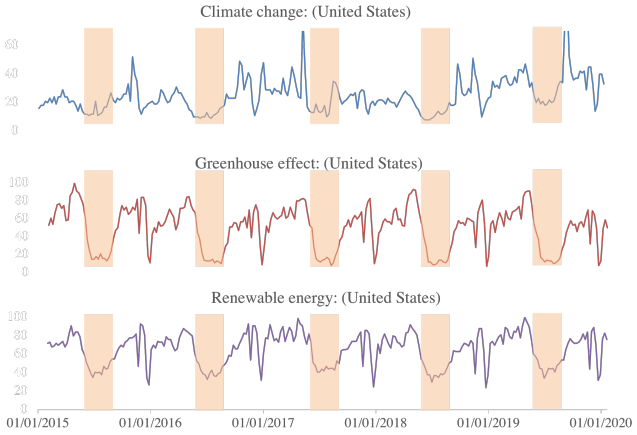

Figure 7. Search terms exhibit strong, consistent seasonal patterns. However, they exhibit them to different degrees. The changes in "Climate Change" are less drastic between Summer and School than "Greenhouse Effect".

Interestingly, searches for climate change-related topics in the United States are lowest in the summer, when temperatures are highest. Rather, the frequency of climate change-related searches appears to be strongly linked to the school calendar. These seasonal trends suggest that a similar analysis might be applicable to climate change, though it is not clear if the assumptions described in Section 2 are applicable. In particular, Assumption 2 likely does not hold, which would imply that search behavior does not reflect subject "literacy." As such, we do not propose an analogous Climate Change Index. However, our initial experiments reveal that climate change concepts have more substantial search volume for countries in the Southern Hemisphere. See Figure 1, for a visualization of this geographic pattern. While less people search for climate change topics in the Northern Hemisphere, those searches tend to be during the school year suggest that climate change is primarily learned through formal education.

7.3. Curricula Timing

Additionally, we observe that Internet search trends reveal the curricula timing of concepts in different countries. For search terms which are more commonly searched during the school year, temporal search patterns indicate the time within the school year that concepts are being taught. For example, in the U.S., search volume for "Global Warming" consistently peaks around the 105th day of the year (April 16th), "Fossil Fuel" consistently peaks around the 111th day of the year (April 22nd), and "Ocean Accidification" peaks on average on the 118th day of the year (April 29th). These dates seem to agree with the order of topics presented in a typical class on climate change. These results suggest that the seasonal trends we observe likely reflect which concepts are being learned in school.

Table 2. Days into the year when different concepts have their peak popularity in the United States.

Concept	Peak Search Day
Global Warming	105
Fossil Fuel	111
Sea Level Rise	111
Greenhouse Effect	114
Ocean Acidification	118

8. CONCLUSION

Assessing the quality of learning worldwide in a regionally specific, continuous, and inexpensive way remains an important problem towards achieving the United Nations' Sustainable Development Goal 4 to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all." In this paper, we demonstrated that Google Trends data can be used as a meaningful proxy for measuring worldwide progress on this goal. To the best of our knowledge, this is the first paper to analyze Google Trends data to understand learning. This is likely because, until now, it has been conceptually overlooked to derive meaningful information from the data that Google publicly releases.

Using Google Trends data, we presented several methodologies to (1) calculate country-level per capita statistics of educational quality, (2) measure the extent to which topics are learned in school, and (3) find curriculum-level patterns across geographically regions.

More specifically, we introduced a new metric to assess educational quality, the Computer Science Literacy-proxy Index by Search (CSLI-s), using these methodologies and show that this statistic is correlated with several other validation measures.

Using this metric, we quantified that computer science literacy is growing around 1.5% per year around the world and observed that Oceania was well ahead of the world in per capita computer science literacy. We observed that climate change topics tend to be much more searched in the Southern Hemisphere. While fewer people search for climate change topics in the Northern Hemisphere, those searches tend to be during the school year.

As Internet accessibility rates continue to increase globally, we anticipate that digital traces of user behavior will continue to present a wealth of information from which we can learn about human behavior. We hope that the methods presented in this paper will allow researchers, educators, and policymakers to tackle a difficult problem: an assessment of informal learning quality around the world.

REFERENCES

[1] ANDERSON, T. *The Theory and Practice of Online Learning*, 2nd ed. AU Press, Canada, 2009.

[2] ARSLAN, S., TIWARI, M., AND PIECH, C. Global cs fluency ranking. `https://github.com/serhatarslan-hub/global_cs_fluency_ranking`, 2020.

[3] ASKITAS, N., AND ZIMMERMANN, K. F. Google Econometrics and Unemployment Forecasting. IZA Discussion Papers 4201, Institute of Labor Economics (IZA), June 2009.

[4] ASTRACHAN, O., MORELLI, R., CHAPMAN, G., AND GRAY, J. Scaling high school computer science: Exploring computer science and computer science principles. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (New York, NY, USA, 2015), SIGCSE '15, ACM, pp. 593–594.

[5] CHOI, H., AND VARIAN, H. Predicting the present with google trends. *Economic Record 88*, s1 (2012), 2–9.

[6] CODE.ORG, 2019. https://code.org.

[7] CONDE, M. A., GARCIA-PENALVO, F. J., AND ALIER, M. Interoperability scenarios to measure informal learning carried out in ples. In *2011 Third International Conference on Intelligent Networking and Collaborative Systems* (Fukuoka, Japan, Nov 2011), IEEE, pp. 801–806.

[8] DOWNES, S., ET AL. New technology supporting informal learning. *Journal of emerging technologies in web intelligence 2*, 1 (2010), 27–33.

[9] EPSTEIN, J. L., AND MCPARTLAND, J. M. The concept and measurement of the quality of school life. *American Educational Research Journal 13*, 1 (1976), 15–30.

[10] ETTREDGE, M., GERDES, J., AND KARUGA, G. Using web-based search data to predict macroeconomic statistics. *Commun. ACM 48*, 11 (Nov. 2005), 87–92.

[11] FRENKEL, K. A. Cs enrollments rise at the expense of the humanities? *Commun. ACM 56*, 12 (Dec. 2013), 19–21.

[12] GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S., AND BRILLIANT, L. Detecting influenza epidemics using search engine query data. *Nature 457* (Nov 2008), 1012 EP –.

[13] GITTELMAN, S., LANGE, V., A. GOTWAY CRAWFORD, C., OKORO, C., LIEB, E., DHINGRA, S., AND TRIMARCHI, E. A new source of data for public health surveillance: Facebook likes. *Journal of medical Internet research 17* (04 2015), e98.

[14] GOEL, S., HOFMAN, J. M., LAHAIE, S., PENNOCK, D. M., AND WATTS, D. J. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences 107*, 41 (2010), 17486–17490.

[15] GOOGLE. Trends, 2019. https://www.google.com/trends.

[16] GOUSIOS, G. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (Piscataway, NJ, USA, 2013), MSR '13, IEEE Press, pp. 233–236.

[17] GOVE, A., AND BLACK, M. M. Measurement of early childhood development and learning under the sustainable development goals. *Journal of Human Development and Capabilities 17*, 4 (2016), 599–605.

[18] GUZMAN, G. Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement 36* (11 2011), 119–167.

[19] HALLIDAY-WYNES, S., AND BEDDIE, F. Informal learning. at a glance. Tech. rep., National Centre for Vocational Education Research (NCVER), Adelaide, Australia, 05 2009.

[20] HERTZ, M. What do "cs1" and "cs2" mean?: Investigating differences in the early courses. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education* (New York, NY, USA, 2010), SIGCSE '10, ACM, pp. 199–203.

[21] INTERNATIONAL TELECOMMUNICATION UNION. Individuals using internet, 2019.

[22] JORDAN, K. Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning 15*, 1 (Jan. 2014).

[23] KALELIOĞLU, F. A new way of teaching programming skills to k-12 students: Code. org. *Computers in Human Behavior 52* (2015), 200–210.

[24] KIZILCEC, R. F., PIECH, C., AND SCHNEIDER, E. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (New York, NY, USA, 2013), LAK '13, ACM, pp. 170–179.

[25] LAZER, D., KENNEDY, R., KING, G., AND VESPIGNANI, A. The parable of google flu: traps in big data analysis. *Science 343*, 6176 (2014), 1203–1205.

[26] LEON-GARCIA, A. *Probability, Statistics, and Random Processes for Electrical Engineering*. Pearson/Prentice Hall, Upper Saddle River, NJ, USA, 2008.

[27] MARGULIEUX, L., KETENCI, T. A., AND DECKER, A. Review of measurements used in computing education research and suggestions for increasing standardization. *Computer Science Education 29*, 1 (2019), 49–78.

[28] MCDONALD, P., MOHEBBI, M., AND SLATKIN, B. Comparing google consumer surveys to existing probability and non-probability based internet surveys, 2012.

[29] MCLAREN, N., AND SHANBHOGUE, R. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin 51*, 2 (2011), 134–140.

[30] MORRISON, B. B., AND DISALVO, B. Khan academy gamifies computer science. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* (New York, NY, USA, 2014), SIGCSE '14, ACM, pp. 39–44.

[31] MULLER, C. Measuring education and skill. *The ANNALS of the American Academy of Political and Social Science 657*, 1 (2015), 136–148. PMID: 25983334.

[32] NATIONAL SCIENCE BOARD. Instructional technology and digital learning, 2018.

[33] OECD. Infe report on adult financial literacy in g20 countries, 2017.

[34] OECD. Programme for international student assessment, 2018. http://www.oecd.org/pisa/.

[35] PAPPANO, L. The year of the mooc. *The New York Times 2*, 12 (2012), 2012.

[36] PIECH, C., SAHAMI, M., HUANG, J., AND GUIBAS, L. Autonomously generating hints by inferring problem solving policies. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (New York, NY, USA, 2015), L@S '15, ACM, pp. 195–204.

[37] POLGREEN, P. M., CHEN, Y., PENNOCK, D. M., NELSON, F. D., AND WEINSTEIN, R. A. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases 47*, 11 (12 2008), 1443–1448.

[38] PRITCHARD, D., AND VASIGA, T. Cs circles: An in-browser python course for beginners. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education* (New York, NY, USA, 2013), SIGCSE '13, ACM, pp. 591–596.

[39] REILLY, S., RICHEY, S., AND TAYLOR, J. B. Using google search data for state politics research: An empirical validity test using roll-off data. *State Politics & Policy Quarterly 12*, 2 (2012), 146–159.

[40] ROBINSON, M. A. An empirical analysis of engineers' information behaviors. *Journal of the American Society for information Science and technology 61*, 4 (2010), 640–658.

[41] SMITH, T. Some aspects of measuring education. *Social Science Research 24*, 3 (1995), 215 – 242.

[42] STATCOUNTER GLOBALSTATS. Search engine market share worldwide, 2019.

[43] UNITED NATIONS DEVELOPMENT PROGRAMME. Human development index (hdi), 2019. http://hdr.undp.org/en/content/human-development-index-hdi.

[44] VAN DER MAATEN, L., AND HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research 9* (2008), 2579–2605.

[45] WATKINS, K. E., AND MARSICK, V. J. Towards a theory of informal and incidental learning in organizations. *International Journal of Lifelong Education 11*, 4 (1992), 287–300.