Mahidol University
Faculty of Medicine Ramathibodi Hospital

RMJ Ramathibodi Medical Journal

Rama Med J | Special Article

The Role and Responsibilities of Data Professionals in Healthcare Organization

Ratchainant Thammasudjarit

Department of Computer Science, Faculty of Science, Srinakharinwirot University, Bangkok, Thailand

Healthcare organizations are increasingly embracing data-driven approaches to enhance patient care, reduce costs, comply with regulations, and drive innovation through data analytics. Such approaches require collaboration between data professionals. Having data professionals such as data scientists, data analysts, data engineers, and machine learning engineers together and managing them to work on a given task is the new challenge for healthcare organization that usually employs domain experts who have data skills at the beginner to intermediate level. Such a practice might work on a small scale. However, for the enterprise level, the large scale of data in cloud environment requires much more than just an intermediate level. This article describes the role and responsibilities of data professionals to contribute to a healthcare organization with illustrative examples from an analytic project. This article would guide healthcare organizations to acquire the right data professionals to the right tasks.

Keywords: Data scientist, Data analyst, Data engineer, Machine learning engineer, Healthcare organization

Rama Med J: doi:10.33165/rmj.2024.47.4.267918

Received: February 23, 2024 **Revised:** October 10, 2024 **Accepted:** October 11, 2024

Corresponding Author:
Ratchainant Thammasudjarit
Department of Computer Science,
Faculty of Science,
Srinakharinwirot University,
114 Sukhumvit 23 Road,
Khlong Toei Nuea, Watthana,
Bangkok 10110, Thailand.
Telephone: +66 2649 5000
Email: ratchainant@g.swu.ac.th

Introduction

In the rapidly evolving landscape of the healthcare industry, the role of data professionals has become pivotal in spearheading transformative changes. Data scientists,[1] data snalysts,[2] data engineers,[3] and machine learning (ML) engineers[4] are at the forefront of this revolution, harnessing the power of data to drive informed decision-making and enhance patient care. These professionals bring a unique blend of technical expertise and healthcare knowledge, enabling the extraction of meaningful insights from vast and complex datasets. Their work ranges from analyzing patient data for improved treatment plans,[5] to implement predictive models for disease outbreaks[5] and ensure the seamless integration of emerging technologies like artificial intelligence (AI) and ML in clinical settings. As healthcare continues to shift towards a more data-centric approach, the contributions of these data experts become increasingly vital, not only in optimizing operational efficiency and managing healthcare costs, but also in paving the way for innovative medical breakthroughs and personalized patient care.[6]

In the healthcare industry, the specialized roles of data scientists, data analysts, data engineers, and ML engineers are integral to effectively utilizing data to improve patient outcomes and operational efficiencies.[5] Data scientists in healthcare focus on methodologies of modeling as the proof of concepts to solve healthcare problems.[5] Data analysts play a crucial role in interpreting healthcare data[7], providing actionable insights for patient care optimization and healthcare management. Data engineers are responsible for constructing robust and secure data pipelines, including infrastructures, and ensuring the integration and accessibility of diverse healthcare data sources.[8] ML engineers,[9]on the other hand, focus on developing intelligent systems and applications deployed from the proof of concepts on the robust infrastructure such as cloud computing. Together, these professionals form a dynamic team, driving the digital transformation in healthcare and enabling a more data-driven approach to patient care and healthcare management.

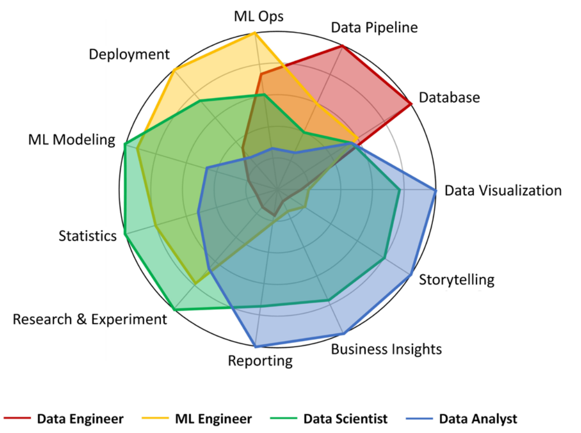
Specific Roles and Skill Spectrum

In a healthcare organization, the collaborative efforts of data scientists, data analysts, data engineers, and ML engineers are essential for harnessing the full potential of data-driven healthcare. Data engineers lay the foundational groundwork by building and maintaining a robust data infrastructure, which ensures the seamless collection, storage, and accessibility of vast healthcare data sets. Utilizing this infrastructure, data scientists aid in research and development with solid scientific methodologies. Data analysts then take these insights and translate them into understandable and actionable information, helping healthcare administrators and practitioners make informed decisions. Meanwhile, ML engineers develop intelligent systems and tools on top of the research and development from data scientists to make production ready. This synergistic approach leverages the strengths of each role to improve patient care, enhance operational efficiency, and drive innovation in healthcare solutions. There are various discussions about the skill spectrum of such data professionals. Since the skill spectrums are slowly changing over time due to the changes in business models, competition in industries, and technological readiness, we only picked one interesting skill spectrum in 2023 obtained from the AI Powered Data Products.[10] Such skill spectrums were discussed broadly in professional communities such as LinkedIn. Figure 1 illustrates the skill spectrum of each data professional. Each category was aligned based on the responsibilities.

Data Engineer

Data engineers play a pivotal role in the healthcare industry by constructing sophisticated data pipelines that

Figure 1. Skill Spectrum of Data Professionals



Abbreviations: ML, machine learning; Ops, operations.

ensure seamless data integration from multiple sources, maintain data quality and consistency essential for patient care, and automate data processes for efficiency and accuracy.[11] Their expertise in designing scalable systems is crucial for handling the increasing volume of healthcare data, enabling real-time data analytics for urgent decision-making, and ensuring compliance with healthcare regulations like the Health Insurance Portability and Accountability Act (HIPAA).[12] Moreover, they are responsible for implementing secure data practices to protect patient privacy and managing the data lifecycle to support cost-effective healthcare operations.Therefore, databases and data pipelines are highly important skills for data engineering.

Data engineers must be versed in ML operations (Ops), deployment, and ML modeling to effectively bridge the gap between data science and production. This ensures the seamless transition of ML models from concept to operational systems. They require an acceptable proficiency level in ML Ops to automate and scale ML pipelines, manage version control, and maintain the infrastructure necessary for monitoring and updating model postdeployment. Furthermore, a basic understanding of ML modeling helps not only for supporting data scientists but also for optimizing data pipelines to work together with ML engineers for feature engineering to enhance the performance of ML algorithms.

Data Scientist

Data scientists are required to possess strong capabilities in research, experimentation, ML modeling, and statistics, as these competencies are foundational to their role.[5] Research prowess is essential not just for innovation and staying current with emerging data science trends but also for understanding the specific challenges and nuances of the domain in which they operate. Experimentation skills are critical for methodically testing hypotheses and refining models through an iterative process that leverages empirical data. Proficiency in ML modeling is at the heart of a data scientist's toolkit, necessary for developing predictive models that drive decision-making and uncovering patterns within complex datasets. Lastly, a deep understanding of statistics underpins a data scientist's ability to interpret data accurately, conduct sound inferences, and validate the performance of models with statistical rigor. These skills intertwine to enable data scientists to convert raw data into actionable insights, driving strategic outcomes and innovation within their organizations.

Data scientists must also be proficient in data visualization, reporting, storytelling, and deriving business insights. Mastery in data visualization allows them to present intricate data in an accessible and intuitive format, simplifying complex concepts for better comprehension by both technical and non-technical stakeholders. Reporting skills are vital for regularly updating project progress and model performance, ensuring transparency and accountability in their work. The art of storytelling is still important to effectively communicating scientific findings to broader audiences. It may not be as good as data analysts but it is practically understandable between project stakeholders.

Data scientists benefit from experience in ML Ops and deployment, areas typically prioritized by ML engineers, as it enhances their ability to collaborate efficiently with the engineering teams, manage the full life cycle of ML models, and take the end-to-end ownership of projects. This knowledge empowers them to develop models that are not only theoretically sound but also practical and scalable within the production environments. Understanding the operational aspects of model deployment helps data scientists build models that are aligned with real-world constraints, such as computational efficiency and regulatory compliance. Additionally, familiarity with ML Ops fosters an appreciation for the importance of a feedback loop in model deployment, allowing for continuous improvement based on real-world performance. Such experience is invaluable in rapidly prototyping and testing models in a production-like environment, leading to more robust and effective solutions. In essence, while not their primary focus, proficiency in ML Ops and deployment equips data scientists to create models more aligned with technical and business realities, thereby enhancing the overall impact of their work.

Data Analyst

Data analysts must be proficient in data visualization, business insights, reporting, and storytelling to effectively translate complex data into actionable insights for business decision-making. Mastery in data visualization is crucial for them to communicate intricate findings clearly, enabling stakeholders to easily grasp patterns, trends, and outliers. Their ability to extract and convey meaningful business insights ensures that data analysis is not just an academic exercise but directly contributes to achieving organizational goals. Reporting skills are equally important, as they provide regular, accurate updates on key metrics and progress, maintaining transparency and accountability in business operations. Furthermore, storytelling is a powerful tool in their arsenal, allowing them to present data in an engaging narrative form. This not only captures the attention of stakeholders but also contextualizes the data, making it more relatable and understandable. Through these combined skills, data analysts bridge the gap between raw data and strategic business decisions, ensuring that organizations can capitalize on their data assets to drive success and growth.[13]

Although not as deeply involved in research, experimentation, and statistics as data scientists, data analysts still benefit from a foundational understanding of these areas to enhance their analytical capabilities. A grasp of research methodologies equips them with the skills to conduct informed analyses, critically evaluate relevant studies, and stay abreast of evolving trends and techniques in the field. Knowledge in experimentation, such as understanding basic experimental design and hypothesis testing, allows them to validate their findings and employ iterative approaches for refinement, crucial in scenarios like A/B testing.[14] Additionally, a fundamental understanding of statistics is essential for accurate data interpretation and quantitative analysis. It enables data analysts to understand data distributions, apply appropriate statistical tests, and make data-driven recommendations with confidence. This blend of skills bolsters the quality of their analysis and facilitates more effective collaboration with data scientists, contributing to robust, insightful decision-making processes within their organizations.

ML Engineer

ML engineers require proficiency in ML Ops, deployment, and ML modeling to effectively operationalize and manage ML projects. Expertise in ML Ops is essential for automating and optimizing the ML workflow, ensuring streamlined development, testing, deployment processes, maintaining reproducibility, and version control of models and datasets. Deployment skills are critical for not just placing models into production environment, but also guaranteeing their performance, efficiency, and reliability in real-world application, along with continuous monitoring and maintenance. Furthermore, a solid grounding in ML modeling, including feature engineering and selection, is vital for creating accurate, efficient, and scalable models. This comprehensive skill set enables ML engineers to not only develop technically sound models but also ensure these models are practical, maintainable, and effectively integrated within the operational infrastructure, thereby playing a crucial role in transforming ML concepts into tangible, value-adding applications.[15]

In summary, each skill in the spectrum into 5 main groups: science (statistics, ML modeling, research, and experiment), infrastructure (databases and data pipelines), production (deployment and ML Ops), knowledge discovery (data visualization and business insights), and communication (reporting and storytelling). Data engineers focus more on infrastructure and production while science, knowledge discovery, and communication are the second priority. Data scientists focus heavily on science skills, while other skills are required at a practical level to bridge with other data professionals and healthcare stakeholders. Data analysts focus on knowledge discovery and communication to transform insights into actions. Lastly, ML engineers focus on deploying scientific projects to production that require skills related to science, infrastructure and production environments.

Collaboration of Data Professionals in a Data Project

We use a real-world problem, COVID-19 screening from a chest x-ray, to explain how data engineers, data scientists, data analysts, and ML engineers work together based on the following flow: problem understanding, data understanding, and implementation.

Problem Understanding

The first task to conduct a data project is problem understanding. Data scientists perform problem abstraction. COVID-19 screening from chest x-ray is considered an image classification task where the input is a chest x-ray image, and the output is radiology findings related to COVID-19 with predictive probability. The heatmap was overlayed on a chest x-ray image to locate such findings.

Data analysts may investigate the prevalence of disease over time categorized by locations by working with various data sources in both internal and external data. They may also investigate the throughput of current screening methods in healthcare institutions. Estimating the workload of a particular healthcare unit to handle such COVID-19 screening cases, including evaluating technological readiness for the proposed screening method by data scientists. These works link to the policy maker of healthcare organizations to decide which site to be deployed the screening model proposed by data scientists.

Data engineers translate the proposed solution from data scientists and analysis plan from data analysts to identify internal and external data sources. Possible payload of the system from normal situations to peak situations is investigated, and designing which data infrastructure suits the problem is the following step.

ML engineers explore the objective and constraints from users to design intelligence solutions that maximize, objective, in this case, is to screen COVID-19 efficiently, subject to constraints such as how fast the system runtime required by the users. The last step is to translate objectives and constraints into proper solution architectures for production.

Data Understanding

Data understanding is the prerequisite task before the experiment and implementation. To create high performance COVID-19 screening model from chest x-ray images, data scientists delve deep into the dataset using advanced analytical techniques such as image segmentation to understand the characteristics and distribution of the x-ray images, formulating hypotheses about features that could indicate specific conditions. They focus on feature analysis, identifying key aspects of the images crucial for accurate classification, and engage in feature engineering to enhance the effectiveness of their models. Evaluating the quality of the image data is another critical task, where they look for issues like varying image resolutions, artifacts, or missing information and strategize on data preprocessing methods to address these challenges. Through exploratory data analysis, including visual examination and statistical methods, data scientists gather insights into the image data, documenting these findings to inform their modeling approach. This phase also involves close collaboration with other team members to ensure a comprehensive understanding of the data and its nuances and to communicate initial insights, laying a solid foundation for developing robust and accurate image classification models.

Data analysts perform a series of essential tasks starting with an initial data exploration to assess the quality and completeness of the dataset, employing descriptive statistics to grasp its basic characteristics. They focus on identifying and selecting key variables crucial for the project's objectives, and conduct correlation analysis to understand the relationships and patterns within the data. Data visualization plays a significant role at this stage, as analysts create various charts and graphs to aid in understanding complex data structures and communicating preliminary findings effectively. They also document methodologies, initial observations, and any potential limitations or assumptions made during the analysis. Throughout this phase, data analysts collaborate closely with data engineers and scientists to ensure alignment on data structures, quality, and analytical approaches, providing valuable feedback that informs subsequent iterations of data collection and preparation. This comprehensive approach allows them to set a solid foundation for more detailed analysis and model development in the later stages of the project.

Data engineers encompass a range of critical tasks that lay the foundation for successful data analysis. They begin by sourcing and extracting data from various internal and external sources, followed by a thorough assessment of data quality, including profiling, to understand its structure, content, and relationships. Data engineers are instrumental in integrating and consolidating data from multiple sources, ensuring consistency and coherence across the dataset. They engage in data cleaning and transformation, rectifying inconsistencies, handling missing values, and reformatting data for optimal usability. Additionally, they set up and manage the necessary data infrastructure, such as databases or data warehouses, and select appropriate tools for data exploration and analysis. Throughout this process, they also create detailed documentation and metadata, clarifying on data sources, preparation methods, and the overall data environment. This comprehensive approach by data engineers ensures that the data is high-quality, reliable, and well-prepared, and is documented for subsequent analysis and insight generation.

ML engineer involves a technical evaluation of the image dataset to ensure its suitability for production. They assess key aspects such as image quality, resolution, and format, identifying any potential limitations in deployment, such as images obtained from different x-ray machine configuration that could affect model performance. ML engineers also plan and implement preprocessing pipelines in production, such as image normalization and augmentation, often automating these processes for efficiency and consistency. They work closely with data scientists to align model architecture appropriate for the classification task.

Implementation

During the implementation phase, data scientists conduct scientific experiments to come up with the x-ray image classification model that yields acceptable performance for practical uses. Working closely with ML engineers for seamless model deployment and with data engineers to ensure the data pipeline supports model training and validation needs. Data scientists conduct iterative testing and refinement of the models, experimenting with different architectures and hyperparameters to fine-tune performance. Throughout this process, they document their methodologies and findings and prepare detailed reports on model performance, providing insights and recommendations for further enhancements. This holistic approach ensures the development of a reliable, efficient x-ray image classification system, primed for deployment in a healthcare setting.

Data analysts implement analytic dashboards to report various key performance indicators (KPIs) related to operation and management when the screening system is deployed. They focus on user experience, ensuring the dashboard is intuitive and meets the specific needs of radiologists and healthcare administrators. This involves extensive data preparation and integration, where analysts gather, clean, and structure data from various sources, including radiology information systems and patient records, by working closely with data engineers. Collaboration is a key aspect, with analysts working closely with stakeholders to align the dashboard with user needs and incorporating feedback for continuous refinement. Rigorous testing is conducted to ensure data accuracy and performance optimization, and they also handle documentation and user training. After implementation, data analysts monitor and update the dashboard based on user feedback and evolving requirements, ensuring it remains a valuable, data-driven decision-making tool in the radiology department.

Data engineers play a pivotal role in building and refining the data infrastructure necessary for effective model development. They design and optimize data pipelines specifically tailored for handling large volumes of image data, ensuring efficient ingestion, processing, and storage of x-ray images. Data engineers are also responsible for managing databases and storage solutions optimized for storing large image datasets, ensuring quick retrieval and efficient handling of these data types. They focus on implementing robust security measures and maintaining compliance with healthcare data regulations to protect patient privacy and data integrity. Setting up and maintaining the necessary computational infrastructure is another key responsibility, which involves configuring high-performance computing resources to handle the intensive demands of image processing and deep learning model training. Collaborating closely with data scientists and ML engineers, data engineers ensure that the infrastructure supports the specific requirements of x-ray image classification, such as graphics processing unit (GPU) optimization for deep learning tasks. They provide ongoing technical support and guidance throughout the project and document the data infrastructure and processes to ensure continuity and clarity. This foundational work by data engineers is crucial for the project's success, enabling the efficient processing and analysis of x-ray images and the development of accurate and reliable classification models.

ML engineers are crucial in transitioning the ML models from the development stage to a real-world clinical setting. Their primary responsibility is deploying these models into production environments, ensuring they are integrated seamlessly with existing healthcare systems or established as standalone applications. This involves optimizing the models for operational efficiency and scalability, managing the required computational resources, and integrating the models with data pipelines for consistent image input and output management. ML engineers are also tasked with ongoing monitoring and maintenance of the models, keeping an eye on performance metrics and updating the models as necessary to adapt to new data or changing requirements.

They work closely with data scientists for technical insights and with information technology (IT) and healthcare professionals for effective system integration while also ensuring that the deployment adheres to healthcare regulations and ethical standards. The documentation of the deployment process and maintenance protocols is also a crucial part of their role, providing clarity and continuity in the project. Through these efforts, ML engineers ensure that the x-ray image classification system is not only technically sound but also practical, reliable, and ready for use in healthcare diagnostics.

Impact and Challenges

Healthcare organizations encounter several challenges in incorporating data professionals such as data scientists, data analysts, data engineers, and ML engineers. One of the primary hurdles is recruitment and talent acquisition. The demand for skilled data professionals is exceptionally high, and healthcare sectors often compete with other industries like technology or finance, which may offer more lucrative salaries and benefits. This competition is compounded by the need for professionals who have expertise in data science and understand the unique nuances of healthcare data and its regulations. Ensuring compliance with stringent healthcare regulations, such as HIPAA in the US, is critical and necessitates professionals who are adept in data management and knowledgeable in specific healthcare data privacy and security practices. This requirement for specialized knowledge creates a significant barrier to entry.

Additionally, integrating data professionals into healthcare organizations involves overcoming cultural and communication barriers. These professionals must effectively collaborate with medical practitioners, who may not have a deep understanding of data science, necessitating a unique combination of communication and technical skills. The challenge here is technical and organizational, as integrating data-driven decision-making into healthcare settings that may traditionally rely on different operational approaches requires careful change management. The complexity and diversity of healthcare data add another layer of difficulty. Healthcare data encompasses a wide array of types, from structured electronic health records to unstructured clinical notes and imaging data. Managing, integrating, and standardizing these diverse data sources for effective analysis and model development is complex. Budget constraints can further complicate these issues, especially for public and nonprofit healthcare entities, limiting their ability to invest in advanced data analytics initiatives and necessary technologies. Lastly, the rapid evolution of data science and AI fields presents a challenge in keeping up with the latest advancements and ensuring continuous professional development of the data workforce. Healthcare organizations must also contend with potential limitations in IT infrastructure and computing resources, essential for supporting large-scale data analytics and ML projects. Overcoming these challenges requires a strategic approach that involves attracting the right talent, fostering an environment conducive to continuous learning and technological adaptation, and effectively integrating data-driven methodologies into healthcare practices.

Conclusions

In healthcare organizations, data scientists focus on developing predictive models and advanced analytics to derive insights from complex healthcare data, often leveraging ML and AI to improve patient outcomes and operational efficiency. Data analysts interpret and visualize data, generating reports to support decision-making, track performance metrics, and identify trends in patient care or resource utilization. Data engineers are responsible for building and maintaining the infrastructure required for data generation, storage, and processing, ensuring data is accessible, clean, and well-organized for analysis. Meanwhile, ML engineers design and implement algorithms and models to automate decision-making and optimize processes, such as diagnostic tools, predictive models for disease progression, or personalized treatment recommendations, ensuring these models are scalable and integrated with existing healthcare systems.

Article Information

Financial Support

No financial support was provided for this study.

Conflict of Interest

The author declares no potential conflicts of interest concerning research, authorship, and/or publication of this article.

References

1. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. *Harv Bus Rev.* 2012;90(10):70-76
2. Kozyrkov C. What great data analysts do-and why every organization needs them. *Harv Bus Rev.* 2018.
3. Desai V, Fountaine T, Rowshankish K. A better way to put your data to work. *Harv Bus Rev.* 2022;100(4):100-107.
4. Pacific Northwest National Laboratory. Machine Learning Engineering: What is Machine Learning Engineering? November 18, 2022. Accessed February 23, 2024. https://www.pnnl.gov/explainer-articles/machine-learning-engineering
5. Subrahmanya SVG, Shetty DK, Patil V, et al. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Ir J Med Sci.* 2022;191(4):1473-1483. doi:10.1007/s11845-021-02730-z
6. Batko K, Ślęzak A. The use of big data analytics in healthcare. *J Big Data.* 2022;9(1):3. doi:10.1186/s40537-021-00553-4
7. Haughom J, Horstmeier P, Wadsworth J, Staheli R, Falk LH. The Changing Role of Healthcare Data Analysts – How Our Most Successful Clients Are Embracing Healthcare Transformation. Health Catalyst; 2017. Accessed February 23, 2024. https://www.healthcatalyst.com/wp-content/uploads/2014/12/whitepaper-Changing-Role-Healthcare-Data-Analysts.pdf
8. Kock-Schoppenhauer, AK. Schreiweis B, Ulrich H, et al. Medical Data Engineering–Theory and Practice. In: Bellatreche L, Chernishev G, Corral A, Ouchani, S, Vain J, eds. *Advances in Model and Data Engineering in the Digitalization Era (MEDI 2021).* Springer Cham; 2021:269-284. doi:10.1007/978-3-030-87657-9_21
9. Habehh H, Gohel S. Machine Learning in Healthcare. *Curr Genomics.* 2021;22(4):291-300. doi:10.2174/1389202922666210705124359
10. Moilanen J, Tatiraju VP. *AI-Powered Data Products: Transforming Data into Profit: A C-Suite Handbook.* MindMote Oy; 2023.
11. Oram A. *The Evolving Role of the Data Engineer: Change and Continuity in Data Practices.* O'Reilly Media Inc; 2020. Accessed February 23, 2024. https://www.qubole.com/wp-content/uploads/2021/03/The-Evolving-Role-of-the-Data-Engineer.pdf
12. Theodos K, Sittig S. Health information privacy laws in the digital age: HIPAA doesn't apply. *Perspect Health Inf Manag.* 2020;18(Winter):1l.
13. Scheider S, Ostermann FO, Adams B. Why good data analysts need to be critical synthesists.

Determining the role of semantics in data analysis. *Future Gener Comput Syst.* 2017;72:11-22. doi:10.1016/j.future.2017.02.046

14. Quin F, Weyns D, Galster M, Silva CC. A/B testing: a systematic literature review. *J Syst Softw.* 2024;211:112011. doi:10.1016/j.jss.2024.112011
15. Ishikawa F, Yoshioka N. How Do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems?-Questionnaire Survey. In: *Proceedings of the 2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP).* IEEE; 2019:2-9. doi:10.1109/ CESSER-IP.2019.00009