# Position Paper: HUSDAT Workshop, CHI 2019

**Krishna Subramanian**
RWTH Aachen University
Aachen 52056, Germany
krishna@cs.rwth-aachen.de

## Abstract

Research in data science has several interesting challenges that facilitate good discussion. In this position paper for the HUSDAT workshop at CHI 2019, I discuss my research background and propose two challenges for discussion.

## Author Keywords

Authors' choice; of terms; separated; by semicolons; include commas, within terms only; required.

## ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous; See [http://acm.org/about/class/1998/]: for full list of ACM classifiers. This section is required.

## Background

I am a PhD candidate at the Media Computing Group, RWTH Aachen University, headed by Prof. Jan Borchers. My research interests align with the topic of this workshop. A part of my research involves investigation of issues that data workers face with learning or performing data science tasks and building potential solutions to mitigate these issues.

**Statsplorer:** In Statsplorer [5], we took up the challenge of making statistical significance testing easier for novices. Novices often need to spend a substantial period of time learning the procedure for selecting the correct significance test. This results in an incomplete just-in-time learning, leading to issues in statistical practice. We showed how Statsplorer, an artifact that limited the need for a data worker to learn the analysis procedure, could help kickstart analysis and help budding students in learning analysis procedures.

**StatPlayground** In StatPlayground [4], a demo at CHI '17, we showed how a two-way direct manipulation interface could help novices learn statistical analysis through discovery learning. StatPlayground allows users to directly manipulate visualizations by direct manipulation, to view how the data properties such as the distribution shape change. In addition, users can also manipulate the resulting statistics to view which data could result in this statistic.

**Study: Problems faced by data workers** Finally, for CHI '19, we looked into how data workers use various scripting languages like R and Python to perform various data science tasks. We tried to model workflows across different modalaties: Notebooks, script files, and interactive consoles. We found that exploratory programming involves excessive code cloning and task switching. We propose to augment existing IDEs with a high-level source code visualization to alleviate these issues.

Over the course of my research, I have identified a few open challenges. I believe that this workshop could be a great opportunity to present them and benefit from the ensuing discussion!

## Challenges #1: How to holistically evaluate artifacts in data science?

Most artifacts that improve data science in one way or another have been validated via classic A/B testing that are based on concrete RQs e.g., Kinetica [3], preliminary usability studies e.g., Variolite [2], and/or qualitative measures of usability or ease of use e.g., Kinetica [3]. While these validation techniques are useful, how can we holistically measure the benefit of an artifact? Often times, there are changes that are more open and hard to measure, such as data worker's change in workflow, which often require long-term studies. Since it is hard for one to pre-emptively know what changes to expect, how can we study this? Do we apply coding techniques from Grounded Theory methodology, e.g., Hypothesis coding, to study this? Or do we utilize the more mathematically grounded user modeling techniques? Of course, another interesting question would be how this influences the reception of such research at CHI and other top venues for HCI?

## Challenges #2: How can we better deal with users?

In my research, I found that performing contextual inquiries is a challenge. Data workers often do not want to be observed as they feel it is intrusive. This leads to compromises e.g., we observe data workers working on fabricated tasks. How can we observe users in a non-intrusive manner? Can remote data analytics be as valid as a live observation? How effective are systems like the 'Me Hate This' button [1] in capturing data workers' issues? Can we do better?

The latter problem is at a more primitive stage and calls for a more open, creative discussion.

## Summary

I have outlined my research background and proposed two challenges that I believe are important and interesting for the data science community. I am looking forward to a great discussion of these challenges and other similar ones at the CHI '19 data science workshop!

## REFERENCES

1. Florian Heller, Leonhard Lichtschlag, Moritz Wittenhagen, Thorsten Karrer, and Jan Borchers. 2011. Me Hates This: Exploring Different Levels of User Feedback for (Usability) Bug Reporting. In *CHI '11: Extended Abstracts of the CHI 2011 Conference on Human Factors in Computing Systems*. 1357–1362. DOI:http://dx.doi.org/10.1145/1979742.1979774

2. Mary Beth Kery, Amber Horvath, and Brad Myers. 2017. Variolite: Supporting Exploratory Programming by Data Scientists. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1265–1276. DOI: http://dx.doi.org/10.1145/3025453.3025626

3. Jeffrey M. Rzeszotarski and Aniket Kittur. 2014. Kinetica: Naturalistic Multi-touch Data Visualization. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 897–906. DOI: http://dx.doi.org/10.1145/2556288.2557231

4. Krishna Subramanian and Jan Borchers. 2017. StatPlayground: Exploring Statistics Through Visualizations. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 401–404. DOI: http://dx.doi.org/10.1145/3027063.3052970

5. Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding Novices in Statistical Analysis. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2693–2702. DOI: http://dx.doi.org/10.1145/2702123.2702347