Learning by Doing versus Learning by Viewing: An Empirical Study of Data Analyst Productivity on a Collaborative Platform at eBay

YUE YIN, Northwestern University, USA
ITAI GURVICH, Cornell Tech, USA
STEPHANIE MCREYNOLDS, Alation Inc., USA
DEBORA SEYS, eBay Inc., USA
JAN A. VAN MIEGHEM, Northwestern University, USA

We investigate how data-analyst productivity benefits from collaborative platforms that facilitate *learning-by-doing* (i.e. analysts learning by writing queries on their own) and *learning-by-viewing* (i.e. analysts learning by viewing queries written by peers). Learning is measured using a behavioral (productivity-improvement) approach. Productivity is measured using the time from creating an empty query to first executing it.

Using a sample of 2,001 data analysts at eBay Inc. who have written 79,797 queries from 2014 to 2018, we find that: 1) *learning-by-doing* is associated with significant productivity improvement when the analyst's prior experience focuses on the focally queried database; 2) only *learning-by-viewing* queries that are authored by analysts with high output rate (average number of queries written per month) is associated with significant improvement in the *viewer's* productivity; 3) *learning-by-viewing* also depends on the "social influence" of the author of the viewed query, which we measure 'locally' based on the number of the author's direct viewers per month or 'globally' based on the how the author's queries propagate to peers in the overall collaboration network. Combining results 2 and 3, when segmenting analysts based on output rate and 'local' social influence, the viewing of queries authored by analysts with high output but low local influence is associated with the largest improvement in the viewer's productivity; whereas when segmenting based on output rate and 'global' social influence, the viewing of queries authored analysts with high output and high global influence is associated with the largest improvement in the viewer's productivity.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**;

Additional Key Words and Phrases: Learning-by-doing; Learning-by-viewing; Productivity; Data analysts; SQL Query; Expert roles; Segmentation; Collaborative data platform; Alation; eBay

ACM Reference Format:
Yue Yin, Itai Gurvich, Stephanie McReynolds, Debora Seys, and Jan A. Van Mieghem. 2018. Learning by Doing versus Learning by Viewing: An Empirical Study of Data Analyst Productivity on a Collaborative Platform at eBay. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2, CSCW, Article 193 (November 2018). ACM, New York, NY. 27 pages. https://doi.org/10.1145/3274462

Authors' addresses: Yue Yin, Northwestern University, 2211 Campus Dr, Evanston, Illinois, 60201, USA, yue-yin@kellogg.northwestern.edu; Itai Gurvich, Cornell Tech, New York City, New York, USA, gurvich@cornell.edu; Stephanie McReynolds, Alation Inc. Redwood City, California, USA; Debora Seys, eBay Inc. 2025 Hamilton Avenue, San Jose, California, USA; Jan A. Van Mieghem, Northwestern University, 2211 Campus Dr, Evanston, Illinois, USA, vanmieghem@kellogg.northwestern.edu.

2573-0142/2018/11-ART193 $15.00
https://doi.org/10.1145/3274462

1 INTRODUCTION

Effective data analytics drives business success by enhancing managerial decision-making. Companies often struggle to maintain growth in the productivity of their data analysts. In this paper, we investigate how data-analyst productivity benefits from collaborative platforms that, in addition to providing a query-writing environment for the analysts, also facilitate access to queries authored by their peers. Productivity is measured using the time from creating an empty query to first executing it.

Companies in various industries employ data analysts. From financial services to retailing to social media, data analysts aim to transform data into valuable information for decision-makers. By retrieving, organizing and narrating raw data, data analysts can spot trends in the market, enable managers to know more about their consumers, and develop recommendations that are aligned with the company's strategy. The current demand for skilled data analysts out-paces the supply. A 2016 McKinsey Global Institute report concludes that by 2024 the U.S. economy would be in shortage of 250,000 analytics professionals [40]. Furthermore, surveyed business leaders find it challenging to recruit and retain proficient data analysts [39].

A good data analyst must be technologically up-to-date to bring insightful results swiftly [18]. The writing and execution of queries is central to the work of data analysts with large-scale data. Queries are written using SQL (Structured Query Language) [71] to answer business questions like: What goods are flying off a retailer's shelf? Who prefers to shop in a boutique store versus on-line? What are the ten most frequently searched items on eBay Motors in London in July 2017? Data analysts answer these questions by extracting data from the proper databases, performing data manipulations (e.g., sorting, grouping, filtering and joining), and finally reporting the results.

Proficiency in programming SQL queries is a learned skill. Organizational learning theory defines learning as a change in the knowledge that occurs as a function of experience [31]. Such knowledge transformation occurs at different levels in organizations — individual, group, organizational, and inter-organizational [21]. Studies have demonstrated that much of the programming knowledge is *tacit knowledge,* i.e., knowledge that usually is not openly expressed or taught [83, 94]. It also has been established that expert programmers know more than mere syntax and semantics of a particular language; compared to novices, their knowledge is better organized. For example, an expert programmer would be able to see the underlying commonalities and the differences among various problems and programs. Numerous researchers who aim to characterize expert programming also have suggested the existence of reusable "chunks" of knowledge representing solution patterns that achieve different kinds of goals [56].

Most current approaches measure learning by assessing changes in cognition. Qualitative methods like questionnaires, interviews and verbal-protocol analyses are typically used to that end [30, 44]. Such cognitive approaches are nonetheless unable to capture tacit knowledge [41]. Behavioral approaches measure learning by assessing changes in practices or performance and have been shown to capture the tacit knowledge well [6, 7, 27]. Recently more researchers start exploring a behavioral approach to quantitatively measure such 'informal learning' from a large online community. For example, Yang et al. measured learning of a Scratch user as growth in the cumulative repertoire of weighted vocabulary block use [25, 98]. We thereby deploy a behavioral approach—measuring change in performance—in our study. The expedition of SQL programming indicates the positive learning outcome of data analysts.

In organizational learning theory, experience—typically defined as the total or cumulative number of task completions—underpins learning. The most fundamental characterization of experience is whether it is acquired directly by the focal organizational unit or indirectly from other units [6]. Two modes of organizational learning are derived from this characterization: learning from

direct experience (i.e. from one's own practices) and learning from indirect experience (i.e. from other organizational units' experience) [55]. Learning from direct experience is the embodiment of "practice makes perfect" while learning from indirect experience emphasizes the circulation of knowledge among peers. Studies of the well-known *learning curve* provide considerable evidence of learning from direct experience [27, 99]. There is also extensive work on collaboration and knowledge sharing that investigates learning from indirect experience [8, 36, 42, 59, 65].

We study these two modes of organizational learning when eBay data analysts work on *Alation*. *Alation* is an enterprise collaborative data platform that makes data accessible to individuals across the organization. The platform empowers analysts to write SQL queries using well-curated data, and allows them to publish their own queries or view any public query authored by their peers. We focus on two potential learning processes on *Alation*: *learning-by-doing ("by oneself")* vs. *learning-by-viewing (peers' queries)* . *Learning-by-doing* captures how data analysts become faster the more queries they write; how they learn from direct experience. In contrast, *learning-by-viewing* captures how data analysts improve by viewing queries authored by their peers; it is an instance of learning from indirect experience. We pose the following research questions:

1. ***Learning-by-doing*: How is data analyst productivity associated with self-practice?**
2. ***Learning-by-viewing*: How is data analyst productivity associated with viewing of queries authored by peers?**

Organizational learning theory emphasizes the role of expert in facilitating the diffusion and validation of credible knowledge. Previous studies have demonstrated that group members are likely to accept and put more weight on information from a recognized expert [86]. The retaining of and the interaction with exceptional performers appears to affect organizational outcomes [6, 16]. Given that most of the programming knowledge is tacit knowledge, it is not clear how to characterize expert (or "star") analysts. The conventional approaches to the characterization of stars are based exclusively on individual output [9, 38, 100]. Classic economic growth theories nonetheless claim that human capital externalities (e.g. the influence that an individual has on the performance of others) are also a key input in the generation of knowledge [1, 61, 78]. Recent empirical work adopts this view and expands the traditional characterization of 'star' by adding measurements of social influence [69]. Recent studies on software development gauge expertise identification approaches by considering the code a developer authors and the code that the developer consults during their work [35, 81].

Inspired by the literature on the role of experts in organizational learning, we ask:

3. ***Learning from experts*: How is learning-by-viewing a query associated with the expertise of the query's author?**

Our study is also related to two well-known perspectives in cognitive psychology and learning sciences [5, 19, 37]: the "within-the-human" perspective that is generally attributed to Jean Piaget [96] vs. the situated cognition perspective proposed by Jean Lave and Etienne Wenger [58]. The former focuses on the internalized development within the individual's mental representation of the world, and presumes that knowledge can be constructed through one's own practices [19]. The latter, in contrast, suggests that knowledge is constructed in a social context, and individual participation in valued social practices is critical for successful learning [19, 37]. Read in our context, the within-the-human perspective would suggest that, mostly, the analyst acquires programming knowledge from writing queries by herself (*learning-by-doing*). The situated-cognition perspective suggests, instead, that learning is mostly accomplished through the analyst's interaction with other analysts (*learning-by-viewing*). For situated-cognition perspective it is needed, of course, for newcomers to be able to observe experts. In our context *Alation* is the platform that facilitates this so-called "legitimate peripheral participation" [58].

The rest of the paper is organized as follows. In the next section, we discuss the theory and develop our hypotheses. We then describe the empirical setting, data and measures. Later, we present our analysis strategy and report our results. We conclude with a comprehensive discussion of the results and their potential implications.

2 THEORY AND HYPOTHESES

Organization learning theory has been applied in a broad spectrum of industries from manufacturing to services [10, 24]. Other recent studies focus on knowledge industries like IT consulting and software development [32, 48, 54]. We follow this path, relying on organizational learning theory to develop and test hypotheses related to the learning of data-analysts working, as individuals, on a collaborative platform that facilitates knowledge sharing. Our hypotheses pertain to two modes of organizational learning: learning from direct experience and learning from indirect experience [32, 50, 75]. We use productivity as the main variable of interest and measure how it relates to the accumulated experience that data analysts have gained by writing queries on their own (*learning-by-doing*) and by viewing queries written by their peers (*learning-by-viewing*). In formulating the hypotheses we also rely on two cognitive theories of learning: *learning-by-doing* as it relates to the within-the-human perspective and *learning-by-viewing* as it relates to situated-learning perspective.

2.1 Learning by Doing

2.1.1 Individual Learning from Direct Experience. Recent studies of individual learning from direct experience cover various industries. Kim et al. estimate the learning curve of IT consultants [54]. KC et al. examine the direct impact of a cardiologist's own prior experience on individual learning [50]. Staats and Gino compare the benefits of individual worker's experience in a day or over several days [85]. They find that specialization is related to productivity improvement over a single day. We expect data analysts in our study to benefit from their past experience of writing queries. A positive answer to the following hypothesis further validates the earlier findings and extends them to the context of data analysts.

Hypothesis 1. *Past experience of writing queries is associated with an improvement in the data analyst's productivity.*

2.1.2 Specificity of Direct Experience. Repetition of a given task is likely to improve the performance of an individual more than experience with related (but different) tasks. Boh et al. show that specialized experience with the same system has the greatest impact on productivity for modification requests completed by individual developers [32, 48]. Such findings, identifying the benefit of focal experience, appear in other service industries [28, 51, 85].

On *Alation*, data analysts write queries using different databases. In an interview study by Kandel et al. focusing on the challenges of data analysts, most of the respondents mentioned the difficulty in interpreting certain database fields [46]. The evidence in the literature suggests that, as the analyst practices more with the focal database, she will become more familiar with this database's nuances, including the field definitions, data quality and assumptions. We therefore measure the association between productivity and specificity of experience. A positive answer to the hypothesis below corroborates the value of focal direct experience in writing queries.

Hypothesis 2. *Past experience in querying the focal database is associated with greater improvement in data analyst productivity than past experience querying different databases.*

2.2 Learning by Viewing

2.2.1 Individual Learning from Indirect Experience. "social interaction among individuals, groups and organizations are fundamental to organizational knowledge creation" [68]. Organizational

learning is frequently an interactive, social phenomenon [91]. Such communal processes are important because no one person embodies sufficient knowledge for solving all complex organizational problems. For instance, in the context of machine repair technicians, most of the knowledge is not acquired in the classroom, but comes, rather, from informal story-sharing among technicians and users about their experiences in particular work environments [15, 70]. This finding, that individuals also benefit from their peers' experience (learning from indirect experience) is also confirmed in [36, 42, 43, 45, 59, 65, 67, 76, 84].

In the context of computer programming, Brandt et al. [11] propound that by relying on information and source code fragments provided by other people from the Web, developers engage in just-in-time learning of new skills and approaches, clarify and extend their existing knowledge, and remind themselves of details deemed not worth remembering. Vasilescu et al. [92] argue that participation in on-line programming communities (e.g. StackOverflow) speeds up code development since quick solutions to technical challenges can be provided by peers. Dasgupta et al. confirmed that remixing—defined as the reworking and combination of existing creative artifacts—acts a pathway to learning [25]. They found that a learner's repertoire of programming concepts increases when she engages in remixing.

Yet there is a trade-off. Viewing peers' code may delay programming activities as both *viewing* and *programming* compete for the developer's time and attention. Current empirical evidence for the benefit of learning from indirect experience is inconclusive. Waldinger finds no evidence for peer effects on the productivity of researchers in physics, chemistry and mathematics. In his study, even very high-quality scientists do not affect the productivity of their local peers [95]. KC et al. investigate the relationship between cardiologists' current performance and the performance of their colleagues in the same hospital [50]. But their data does not include detailed "views" information, i.e., what information individual cardiologists actually observe or share among each other. The authors, therefore, call for future research to identify the precise micro-mechanisms at work, exploring how knowledge is shared among individuals and affects their performance. Our fine-grained data include a complete history of each analyst's record of viewing specific peers' queries, offering an opportunity to respond to the authors' call. We test the following hypothesis to measure learning from indirect experience in analysts writing queries. A positive answer to this hypothesis confirms that it is highly possible that analysts who mostly view queries written by peers bear high productivity. A negative answer still leaves the possibility that only viewing queries written by *certain* peers predicts high productivity. This is investigated in section 2.2.2.

Hypothesis 3. *Past experience of viewing queries written by peers is positively associated with data analyst productivity.*

2.2.2 Characterizing Star Data Analysts. Previous studies on knowledge spillover and peer effects among scientists suggest a differential impact of collaborating with different types of individuals. Exceptional performers, or stars, may greatly advance the production of ideas and the innovation process [9, 69]. The situated-cognition theory in learning sciences places and emphasizes on the role of experts [26]. In the context of programming and software development, developers appear to have greater interest in following some prolific developers, who are considered 'coding rockstars' by the overall community [23]. In the on-line programming community, a developer's status can affect decision-making. Tsay et al. find that contributions from higher-status submitters are more readily accepted by project managers [89].

In the context of data analysts, we must first ask how to identify (or characterize) the "rockstars" and, given such a characterization, how are these said stars associated with the changes in productivity of their peers? The characterization we propose in this paper differs from most existing taxonomy by considering not just the individual output [9, 38, 100], but also the individual's social

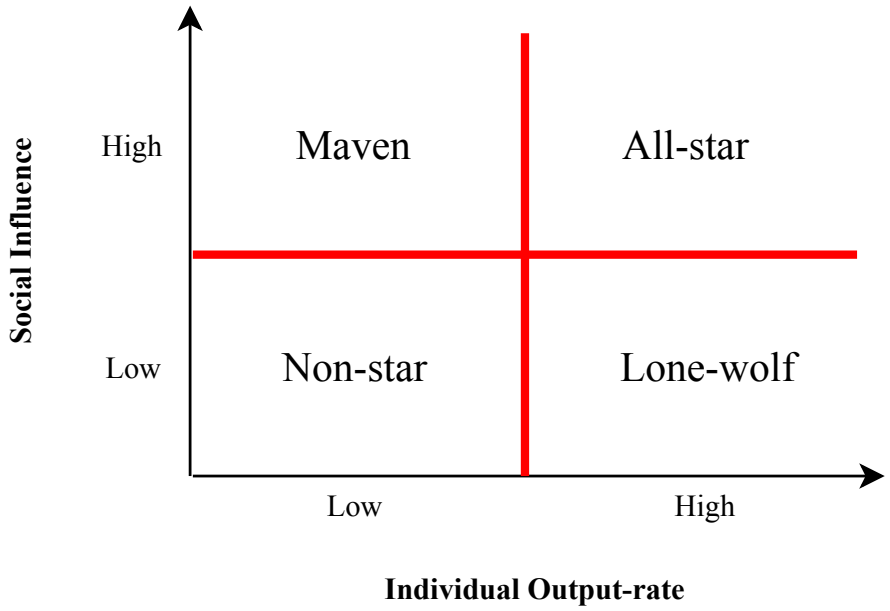

Fig. 1. We segment analysts using two dimensions—output rate and social influence—into four types: All-star, Lone-wolf, Maven and Non-star.

influence on her peers. Such an expanded definition is unavoidable here as we wish to measure *learning-by-viewing* which is an interaction-based construct. This need is also identified in [69, 89] who suggest that the characterization of "stars" should consider the individuals' social influence [69, 89]. We introduce a two-dimensional segmentation of analysts that incorporates both a measure of the individual analyst' output and a measure of her social influence on *Alation*. Relying on Oettl's characterization of star scientists [69] we segment analysts in our study into four types: All-star, Lone-wolf, Maven, Non-star; see Figure 1.

We specify two segmentations in this paper that both use output rate yet each use a different measure of social influence. We adopt the conventional measure of individual output: *output-rate* = the average number of queries created by the analyst i per unit of time. [2, 38] Taking months as unit of time, we will use:

$$\text{Monthly Output of Queries}_i = \frac{\text{Total number of queries } i \text{ has written}}{\text{Number of months since } i \text{ joined in } \textit{Alation}} \tag{1}$$

We adopt two different measures of social influence, *viewership* and *PageRank*, to describe how influential an analyst's queries have been since they were written on *Alation*. *Viewership* captures the average number of distinct viewers per month of all queries authored by analyst i. We define:

$$\text{Monthly Viewers per Query}_i = \frac{\text{Total number of distinct data analysts who have viewed } i\text{'s queries}}{\sum_{\text{query } k \text{ written by } i} \text{Months that query } k \text{ is viewable}} \tag{2}$$

which represents the attention that focal analyst i receives from her peers on *Alation*. Viewership is a measure of the "local" influence of an author on its direct viewers. A qualitative interview study by Dabbish et al. [23] demonstrates that, in large-scale distributed collaborations and communities

of practice (e.g. GitHub), the attention that a developer has received signals her status in the community. Quoting to a representative participant in their study, "[One visible cue is] the number of people watching a project or people interested in the project; obviously it's a better project than versus something that has no one else interested in it."

The second measure of social influence is *PageRank*, which was introduced by Google for weighting the importance of a web page based on the number and quality of links to this page [13]. To compute the *PageRank* of each data analyst in our study, we first build a directed, analyst-to-analyst network that represents the social interactions on *Alation*, which we explain later in 3.2.2. Then, running the PageRank algorithm on this network returns the *PageRank* for every data analyst. The analyst with higher *PageRank* is considered more influential on the overall network herself. While *viewership* captures the 'local' influence, *PageRank* represents a 'global' network-wide influence of an analyst by capturing not only her direct viewers but also the viewers of her viewers etc.

We hypothesize that *learning-by-viewing* queries authored by analysts who outperform in both output-rate and social influence is associated with the largest improvement in productivity. To confirm the expert roles in *learning-by-viewing*, we start with testing the following hypothesis.

Hypothesis 4. *Past experience of viewing queries written by different types of data analysts (All-star, Maven, Lone-wolf or Non-star) is associated with different change in data analyst productivity.*

A rejection of Hypothesis 4 would imply that the predicted productivity improvement through viewing queries are independent of the type of the author. In contrast, support for Hypothesis 4 confirms the superiority of expert roles in our context and can be followed by further investigation: which type of analysts writes the most informative queries that are associated with largest productivity improvement of its viewers?

3 METHODS

3.1 Study Platform

We study eBay data analysts writing and viewing queries on *Alation*. *Alation* is an enterprise collaborative data platform developed by Alation Inc. and used by eBay Inc. As one of its clients. *Alation* serves as an all-inclusive 'resort' for data analysts. First, it provides a repository for all technical meta-data in the analytics data warehouse. Main data services that can connect to *Alation* include Oracle, Teradata, MySQL, SQL Server and Tableau. A data analyst can conveniently access data if she has the proper permissions. Second, *Alation* integrates various analytics tools for data analysts to compose and execute queries, as well as produce comprehensible results. Third, *Alation* advances collaborations and social computing among data analysts inside eBay. A data analyst can share her knowledge with the community by publishing her queries, writing articles about her good practices or participating in conversations on technical issues. A data analyst can also seek knowledge from the community by viewing queries authored and published by other peers, searching for relevant articles or asking for help in the conversation board. Serving as an on-line enterprise community, *Alation* supports collaboration, knowledge sharing, reuse of resources, expertise location, innovation, organizational change and social networking [63, 64, 66, 80]; Everyone in the organization, from data novices to experts, can easily search, collaborate and leverage knowledge on *Alation*.

3.2 Empirical Setting

Our data consists of (1) the usage data of analysts on *Alation* which are automatically collected at the back end; and of (2) the employee information data that we scripted by crawling the eBay personal pages of all analysts in our study. The usage data include the following information

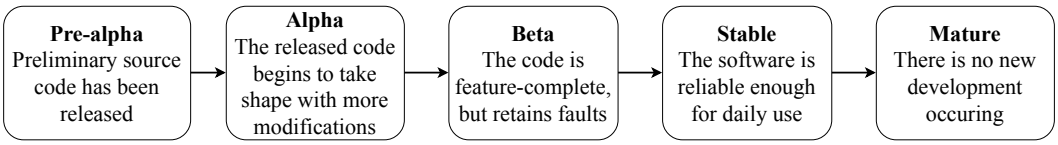

Fig. 2. Similar to query writing, a software development process typically starts with a "Pre-alpha" stage.

spanning four years from January, 2014 to March, 2018: 1) records of entire queries on *Alation*, each item including query id, title, author, time of creation, a brief description of this query and whether this query has been published or not; 2) complete records of query executions on *Alation*, each item including query id, user id, time of execution and number of statements that were executed; 3) complete records of all users viewing query pages on *Alation*; 4) simple personal information for all users on *Alation*, like username, email address, date of joining the *Alation* platform and date of last login. The employee information data track the public information of all data analysts in our study, including employee title, subsidiary area, manager path inside eBay, and location (city campus, country, and building and floor).

To construct our sample of data analysts, we first included all active users who have written at least one query on *Alation* during our study period. We then excluded users who are labeled as Alation employees and users who are authorized as Alation administrators inside eBay Inc. We also excluded users whose eBay employee information is missing on the eBay Intra-net. (That happens when an eBay employee left the company during the study period.) This resulted in an initial set of 2059 users. We then summarized users' hierarchies inside eBay Inc. by parsing their manager paths. Among these 2059 users, there are 17 level-1 employees, 221 level-2, 777 level-3, 748 level-4, 281 level-5 and 15 level-6 (CEO is at level 10). We excluded level-6 and level-1 employees because they are either too senior or too inexperienced to be considered as representative data analysts in our study. The senior product manager who is in charge of *Alation* inside eBay Inc. also confirmed that level 2 - 5 employees are the major users. This resulted in a final data set of 2027 users that have written 101327 queries during the study period. We excluded queries that have never been executed during the study period nor exhibit missing field data; this finally left 79797 queries written by 2001 data analysts for the study. It is important to point out that only about 1 out of 8 queries is ever viewed by an analyst other than the author: out of the 79797 queries, only 10049 queries authored by 1097 data analysts have been viewed by analysts other than the authors.

3.3 Data and Measures

3.3.1 Dependent Variable. To develop a productivity measure for data analysts in our study, we borrow the concept of *Pre-alpha* phase from the software development life cycle [17, 72, 88]. Figure 2 shows a complete development process as summarized by previous researchers [79, 88]. During the *Pre-alpha* phase developers write preliminary source code, which occurs before the Alpha testing.

Figure 3 illustrates the process that a data analyst follows when she is writing a query from scratch on *Alation*. We define $FirstCompletionTime_{i,k}$ as the time interval between the point when the data analyst i clicks the button to create an empty query k and the point when she executes this query for the first time:

$$\text{FirstCompletionTime}_{\text{analyst } i, \text{ query } k} = \text{Timestamp}_{i \text{ first executes } k} - \text{Timestamp}_{i \text{ creates the empty } k} \quad (3)$$

This time interval, comparable to the *Pre-alpha* phase in software development, characterizes the time that a data analyst spends to shape her idea into a testable, prototype query. The longer

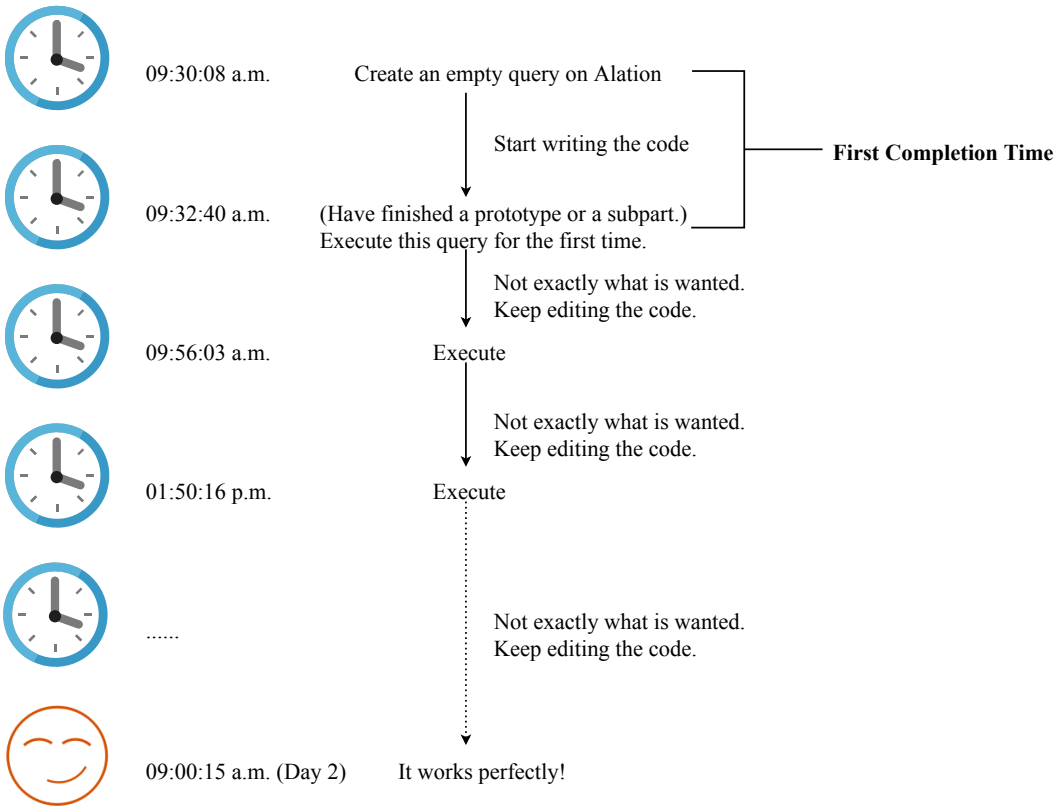

Fig. 3. The First Completion Time is the first stage of a typical query process.

such time interval is, the later this analyst can proceed the following steps. We, therefore, use $FirstCompletionTime_{i,k}$ as a proxy for data analyst productivity. These query writing start time-stamps and the first execution time-stamps are automatically logged at the back end of *Alation*. We believe that the data analysts in our study cannot manipulate this data directly nor have no incentive to act strategically, as only the administrators of *Alation* are informed and have access to these data.

3.3.2 *Explanatory Variables.*

Learning-by-doing. We define $Aggregate\ Direct\ Experience_{i,k}$ as the number of queries that a data analyst i had written by herself on *Alation*, before she clicked the button to create an empty query k during our study period. Because each query uses a particular database, we divide $Aggregate\ Direct\ Experience_{i,k}$ into two parts: 1) $Direct\ Experience\ with\ the\ Focal\ Database_{i,k}$, which is the number of queries using the same database as query k uses, that i has written by herself on *Alation* until she creates query k ; 2) $Direct\ Experience\ with\ Different\ Databases_{i,k}$, which is the number of queries that i has by herself written on *Alation* before she creates k using a database different from the database k uses. Clearly,

$$\begin{aligned} \text{Aggregate Direct Experience}_{i,k} = {} & \text{Direct Experience with the Focal Database}_{i,k} + \\ & \text{Direct Experience with Different Databases}_{i,k} \end{aligned} \tag{4}$$

Learning-by-viewing. We define *Aggregate Indirect Experience*$_{i,k}$ as the number of queries that a data analyst i had viewed from her peers on *Alation* before she clicked the button to create an empty query k during our study period. To differentiate *learning-by-viewing* from different types of data analysts, we further divide *Aggregate Indirect Experience*$_{i,k}$ based on types of the author analysts. For example, *Indirect Experience from All-star*$_{i,k}$ is the number of queries written by other all-star data analysts that i had viewed before she clicked the button to create k. According to our segmentation of data analysts in figure 1, we divide *Aggregate Indirect Experience*$_{i,k}$ as in 5.

$$\begin{aligned}\text{Aggregate Indirect Experience}_{i,k} = {} & \text{Indirect Experience from All-star}_{i,k} + \\ & \text{Indirect Experience from Maven}_{i,k} + \\ & \text{Indirect Experience from Lone-wolf}_{i,k} + \\ & \text{Indirect Experience from Non-star}_{i,k}\end{aligned} \tag{5}$$

To implement the analyst segmentation in Figure 1, we use *Monthly Output of Queries* on *Alation* as a measure of individual data analyst's *output-rate*, and *Monthly Viewers per Query* as a measure of individual data analyst's *viewership*, as defined earlier in equations (1-2). To calculate the *PageRank*, we build a directed, analyst-to-analyst network using the query-viewing data on *Alation*. Each node represents a data analyst in our study. The corresponding graph contains an edge from node A to node B if analyst A has viewed a query written by analyst B. Note that we weigh the edge from node A to node B using the number of distinct queries authored by B that analyst A has viewed. All self-loops have been excluded since we don't consider the behavior that an analyst viewing her own queries as her social interaction. Running the PageRank algorithm implemented in the **R** package **igraph** returns the *PageRank* for all analysts [22, 73].

Figure 4 and 5 are scatter plots illustrating the relationship between *output-rate* and *social influence*. We apply the same segmentation model (see figure 1) to these two plots as follows: in the upper right quadrant we mark data analysts whose *output-rate* and *social influence* are both above the medians as *all-star*; in the lower left quadrant we mark data analysts whose *output-rate* and *social influence* are both below the medians as *non-star*; we mark data analysts who reside in the upper left quadrant as *maven* and data analysts who reside in the lower right quadrant as *lone-wolf*.

3.3.3 Control Variables. Prior work suggests that individual adeptness, multi-tasking and task complexity are likely to affect productivity. A more adept data analyst may write a query faster; a data analyst who has piles of work may become less productive because of stress and pressure; a data analyst may spend more time on writing a complex query that either consists of many statements or uses a complicated database. We incorporate the following control variables to see if *learning-by-doing* and *learning-by-viewing* are associated with additive values over these factors in accelerating data-analyst query-writing. We explain the definitions of these control variables for the scenario in which the analyst i creates and first executes query k.

- *Workload*$_{i,k}$*:* This is the average number of queries that analyst i composes simultaneously together with the focal query k during the *FirstCompletionTime*$_{i,k}$. This definition is inspired by the workload developed in [87] for restaurant workers. For example, suppose *FirstCompletionTime*$_{i,k}$ lasts 40 minutes. During this period, the author data analyst only creates another query that overlaps with the focal query k for 20 minutes. The *Workload*$_{i,k}$, therefore, is (40 min + 20 min)/(40 min) = 1.5 queries.
- *Query Size*$_k$*:* This is the number of statements that were executed in the first execution of query k.
- *Database*$_k$*:* This is a categorical variable that indicates which database query k uses.
- *Saved Query*$_k$*:* This is a binary variable that indicates whether query k has been saved.

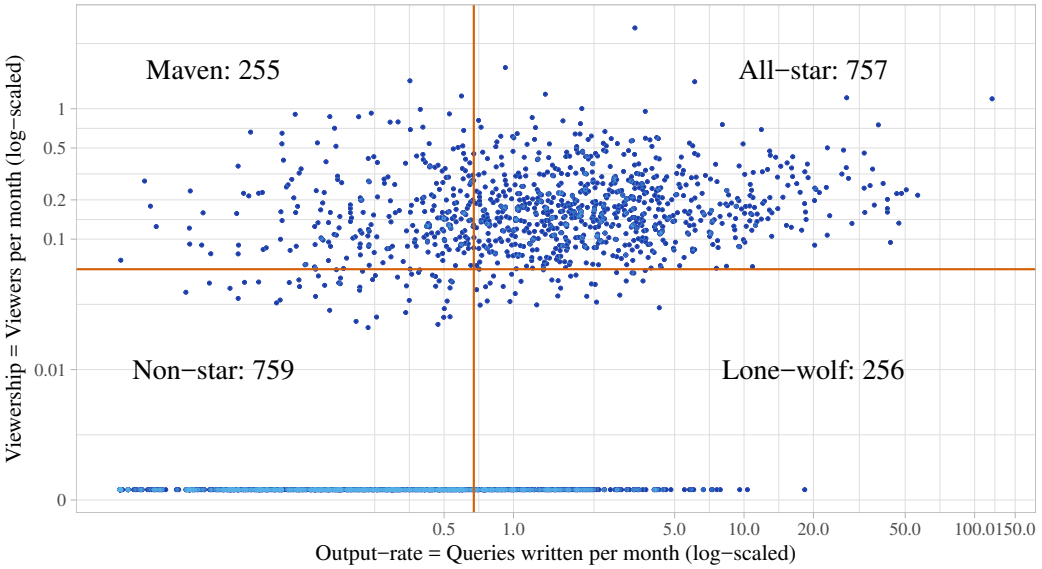

Fig. 4. Segmenting data analysts by Output-rate × Viewership

Note: The median is 0.67 for *Output-rate* and 0.06 for *Viewership*. The correlation between *Output-rate* and *Viewership*is 0.24. Points at the bottom of the plot with zero *Viewership* represent the data analysts whose queries haven't been viewed by any peer; on a log scale these points fall at $-\infty$ but we moved them up for display convenience. These points heavily overlap because only about 1 out of 8 queries is ever viewed by a peer other than the author.

- *Migrated Query*$_k$*:* This is a binary variable that indicates whether part of query k was migrated from a different platform. We acquire such information by parsing the title or description of query k.
- *Tenure on Alation*$_{i,k}$*:* This is the number of months between the date when the author analyst i joined *Alation* and the time she creates query k.
- *eBay Level*$_i$*:* This is a categorical variable that indicates author analyst i's hierarchy in eBay Inc.
- *eBay Sub-area*$_i$*:* This is a categorical variable that indicates author analyst i's subsidiary area in eBay Inc.

We also add month indicators for the number of months since January, 2014 until the creation of query k to control for any environmental difference across time.

4 ANALYSIS AND RESULTS

4.1 Analysis Strategy

We present the descriptive statistics of all variables in the raw data in table 1. Traditional learning curves are typically modeled as exponential forms and are often estimated using a log-linear regression model [3, 6, 57, 60]. Our data are challenging in three ways: (1) our dependent variable (*FirstCompletionTime, in seconds*) is a count variable and is over-dispersed, i.e., the variance is significantly larger than the mean (see table 1); and (2) our dependent variable only takes positive values; (3) our data suffers from potential correlations across observations that are nested within

Table 1. Descriptive statistics of the raw data capturing $N = 79797$ queries written and executed by 2001 data analysts at eBay during 2014-2018.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
FirstCompletionTime	90,711	1,200,935	1	20	262	83,798,568
Aggregate Direct Experience	116.2	200.2	0	16	126	1,775
Aggregate Indirect Experience	35.9	70.8	0	1	35	1,510
Direct Experience with Different Databases	52.4	124.8	0	2	49	1,389
Direct Experience with the focal Database	63.9	113	0	6	68	1,092
Workload	1.1	0.7	1	1	1	28
Tenure on Alation	14.9	13.1	0	4	24	50
Saved Query	0.3	0.5	0	0	1	1
Migrated Query	0.01	0.1	0	0	0	1
Query Size	4.6	117.6	1	1	1	18,095
Output-rate × Viewership Segmentation						
Indirect Experience from All-Star	32.6	64	0	0	32	1,343
Indirect Experience from Lone-Wolf	0.2	0.8	0	0	0	12
Indirect Experience from Maven	2.9	10.5	0	0	1	169
Indirect Experience from Non-star	0.1	0.8	0	0	0	20
Output-rate × PageRank Segmentation						
Indirect Experience from All-Star	32.8	64.4	0	0	32	1,352
Indirect Experience from Lone-wolf	0.01	0.2	0	0	0	11
Indirect Experience from Maven	3	10.7	0	0	1	169
Indirect Experience from Non-star	0.07	0.5	0	0	0	22

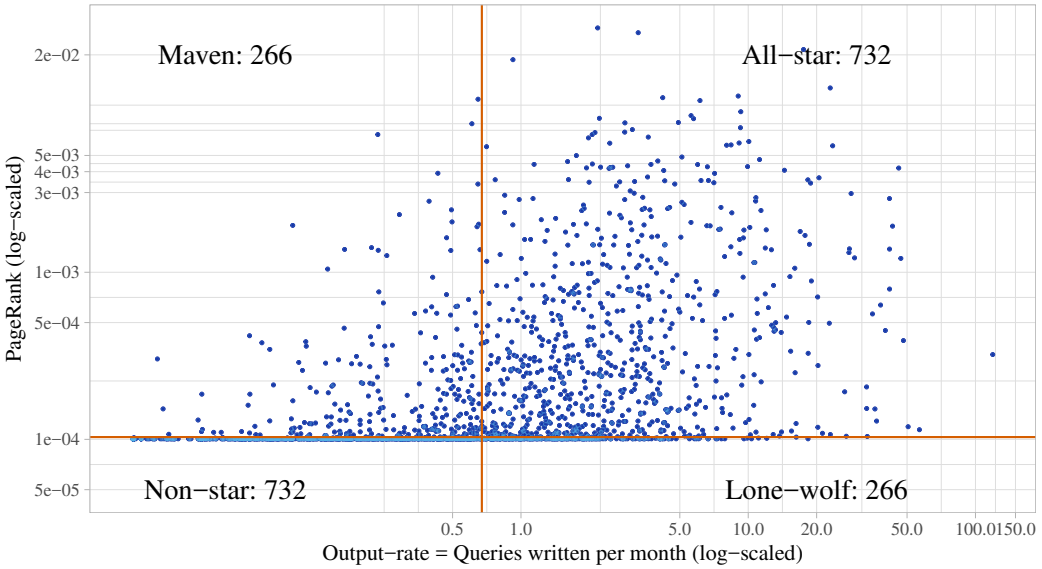

Fig. 5. Segmenting data analysts by Output-rate × PageRank

Note: The median is 0.67 for *Output-rate* and 0.0001 for *PageRank*. The correlation between *Output-rate* and *PageRank*is 0.16.

multiple levels (a data analyst writes multiple queries using different databases over time). To deal with these challenges, we fit a mixed-effects zero-truncated negative binomial regression model. Zero-truncated negative binomial regression models are a class of generalized linear models that are appropriate for non-negative and over-dispersive count data [4]. To account for the three-level nesting of the dataset (from queries to users to databases), we create a mixed-effects model where the experience variables and usage of databases are fixed effects and the unique analyst intercepts are represented as random effects [49]. In R [73], mixed-effects zero-truncated negative binomial models are implemented in the **glmmTMB** package [14].

For ease of comparing the relative importance of the explanatory variables, in the running models we standardize (i.e., normalize to mean zero and unit standard deviation) all variables except for the dependent variable and categorical variables.

4.2 Results

We build six separate models to test our hypotheses. Table 2 presents the model specifications by indicating the inclusive explanatory variables in each model. Particularly, model 1 contains both *Aggregate Direct Experience* and *Aggregate Indirect Experience* ; model 2 contains *Direct Experience with the Focal Database*, *Direct Experience with Different Databases* and *Aggregate Indirect Experience*. Model 3 and Model 4 are built under the *Output-rate* × *Viewership* segmentation of data analysts: model 3 contains *Aggregate Direct Experience* and indirect experience respectively from four types of author analysts: all-star, non-star, maven and lone-wolf; model 4 contains *Direct Experience with the Focal Database*, *Direct Experience with Different Databases*, and indirect experience respectively from four types of author analysts. Except for being built under the *Output-rate* × *PageRank*

Table 2. Results of Zero-truncated Negative Binomial Regressions for Model 1- 6

	Dependent variable: FirstCompletionTime					
			Output-rate × Viewership Segmentation		Output-rate × PageRank Segmentation	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Aggregate direct experience	0.015 (0.032)		−0.071* (0.034)		−0.045 (0.033)	
Direct experience with focal database		−0.057* (0.025)		−0.095*** (0.025)		−0.079*** (0.025)
Direct experience with different database		0.046* (0.022)		−0.012 (0.023)		−0.003 (0.023)
Aggregate indirect experience	−0.029 (0.032)	−0.025 (0.031)				
Indirect experience from all-star			−0.144*** (0.040)	−0.143*** (0.040)	−0.199*** (0.041)	−0.196*** (0.040)
Indirect experience from non-star			0.143*** (0.031)	0.151*** (0.050)	0.077* (0.031)	0.083** (0.03)
Indirect experience from maven			0.277*** (0.033)	0.269*** (0.034)	0.275*** (0.026)	0.268*** (0.033)
Indirect experience from lone-wolf			−0.182*** (0.039)	−0.178*** (0.039)	-0.169*** (0.033)	-0.171*** (0.033)
Constant	-1.36 (24.39)	-0.89 (19.93)	-1.2 (19.54)	-0.74 (16.09)	-0.38 (13.81)	-0.89 (17.96)
Observations	79,797	79,797	79,797	79,797	79,797	79,797
Log Likelihood	−581,467	−581,454	−581,401	−581,396	−581,398	−581,393
Akaike Inf. Crit.	1,163,078	1,163,066	1,162,964	1,162,955	1,162,958	1,162,950
Bayesian Inf. Crit.	1,163,802	1,163,800	1,163,716	1,163,717	1,163,711	1,163,712

Note: *p<0.05; **p<0.01; ***p<0.001

Control variables are omitted in the table.

Table 3. ANOVA results reject both $\beta_1 = \beta_2$ and $\beta_3 = \beta_4 = \beta_5 = \beta_6$

	ΔDf	χ^2	P-Value
Model 1: Model 2	1	14	$< 0.001^{***}$
Model 1: Model 3	3	119.64	$< 0.001^{***}$
Model 1: Model 5	3	125.31	$< 0.001^{***}$
Model 1: Model 4	4	130.62	$< 0.001^{***}$
Model 1: Model 6	4	135.48	$< 0.001^{***}$
Model 2: Model 4	3	116.26	$< 0.001^{***}$
Model 2: Model 6	3	121.48	$< 0.001^{***}$
Model 3: Model 4	1	10.625	0.001^{**}
Model 5: Model 6	1	10.171	0.001^{**}

segmentation, Model 5 and Model 6 have the same specifications as Model 3 and Model 4 have. All six models contain control variables that are listed in 3.3.3.

Note that model {1, 2, 3, 4} and model {1, 2, 5, 6} are nested: if we write Model 4 or 6 as

$$\begin{aligned} g(\mathrm{E}[\mathrm{FirstCompletionTime}_{i,k}]) = & \beta_0 + \beta_1 \mathrm{Direct\ Experience\ with\ Different\ Databases}_{i,k} + \\ & \beta_2 \mathrm{Direct\ Experience\ with\ the\ focal\ Database}_{i,k} + \\ & \beta_3 \mathrm{Indirect\ Experience\ from\ All\text{-}star}_{i,k} + \\ & \beta_4 \mathrm{Indirect\ Experience\ from\ Maven}_{i,k} + \\ & \beta_5 \mathrm{Indirect\ Experience\ from\ Lone\text{-}wolf}_{i,k} + \\ & \beta_6 \mathrm{Indirect\ Experience\ from\ Non\text{-}star}_{i,k} + \\ & \gamma \mathrm{Control\ Variables}_{i,k} \end{aligned} \tag{6}$$

where $g(\cdot)$ is the general link function and is the natural logarithm in our model. Clearly, Model 1 is a special case of Model 4 or 6 (under different segmentations) where $\beta_1 = \beta_2$ and $\beta_3 = \beta_4 = \beta_5 = \beta_6$; Model 2 is a special case of Model 4 or 6 where $\beta_3 = \beta_4 = \beta_5 = \beta_6$; Model 3 is a special case of Model 4 where $\beta_1 = \beta_2$. Model 5 is a special case of Model 6 where $\beta_1 = \beta_2$. Similarly, Model 1 is a special case of Model 2 where $\beta_1 = \beta_2$ and a special case of Model 3 where $\beta_3 = \beta_4 = \beta_5 = \beta_6$; Model 2 is a special case of Model 4 or 6 where $\beta_3 = \beta_4 = \beta_5 = \beta_6$ (under different segmentations).

We use Analysis of Variance (ANOVA) to test the nested models. Table 3 summarizes results of comparisons between nested pairs. We find that the difference in log-likelihoods of Model 1 and Model 2 is statistically significant. So is the difference in log-likelihoods of Model 1 and Model 3. Thus, we can reject both $\beta_1 = \beta_2$ and $\beta_3 = \beta_4 = \beta_5 = \beta_6$ (under output-rate × viewership segmentation) with sufficient confidence. Other comparison results in Table 3 all support this finding. Similarly, the comparison results in Table 3 suggests that under output-rate × PageRank characterization we can reject both $\beta_1 = \beta_2$ and $\beta_3 = \beta_4 = \beta_5 = \beta_6$.

4.2.1 Choosing the best model. The main results of the mixed-effects zero-truncated negative binomial regression on *FirstCompletionTime* are reported in table 2 (see Appendix A. for the full

results). Because we performed mean centering, our baseline was the mean values of all explanatory variables (except for the categorical ones). The coefficients can be interpreted as follows: for an explanatory variable change by one unit (in our case, by one standard deviation), the difference in the logs of expected counts of the dependent variable (*FirstCompletionTime*) is expected to change by the corresponding coefficient, given all the other variables in the model are held constant.

We then use the Akaika Information Criterion (AIC) to evaluate the goodness of fit for each model. Generally, the smaller the AIC, the better the corresponding model over other competing models. Under this criterion, Model 4 is the best among the four models {1, 2, 3, 4} that apply the *output-rate* × *viewership* segmentation of data analysts; Model 6 is the best among the four models {1, 2, 5, 6} that apply the *output-rate* × *PageRank* segmentation of data analysts. We thereby adopt Model 4 and Model 6 to understand *learning-by-doing* and *learning-by-viewing* on data analysts in our study, under the respective segmentation of data analysts. To avoid multicollinearity between explanatory variables in these two final models, we examine the VIF (variance inflation factor) of the set of explanatory variables, comparing against the recommended maximum of 5 [20]. Table 4 shows in our case the VIFs of all explanatory variables in the final models remain well below 2, indicating the absence of multicollinearity.

Table 4. Variance Inflation Factor of Explanatory Variables

Variables	VIF	
	Output-rate × Viewership Segmentation	Output-rate × PageRank Segmentation
Direct experience with the focal database	1.28	1.29
Direct experience with different databases	1.29	1.29
Indirect experience from All-Star	1.52	1.50
Indirect experience from Lone-Wolf	1.30	1.17
Indirect experience from Maven	1.58	1.62
Indirect experience from Non-Star	1.32	1.36
Workload	1.02	1.04
Tenure on Alation	1.04	1.05
Saved Query	1.05	1.05
Migrated Query	1.04	1.05
Query Size	1.01	1.01
Mean VIF	1.26	1.23

4.2.2 Output-rate × viewership segmentation. From Model 4 in Table 2 we find that analysts who have more prior direct experience with the focal database are more likely to have shorter expected *FirstCompletionTime* of writing new queries. Holding all other variables constant, a unit (in our case,

by one standard deviation) increase in a data analyst's direct experience with the focal database is significantly associated with an average 9.5% decrease in the expected *FirstCompletionTime* of a new query. Estimated by the mean of *FirstCompletionTime*, 9.5% decrease is equivalent to 2.4 hours less. A unit increase in direct experience with different databases predicts a 1.2% shorter *FirstComletionTime*, yet not statistically significant.

We also find that viewing queries authored by different types analysts predicts different changes in the expected *FirstCompletionTime* of new queries. Holding all other variables constant, a unit increase in a data analyst's indirect experience (the number of queries she has viewed) from all-star analysts is significantly associated with an average 14.3% decrease (equivalent to 3.6 hours less if estimated by the mean) in the expected time she would spend between creating and first executing a new query; a unit increase in the number of queries the focal data analyst has viewed from lone-wolf is significantly associated with an average 17.8% decrease (equivalent to 4.4 hours less if estimated by the mean). Both indirect experience from non-star and maven is associated with a significant increase in the expected *FirstCompletionTime* of a new query.

4.2.3 Output-rate × PageRank characterization. From Model 6 in Table 2 we find that more prior direct experience with the focal database is associated with a shorter expected *FirstCompletionTime* of new queries. Holding all other variables constant, a unit increase in a data analyst's direct experience with the focal database is significantly associated with an average 7.9% decrease (equivalent to 2.0 hours less if estimated by the mean) in the expected *FirstCompletionTime* of a new query; a unit increase in her direct experience with different databases is linked with the decrease that is neither statistically nor piratically significant (0.3% on average).

We also find that viewing queries authored by different types of analysts is associated with different changes in the expected *FirstCompletionTime* of a new query. Holding all other variables constant, a unit increase in a data analyst's indirect experience from all-star analysts is significantly associated with an average 19.6% decrease (equivalent to 4.9 hours less if estimated by the mean) in the expected *FirstCompletionTime* of a new query; a unit increase in the number of queries the focal data analyst has viewed from lone-wolf authors is significantly associated with an average 17.1% decrease (equivalent to 4.3 hours less if estimated by the mean). Both indirect experience from non-star and maven are related with a significant increase in the expected *FirstCompletionTime* of a new query.

To summarize, our results provide robust support for hypothesis 2 and 4 under both segmentations of data analysts. We have partial support for hypothesis 1 because our results suggest only the direct experience with the focal database is associated with significant improvement in data-analyst productivity. We also have partial support for hypothesis 3 because our results suggest that only viewing queries authored by all-star and lone-wolf analysts is associated with significant improvement in data analyst productivity. Analysts who have viewed more queries authored by maven and non-star analysts are more likely to spend longer *FirstCompletionTime* on a new query in our study.

Furthermore, we find that under the *Output-rate* × *PageRank* segmentation, viewing queries authored by all-star analysts is associated with the largest improvement in data-analyst productivity. In contrast, under the *Output-rate* × *viewership* segmentation, viewing queries authored by lone-wolf analysts is associated with the largest improvement.

5 SUMMARY, DISCUSSION, AND LIMITATIONS

5.1 Summary and Discussion

Our results provide evidence of a statistically-significant association between both *learning-by-doing* and *learning-by-viewing* and data-analyst productivity. Greater direct experience with the focal

database and greater indirect experience from viewing queries authored by 'all-star' and 'lone-wolf' analysts both predict significant improvement in data analyst's productivity. The magnitude of the coefficients in table 2 underscore the practical significance of these associated value. For instance, an increase by one standard deviation in a data analyst's direct experience with the focal database predicts a significant decrease of 9.5% in the expected *FirstCompletionTime*, the equivalent of 2.4 hours average reduction. One standard deviation increase in the number of queries a data analyst has viewed from 'all-star' authors is associated with a significant decrease of 14.3% in the expected *FirstCompletionTime*, under the output-rate × viewership characterization, or a significant decrease of 19.6% under the output-rate × PageRank characterization. These are equivalent to a reduction of 3.6 or 4.9 hours respectively in the expected *FirstCompletionTime*. With 1600 queries created on *Alation* every month in our study period, these numbers accumulate to substantial numbers.

Our study builds on and contributes to the literature of organizational learning and computational social science, as well as that of cognitive psychology literature and learning sciences. Although the relationship between experience and productivity have been extensively studied in manufacturing and service industries, there are only few empirical studies of learning processes among data analysts and none with this granularity of field data. Most existing studies are qualitative studies, such as interviews and surveys, and tend to measure learning using cognitive approaches [12, 46, 47, 52, 53]. To the best of our knowledge, our study may be the first to empirically examine how data analysts learn to speed up writing queries using behavioral (performance-measure) approaches.

First, our results support association between learning from direct experience and productivity. More experience in writing queries is associated with faster *FirstCompletionTime*. This finding is consistent with the study by Kim et. al. that surveyed 793 professional data scientists at Microsoft [53]. Respondents to the survey spoke of 'getting their hands dirty' as one of the best practices to improve data science analysis, and frequently mentioned the desire for hands-on training and practical case studies. We also find that prior experience in using the focal database is associated with greater improvement in productivity than the experience of using a different database. This implies that data analysts obtain more domain knowledge of a specific database through self-practice [48]. This finding also echoes previous studies that demonstrate the greater impact of related experience on productivity [28, 51, 85, 97].

Second, our results provide evidence for the role of "experts" in *learning-by-viewing*. We find that only viewing queries authored by analysts with high output rate is associated with significant improvement in data analyst productivity. This finding not only confirms that circulating the query code can be an effective social practice, but also suggests that interacting with the 'community of practices' is critical for the advance of an analyst's knowledge. This is consistent with the finding of Dabbish et al. that being able to view code authored by others supports better programming [23]. Kim et al. also reported that "respondents expressed the goal of fostering a 'community of practice' across the company" [53]. Despite of this, Kandel et al. found in their interview study that "the least commonly shared resource among data analysts was the analysis code" [46]. Our results give support to the value of collaborative platforms in providing an unfenced channel for collaboration.

Third, our findings suggest that different types of social influence, namely the 'local' influence (as captured by the viewership) vs. the 'global' influence (as captured by the page-rank), are associated with different changes in viewer analysts' productivity. Under both segmentations of analysts , only viewing queries authored by 'all-star' and 'lone-wolf' is associated with a significant decrease in the *FirstCompletionTime* of new queries. Analysts who mostly queries authored by 'non-star' and 'maven' analysts, in contrast, are more likely to spend more of that time. Using the output-rate × viewership segmentation, we find that viewing queries authored by 'lone-wolf' analysts is associated with the largest improvement in *FirstCompletionTime*; using the output-rate × page-rank segmentation, we find that viewing queries authored by 'all-star' analysts is associated with the

largest improvement. These two findings together indicate that the most influential analysts might be those that have few incoming links (fewer direct viewers) but a relatively large contingency of viewers with high *PageRank*. Such analysts could be the 'ultimate stars'.

Overall, regardless of the segmentation of analysts, *learning-by-viewing* is associated with greater productivity improvement than *learning-by-doing* in our study. This result agrees with Lave and co-authors' conclusion that 'engaging in practice may well be a condition for the effectiveness of learning' and is consistent with the anecdotal evidence in their paper that apprentices learn mostly in relations with other apprentices [58].

5.2 Implications

While one must be careful with extrapolations, our results suggest directions for further explorations by managers and designers.

5.2.1 Managerial Implications. First, our findings about *learning-by-doing* provide guidance for developing training plans for data analysts. Based on the result that *learning-by-doing* with the same data bases is associated with significant productivity improvement, managers can encourage data analysts to practice with focus. Similarly, individual contributors on peer production community (e.g. Wikipedia and GitHub) can choose to work on related projects in order to be more productive.

Second, our findings about *learning-by-viewing* provide suggestive evidence of the value of collaborative platforms. Furthermore, the differential associations of *learning-by-viewing* queries authored by different types of analysts can inform performance evaluation and team-building. Organizations that use collaborative platforms can take their employees' star status on such platforms seriously into evaluation of their performance and contribution. Besides, with the identification of star vs. non-star, project managers can team suitable analysts to work together.

Third, our findings help to identify star workers on collaborative platforms. We show it is useful to leverage articulated social measures and observed code-related activity simultaneously. As Marlow et al. put it [62], "Succinctly summarizing expertise based on behavioral data and incorporating evidence of social interactions may support more nuanced impressions and reduce bias. In any type of peer production site where a person shares their work for others to build on, dealing with contributions from others is necessary and important." Collaborative platforms or peer production communities may consider adopting measurements such as viewership and *PageRank* that are discussed in our study or other measurements that fit their context better.

5.2.2 Design Implications. Our results particularly inform the future design of collaborative platforms to support learning. We emphasize providing cues about expertise and star status, by showing that a data analyst learns better if she has more interactions with certain groups of star analysts. Accessible cues about expertise and star status can be very critical for the legitimate periphery participation of the beginners or novice [58]. Researches in corporate domain also suggest that expertise finding is an important task (e.g.[74]) and show that many internal tools have been developed to help people tag their own and others' expertise [82]. Our results suggest that collaborative platforms invest in designing tools to deliberately highlight certain information thereby providing social signals. For example, dashboard pages and activity feeds could signal to data analysts the social significance and technical merits of their peers. The theory of social translucence also confirms the potential for this transparency to radically improve collaboration and learning in complex knowledge-based activities [23]. For learners, analysts or developers can take these cues into effective strategies for coordinating projects and advancing their technical skills [29]. For knowledge contributors, signaling their social influence among the communities may motivate desired behavior yet one always should anticipate unintended consequences.

5.3 Limitations and Future Work

First, our study focuses on the behavioral approach to measure learning. Qualitative studies that describe the cognitive phenomena of analysts are lacking. Central questions like, "what are the analysts learning" or "are they really learning anything" remain to be answered directly. Future work to interview, observe, or survey data analysts will bring insightful answers to these questions. These follow-up qualitative studies asking data analysts to explain their behaviors in creating/viewing queries could complement our findings towards a better understanding of the mechanisms underlying these two learning processes. For example, researchers who find the negative association between viewing queries authored by 'maven' and the viewer's productivity rather intriguing can survey the subgroup of analysts in our study who have viewed most queries from 'mavens'. Perhaps these analysts found it very difficult to match their existing knowledge with those queries authored by 'maven' [93]. Furthermore, we only report the analyst's activity of writing/viewing a query without further reporting which part of the query that the focal analyst has indeed focused on. Such additional information may help us identify what domain-specific aspects of viewing peers' queries or what type of self-practice advance learning efficiency more [33, 34]. Still this requires more qualitative evidence in the future work.

Second, our results show associations without supporting causal inferences. Although we apply mixed-effects models to address potential correlations and incorporate control variables to deal with some confounding factors, our analysis may still suffer from endogeneity issues such as self-selection bias. Nonetheless, the associations we find do motivate randomized experiments in future research to carefully measure causal effects. One possible way to do this is to bring exogenous variations in analysts' exposure to queries authored by peers. Randomizing query-searching results and arbitrarily mask peers' queries to analysts are both practical.

Third, we use *FirstCompletionTime* to measure data-analyst productivity, which is the period of time from a data analyst creating an empty query to executing it for the first time. Other (aggregate) metrics, like the number of queries a data analyst creates in a week or month, may highly depend on the assignment that the analysts has received. Nevertheless, it is possible that *FirstCompletionTime* depends on individual work style. Some empirical evidence in the learning sciences shows two distinct programming styles: tinkering vs. planning [90]. Tinkerers keep exploring new ideas by making adjustments step by step. Planners first develop a clear plan, then do it once and right [77]. Tinkering data analysts may be making small and frequent code changes and thus have shorter *FirstCompletionTime* than planning analysts. Such caveat may be addressed by using a different measure of the coding time: the period of time that the data analyst remains active on the composing page. To obtain such time requires detailed click-stream data that is not accessible to us.

Fourth, our study focuses exclusively on the quantity (productivity) without discussing the quality of queries. Including the quality of a query is not straightforward as different people may have different opinions on what constitutes a *good* query: teammates or project managers may want the query to be well-documented so that it can be reused; IT staff may want the query to be efficient when running on large-scale data; stakeholders or decision-makers may want the query to deliver accurate and insightful answers to their questions; the author may want the query to run smoothly without errors. Again, to operationalize these multiple perspectives on quality requires different levels of data, both qualitative and quantitative, which are not accessible in our data. Future research could extend our study by generating measures of quality (e.g., the number of errors during executions, or the rate of positive responses from peers who have viewed the query) to examine the results with quality-adjusted output.

ACKNOWLEDGMENTS

The authors would like to thank Qian Chang, Rena Fired-Chung, Matt Gardner, Aaron Kalb, Jake Magner, and Dennis Zhang for valuable discussions and helpful feedback. Itai Gurvich, Yue Yin and Jan Van Mieghem are grateful for the open collaboration with Alation and eBay; yet they have no financial interest in, nor received any financial remuneration or funding from, Alation or eBay.

A FULL REGRESSION RESULTS FOR MODEL 1-6 IN TABLE 2

Table 5. Zero-truncated Negative Binomial Regressions for Model 1-6: Full results

	Dependent variable: FirstCompletionTime					
			Output-rate ×Viewership Segmentation		Output-rate×PageRank Segmentation	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Aggregate direct experience	0.015 (0.032)		−0.071* (0.034)		−0.045 (0.033)	
Direct experience with the focal database		−0.057* (0.025)		−0.095*** (0.025)		−0.079*** (0.025)
Direct experience with different database		0.046* (0.022)		−0.012 (0.023)		−0.003 (0.023)
Aggregate indirect experience	−0.029 (0.032)	−0.025 (0.031)				
Indirect experience from all-star			−0.144*** (0.040)	−0.143*** (0.040)	−0.199*** (0.041)	−0.196*** (0.040)
Indirect experience from non-star			0.143*** (0.031)	0.151*** (0.050)	0.077* (0.031)	0.083** (0.03)
Indirect experience from maven			0.277*** (0.033)	0.269*** (0.034)	0.275*** (0.026)	0.268*** (0.033)
Indirect experience from lone-wolf			−0.182*** (0.039)	−0.178*** (0.039)	-0.169*** (0.033)	-0.171*** (0.033)
Workload	1.32*** (0.08)	1.28*** (0.08)	1.33*** (0.084)	1.31*** (0.084)	1.32*** (0.084)	1.31*** (0.084)
Query size	-0.031*** (0.01)	-0.036*** (0.01)	-0.037*** (0.01)	-0.037*** (0.01)	-0.037*** (0.01)	-0.037*** (0.01)
Saved query	2.89*** (0.05)	2.91*** (0.052)	2.89*** (0.05)	2.89*** (0.052)	2.90*** (0.052)	2.90*** (0.052)
Migrated query	1.88*** (0.17)	1.87*** (0.174)	1.88*** (0.175)	1.88*** (0.174)	1.90*** (0.176)	1.89 (0.176)
Tenure on Alation	-0.061 (0.069)	-0.061 (0.069)	-0.063 (0.069)	-0.061 (0.069)	-0.072 (0.069)	-0.071 (0.069)

Continued on next page

Table 5 – continued from previous page

	Dependent variable: FirstCompletionTime					
			Output-rate×Viewership Segmentation		Output-rate×PageRank Segmentation	
	(1)	(2)	(3)	(4)	(5)	(6)
eBay Level: 7	0.011 (0.246)	0.015 (0.246)	0.011 (0.246)	0.011 (0.246)	0.019 (0.246)	0.019 (0.246)
eBay Level: 8	-0.153 (0.25)	-0.156 (0.25)	-0.147 (0.25)	-0.153 (0.25)	-0.141 (0.25)	-0.146 (0.25)
eBay Level: 9	0.183 (0.322)	0.173 (0.321)	0.196 (0.322)	0.183 (0.322)	0.197 (0.322)	0.184 (0.322)
eBay Sub-area: Global Function - HR	-0.629 (2.34)	-0.729 (2.34)	-0.599 (2.34)	-0.629 (2.34)	-0.664 (2.35)	-0.69 (2.34)
eBay Sub-area: Global Function - Legal	0.013 (0.975)	0.107 (0.974)	0.034 (0.974)	0.013 (0.973)	0.102 (0.976)	0.085 (0.976)
eBay Sub-area: Marketplaces	-0.274 (0.202)	-0.269 (0.202)	-0.275 (0.202)	-0.275 (0.202)	-0.276 (0.202)	-0.275 (0.202)
Database	Yes	Yes	Yes	Yes	Yes	Yes
Month	Yes	Yes	Yes	Yes	Yes	Yes
Constant	-1.36 (24.39)	-0.89 (19.93)	-1.2 (19.54)	-0.74 (16.09)	-0.38 (13.81)	-0.89 (17.96)
Observations	79,797	79,797	79,797	79,797	79,797	79,797
Log Likelihood	−581,467	−581,454	−581,401	−581,396	−581,398	−581,393
Akaike Inf. Crit.	1,163,078	1,163,066	1,162,964	1,162,955	1,162,958	1,162,950
Bayesian Inf. Crit.	1,163,802	1,163,800	1,163,716	1,163,717	1,163,711	1,163,712
Note:						*p<0.05; **p<0.01; ***p<0.001

REFERENCES

[1] Daron Acemoglu. 1996. A microfoundation for social increasing returns in human capital accumulation. *The Quarterly Journal of Economics* 111, 3 (1996), 779–804.

[2] Paul J Adams, Andrea Capiluppi, and Cornelia Boldyreff. 2009. Coordination and productivity issues in free software: The role of brooks' law. In *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*. IEEE, 319–328.

[3] Armen Alchian. 1963. Reliability of progress curves in airframe production. *Econometrica: Journal of the Econometric Society* (1963), 679–693.

[4] Paul D Allison and Richard P Waterman. 2002. Fixed-effects negative binomial regression models. *Sociological methodology* 32, 1 (2002), 247–265.

[5] John R Anderson, Lynne M Reder, and Herbert A Simon. 1997. Situative versus cognitive perspectives: Form versus substance. *Educational researcher* 26, 1 (1997), 18–21.

[6] Linda Argote. 2012. *Organizational learning: Creating, retaining and transferring knowledge*. Springer Science & Business Media.

[7] Linda Argote and Dennis Epple. 1990. Learning curves in manufacturing. *Science* 247, 4945 (1990), 920–924.

[8] Linda Argote, Paul Ingram, John M Levine, and Richard L Moreland. 2000. Knowledge transfer in organizations: Learning from the experience of others. *Organizational behavior and human decision processes* 82, 1 (2000), 1–8.

[9] Pierre Azoulay, Joshua S Graff Zivin, and Jialan Wang. 2010. Superstar extinction. *The Quarterly Journal of Economics* 125, 2 (2010), 549–589.

[10] C Lanier Benkard. 1999. *Learning and forgetting: The dynamics of aircraft production*. Technical Report. National bureau of economic research.

[11] Joel Brandt, Philip J Guo, Joel Lewenstein, Mira Dontcheva, and Scott R Klemmer. 2009. Two studies of opportunistic programming: interleaving web foraging, learning, and writing code. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1589–1598.

[12] Matthew Brehmer, Michael Sedlmair, Stephen Ingram, and Tamara Munzner. 2014. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. ACM, 1–8.

[13] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1-7 (1998), 107–117.

[14] Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 9, 2 (2017), 378–400. https://journal.r-project.org/archive/2017/RJ-2017-066/index.html

[15] JohnSeely Brown and Paul Duguid. 2000. Organizational learning and communities of practice: Toward a unified view of working, learning, and innovation. In *Knowledge and communities*. Elsevier, 99–121.

[16] Ronald S Burt. 2009. *Structural holes: The social structure of competition*. Harvard university press.

[17] Raymond PL Buse and Westley R Weimer. 2008. A metric for software readability. In *Proceedings of the 2008 international symposium on Software testing and analysis*. ACM, 121–130.

[18] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. 2012. Business intelligence and analytics: from big data to big impact. *MIS quarterly* (2012), 1165–1188.

[19] Paul Cobb and Janet Bowers. 1999. Cognitive and situated learning perspectives in theory and practice. *Educational researcher* 28, 2 (1999), 4–15.

[20] Patricia Cohen, Stephen G West, and Leona S Aiken. 2014. *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology Press.

[21] Mary M Crossan, Henry W Lane, and Roderick E White. 1999. An organizational learning framework: From intuition to institution. *Academy of management review* 24, 3 (1999), 522–537.

[22] Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal* Complex Systems (2006), 1695. http://igraph.org

[23] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1277–1286.

[24] Eric D Darr, Linda Argote, and Dennis Epple. 1995. The acquisition, transfer, and depreciation of knowledge in service organizations: Productivity in franchises. *Management Science* 41, 11 (1995), 1750–1762.

[25] Sayamindu Dasgupta, William Hale, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Remixing as a pathway to computational thinking. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1438–1449.

[26] Chris Dede, Brian Nelson, Diane Jass Ketelhut, Jody Clarke, and Cassie Bowman. 2004. Design-based research strategies for studying situated learning in a multi-user virtual environment. In *Proceedings of the 6th international*

conference on Learning sciences. International Society of the Learning Sciences, 158–165.
[27] John M Dutton and Annie Thomas. 1984. Treating progress functions as a managerial opportunity. *Academy of management review* 9, 2 (1984), 235–247.
[28] Carolyn D Egelman, Dennis Epple, Linda Argote, and Erica RH Fuchs. 2016. Learning by doing in multiproduct manufacturing: Variety, customizations, and overlapping product generations. *Management science* 63, 2 (2016), 405–423.
[29] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.
[30] K Anders Ericsson and Herbert A Simon. 1980. Verbal reports as data. *Psychological review* 87, 3 (1980), 215.
[31] C Marlene Fiol and Marjorie A Lyles. 1985. Organizational learning. *Academy of management review* 10, 4 (1985), 803–813.
[32] Wai Fong Boh, Sandra A Slaughter, and J Alberto Espinosa. 2007. Learning from experience in software development: A multilevel analysis. *Management Science* 53, 8 (2007), 1315–1331.
[33] Thomas Fritz, Gail C Murphy, and Emily Hill. 2007. Does a programmer's activity indicate knowledge of code?. In *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*. ACM, 341–350.
[34] Thomas Fritz, Gail C Murphy, Emerson Murphy-Hill, Jingwen Ou, and Emily Hill. 2014. Degree-of-knowledge: Modeling a developer's knowledge of code. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 23, 2 (2014), 14.
[35] Thomas Fritz, Jingwen Ou, Gail C Murphy, and Emerson Murphy-Hill. 2010. A degree-of-knowledge model to capture source code familiarity. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 385–394.
[36] Francesca Gino, Linda Argote, Ella Miron-Spektor, and Gergana Todorova. 2010. First, get your feet wet: The effects of learning from direct and indirect experience on team creativity. *Organizational Behavior and Human Decision Processes* 111, 2 (2010), 102–115.
[37] James G Greeno. 1997. On claims that answer the wrong questions. *Educational researcher* 26, 1 (1997), 5–17.
[38] Boris Groysberg, Linda-Eling Lee, and Ashish Nanda. 2008. Can they take it with them? The portability of star knowledge workers' performance. *Management Science* 54, 7 (2008), 1213–1230.
[39] Cynthia Harrington. 2017. Speaking Data Science with an Investment Accent. *CFA Institute Magazine* 28, 4 (2017), 4–7.
[40] N Henke, J Bughin, M Chui, et al. 2017. The age of analytics: competing in a data-driven world. McKinsey Global Institute.
[41] Gerard P Hodgkinson and Paul R Sparrow. 2002. *The competent organization: A psychological analysis of the strategic management process*. Open University Press.
[42] George P Huber. 1991. Organizational learning: The contributing processes and the literatures. *Organization Science* 2, 1 (1991), 88–115.
[43] Robert S Huckman and Bradley R Staats. 2011. Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. *Manufacturing & Service Operations Management* 13, 3 (2011), 310–328.
[44] Anne Sigismund Huff and Mark Jenkins. 2002. *Mapping strategic knowledge*. Sage.
[45] Elina H Hwang, Param Vir Singh, and Linda Argote. 2015. Knowledge sharing in online communities: learning to cross geographic and hierarchical boundaries. *Organization Science* 26, 6 (2015), 1593–1611.
[46] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization & Computer Graphics* 12 (2012), 2917–2926.
[47] Eser Kandogan, Aruna Balakrishnan, Eben M Haber, and Jeffrey S Pierce. 2014. From data to insight: work practices of analysts in the enterprise. *IEEE computer graphics and applications* 34, 5 (2014), 42–50.
[48] Keumseok Kang and Jungpil Hahn. 2009. Learning and forgetting curves in software development: Does type of knowledge matter? *ICIS 2009 Proceedings* (2009), 194.
[49] Raghav Pavan Karumur, Bowen Yu, Haiyi Zhu, and Joseph A Konstan. 2018. Content is King, Leadership Lags: Effects of Prior Experience on Newcomer Retention and Productivity in Online Production Groups. (2018).
[50] Diwas Kc, Bradley R Staats, and Francesca Gino. 2013. Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. *Management Science* 59, 11 (2013), 2435–2449.
[51] Diwas Singh Kc and Bradley R Staats. 2012. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* 14, 4 (2012), 618–633.
[52] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 96–107.

[53] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2017. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions on Software Engineering* 1 (2017), 1–1.
[54] Youngsoo Kim, Ramayya Krishnan, and Linda Argote. 2012. The learning curve of IT knowledge workers in a computing call center. *Information Systems Research* 23, 3-part-2 (2012), 887–902.
[55] Steve WJ Kozlowski. 2012. *The Oxford handbook of organizational psychology*. Organizational Learning and knowledge management, Vol. 1. Oxford University Press.
[56] H Chad Lane and Kurt VanLehn. 2005. Teaching the tacit knowledge of programming to noviceswith natural language tutoring. *Computer Science Education* 15, 3 (2005), 183–201.
[57] Michael A Lapré, Amit Shankar Mukherjee, and Luk N Van Wassenhove. 2000. Behind the learning curve: Linking learning activities to waste reduction. *Management Science* 46, 5 (2000), 597–611.
[58] Jean Lave, Etienne Wenger, and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Vol. 521423740. Cambridge university press Cambridge.
[59] Barbara Levitt and James G March. 1988. Organizational learning. *Annual review of sociology* 14, 1 (1988), 319–338.
[60] Ferdinand K Levy. 1965. Adaptation in the production process. *Management Science* 11, 6 (1965), B–136.
[61] Robert E Lucas Jr. 1988. On the mechanics of economic development. *Journal of monetary economics* 22, 1 (1988), 3–42.
[62] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 117–128.
[63] Tara Matthews, Jilin Chen, Steve Whittaker, Aditya Pal, Haiyi Zhu, Hernan Badenes, and Barton Smith. 2014. Goals and perceived success of online enterprise communities: what is important to leaders & members?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 291–300.
[64] Tara Matthews, Steve Whittaker, Hernan Badenes, Barton A Smith, Michael Muller, Kate Ehrlich, Michelle X Zhou, and Tessa Lau. 2013. Community insights: helping community leaders enhance the value of enterprise online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 513–522.
[65] Anne S Miner and Pamela R Haunschild. 1995. Population-level learning. *Research in Organizational Behavior: an annual series of analytical essays and critical reviews* 17 (1995), 115–166.
[66] Michael Muller, Kate Ehrlich, Tara Matthews, Adam Perer, Inbal Ronen, and Ido Guy. 2012. Diversity among enterprise online communities: collaborating, teaming, and innovating through social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2815–2824.
[67] Sriram Narayanan, Jayashankar M Swaminathan, and Srinivas Talluri. 2014. Knowledge diversity, turnover, and organizational-unit productivity: An empirical analysis in a knowledge-intensive context. *Production and Operations Management* 23, 8 (2014), 1332–1351.
[68] Ikujiro Nonaka. 1994. A dynamic theory of organizational knowledge creation. *Organization science* 5, 1 (1994), 14–37.
[69] Alexander Oettl. 2012. Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science* 58, 6 (2012), 1122–1140.
[70] Julian E Orr. 2016. *Talking about machines: An ethnography of a modern job*. Cornell University Press.
[71] Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J Abadi, David J DeWitt, Samuel Madden, and Michael Stonebraker. 2009. A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 165–178.
[72] James Piggot and Chintan Amrit. 2013. How healthy is my project? open source project attributes as indicators of success. In *IFIP International Conference on Open Source Systems*. Springer, 30–44.
[73] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[74] Daphne R Raban, Avinoam Danan, Inbal Ronen, and Ido Guy. 2012. Impression formation in corporate people tagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 569–578.
[75] Ray Reagans, Linda Argote, and Daria Brooks. 2005. Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management Science* 51, 6 (2005), 869–881.
[76] Ray Reagans, Ella Miron-Spektor, and Linda Argote. 2016. Knowledge utilization, coordination, and team performance. *Organization Science* 27, 5 (2016), 1108–1124.
[77] Mitchel Resnick and Eric Rosenbaum. 2013. Designing for tinkerability. *Design, make, play: Growing the next generation of STEM innovators* (2013), 163–181.
[78] Paul M Romer. 1990. Endogenous technological change. *Journal of political Economy* 98, 5, Part 2 (1990), S71–S102.
[79] Gregor J Rothfuss and K Bauknecht. 2002. A framework for open source projects. *Department of Information Technology. Zurich, Switzerland, University of Zurich* 157 (2002).

[80] Matthew Rowe, Miriam Fernandez, Harith Alani, Inbal Ronen, Conor Hayes, and Marcel Karnstedt. 2012. Behaviour analysis across different types of Enterprise Online Communities. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 255–264.
[81] David Schuler and Thomas Zimmermann. 2008. Mining usage expertise from version archives. In *Proceedings of the 2008 international working conference on Mining software repositories*. ACM, 121–124.
[82] N Sadat Shami, Kate Ehrlich, Geri Gay, and Jeffrey T Hancock. 2009. Making sense of strangers' expertise from signals in digital artifacts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 69–78.
[83] Elliot Soloway, Kate Ehrlich, and Jeffrey Bonar. 1982. Tapping into tacit programming knowledge. In *Proceedings of the 1982 conference on Human factors in computing systems*. ACM, 52–57.
[84] Bradley R Staats. 2012. Unpacking team familiarity: The effects of geographic location and hierarchical role. *Production and Operations Management* 21, 3 (2012), 619–635.
[85] Bradley R Staats and Francesca Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science* 58, 6 (2012), 1141–1159.
[86] Garold Stasser, Dennis D Stewart, and Gwen M Wittenbaum. 1995. Expert roles and information exchange during discussion: The importance of knowing who knows what. *Journal of experimental social psychology* (1995).
[87] Tom Fangyun Tan and Serguei Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* 60, 6 (2014), 1574–1593.
[88] Vinay Tiwari. 2010. Some observations on open source software development on software engineering perspectives. *International Journal of Computer Science & Information Technology (IJCSIT)* 2, 6 (2010), 113–125.
[89] Jason Tsay, Laura Dabbish, and James Herbsleb. 2014. Influence of social and technical factors for evaluating contribution in GitHub. In *Proceedings of the 36th international conference on Software engineering*. ACM, 356–366.
[90] Sherry Turkle and Seymour Papert. 1992. Epistemological pluralism and the revaluation of the concrete. *Journal of Mathematical Behavior* 11, 1 (1992), 3–33.
[91] Marcie J Tyre and Eric Von Hippel. 1997. The situated nature of adaptive learning in organizations. *Organization science* 8, 1 (1997), 71–83.
[92] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2013. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In *Social computing (SocialCom), 2013 international conference on*. IEEE, 188–195.
[93] Anneliese Von Mayrhauser and A Marie Vans. 1995. Program comprehension during software maintenance and evolution. *Computer* 8 (1995), 44–55.
[94] Richard K Wagner and Robert J Sternberg. 1985. Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of personality and social psychology* 49, 2 (1985), 436.
[95] Fabian Waldinger. 2011. Peer effects in science: Evidence from the dismissal of scientists in Nazi Germany. *The Review of Economic Studies* 79, 2 (2011), 838–861.
[96] JP WHITE. 1975. PSYCHOLOGY AND EPISTEMOLOGY-TOWARDS A THEORY OF KNOWLEDGE-PIAGET, J.
[97] Sze-Sze Wong. 2004. Distal and local group learning: Performance trade-offs and tensions. *Organization Science* 15, 6 (2004), 645–656.
[98] Seungwon Yang, Carlotta Domeniconi, Matt Revelle, Mack Sweeney, Ben U Gelman, Chris Beckley, and Aditya Johri. 2015. Uncovering trajectories of informal learning in large online communities of creators. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 131–140.
[99] Louis E Yelle. 1979. The learning curve: Historical review and comprehensive survey. *Decision Sciences* 10, 2 (1979), 302–328.
[100] Lynne G. Zucker, Michael R. Darby, and Marilynn B. Brewer. 1998. Intellectual human capital and the birth of US biotechnology enterprises. *The American Economic Review* 88, 1 (1998), 290–306. http://www.jstor.org/stable/116831

Received April 2018; revised July 2018; accepted September 2018