**Special Topics in Software - Empirical Software Engineering**

Winter 2021
Shurui Zhou

## Overview

Recent decades have witnessed the emergence of two important trends in software systems. First, an increasing number of software technologies have a significant social component (e.g. end-user programming, collaborative development tools). Second, software systems have increased significantly in complexity and size challenging traditional development and testing approaches. Empirical research methods play a key role in the evaluation of tools and technologies, and in testing the social and technical theories they embody.

This course provides an overview and hands-on experience with a core of qualitative and quantitative empirical research methods, including interviews, qualitative coding, survey design, and large-scale mining and analysis of data. There will be extensive reading with occasional student presentations about the reading in class, weekly homework assignments, and a semester-long research project for which students must prepare in-class kickoff and final presentations as well as a final report.

We will focus on software engineering related research questions in readings and assignments. Students will mine and integrate data from and across online software repositories (e.g., GitHub and Stack Overflow) and employ a spectrum of data analysis techniques, ranging from statistical modeling to social network analysis. For the final research project, we encourage students to come up with a research question of interest to themselves. The delivery will be a research paper, and one or more empirical methods presented in class have to be part of the paper.

---

## Learning Goals

The learning goals describe what I want students to know or be able to do by the end of the semester. I evaluate whether learning goals have been achieved through assignments, written project reports, and in-class presentations. All learning goals are roughly written in a form "after taking this class, the student should be able to ...".

- Summarize and interpret a body of literature on a particular topic; identify gaps in the literature; write a literature review
- Formulate and motivate research questions
- Understand what research designs and research methods are available for empirical research

- Compare the suitability of different research designs and research methods in different scenarios; explain the relative strengths and weaknesses
- Design empirical studies for different purposes (e.g., evaluating a tool, understanding a phenomenon); choose appropriate methods and defend the choice
- Combine research methods in a mixed-methods design
- Collect and analyze qualitative and quantitative data
- Design interview protocols and user surveys
- Code qualitative data
- Mine data from online repositories
- Run statistical tests and interpret results
- Build, validate, and interpret regression models
- Draw conclusions from empirical data
- Present results verbally and in writing

## Schedule

We cover the following topics (slides or notes posted when available):

| Week | Topic |
|---|---|
| 1 | Introduction |
| 2 | Literature Review and Theory |
| 3 | Interviews |
| 4 | Grounded Theory |
| 5 | Surveys |
| 6 | Introduction to Measurement |
| 7 | Experimentation |
| 8 | Quasi-experimental Design & Linear Regression |
| 9 | Time Series Analysis |
| 10 | Mixed-methods |
| 11 | Text Mining |
| 12 | Social Network Analysis |

## Evaluation

Evaluation will be based on the following approximate percentages:

- 40% assignments
- 50% research project
    - 10% initial project description (proposal)
    - 2% interim report
    - 8% final presentation
    - 30% final report
- 10% participation and in-class presentations

## Syllabus

1. **Contrasting methods**
    a. **Method:**
    - (Ch.1) Creswell, John W., and J. David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

    b. **Reading**
    - Easterbrook, Steve, et al. "**Selecting empirical methods for software engineering research.**" Guide to advanced empirical software engineering. Springer, London, 2008. 285-311.
    - Bogart, Christopher, et al. "**How to break an API: cost negotiation and community values in three software ecosystems**." Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering. 2016..
    - Raemaekers, Steven, Arie Van Deursen, and Joost Visser. "**Semantic versioning versus breaking changes: A study of the maven repository**." *2014 IEEE 14th International Working Conference on Source Code Analysis and Manipulation*. IEEE, 2014.

    c. **Assignment:**
    Read the assigned chapter and papers, and consider the differences in methods, the research questions they addressed, and the evidence they used to reach their conclusions. What does this tell you about the differences between qualitative and quantitative methods? In general, when is each type of method appropriate? What weaknesses does each method suffer from? Summarize your conclusions in 1-2 pages, and be prepared to informally present and discuss them in class.

2. **Literature Review and Theory**
    a. **Method**
    - (Ch.2 &3) Creswell, John W., and J. David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

    b. **Reading**
        - **Theory-oriented papers:**
            - (Ch. 1) Mokyr, Joel. The gifts of Athena: Historical origins of the knowledge economy. Princeton University Press, 2002.
            - Xiao, S., Witschey, J., & Murphy-Hill, E. (2014). Social influences on secure development tool adoption: why security tools spread, Proceedings of the 17th

ACM conference on Computer supported cooperative work & social computing (pp. 1095-1106): ACM.

**c. Assignment**

Critique the Xiao et al paper, specifically with respect to its use of theory. For example, you might consider

- Did the theory play a major role in framing the research questions?
- Was the methodology influenced by the theory?
- How did the theory impact the data that was collected and the way it was analyzed?
- How important was the theory to this research? Does it help explain the results? Does it help generate new questions?

**3. Interviewing**

**a. Method**

- Powell, Martine B., Ron P. Fisher, and Rebecca Wright. "Investigative interviewing." *Psychology* and law: An empirical perspective (2005): 11-42.
- (Ch 4&6).Interviewing as qualitative research: A guide for researchers in education and the social sciences: Teachers college press.
- Cassell, Catherine, and Gillian Symon, eds. *Essential guide to qualitative methods in organizational research*. Sage, 2004.

**b. Example**

- Grinter, Rebecca E., and Leysia Palen. "Instant messaging in teen life." *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. 2002.
- Chattopadhyay, Souti, et al. "What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.

**c. Assignment**

Think about a research idea that related to your own area and you need to collect information from stakeholders through interview. (see examples above)

1) decide what your purpose is and write a sentence describing it.
2) develop an interview protocol. The protocol can be short, focusing on exactly what you are interested in. You should anticipate short interviews, perhaps 15-20 minutes at most.
3) conduct two interviews.
4) be prepared to tell the class what you learned, how the interviews went, any problems or lessons you can share. In future classes, we will learn more structured ways of analyzing qualitative data such as interview transcripts.

**4. Grounded Theory**

**a. Method**

- (Ch 9) Creswell, John W., and J. David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.
- (Ch 4&5) Miles, Matthew B., A. Michael Huberman, and Johnny Saldana. "Qualitative data analysis: A methods sourcebook." (2014).

**b. Example**

- Razavi, M. N., & Iverson, L. (2006). A grounded theory of information sharing behavior in a personal learning space, Proceedings of the ACM Conference on Computer Supported Cooperative Work (pp. 459-468).
- de Souza, C. R., & Redmiles, D. F. (2008). An empirical study of software developers' management of dependencies and changes, Proceedings of the 30th International Conference on Software Engineering (pp. 241-250).

c. **Exercise**

Transcribe the interviews you recorded last week. Write down a research question (or two) that you think you can answer with the interviews. Develop a suitable coding scheme for the collaborative writing interviews you performed, and apply the codes either to your detailed notes or (preferably your transcription of the interviews). Write an analytic memo (2-3 paragraph) based on these codes.

5. **Survey**

a. **Method**

- (Ch 5) Creswell, John W., and J. David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.
- (Ch 1&5) Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons, 2014.

b. **Example**

- Henne, B., Harbach, M., & Smith, M. (2013). Location privacy revisited: factors of privacy decisions. Extended Abstracts on Human Factors in Computing Systems.
- Shklovski, I., Mainwaring, S. D., Skúladóttir, H. H., & Borgthorsson, H. (2014). Leakiness and creepiness in app space: perceptions of privacy and mobile app use, Proceedings of the ACM conference on Human factors in computing systems (pp. 2347-2356): ACM.

c. **Assignment**

Think about what you learned about from the interviews you conducted, and select an issue for further investigation that would be appropriate for a survey method. Design a short (8-10) item questionnaire that is well designed to address this issue, and enter the survey into an online tools such as surveymonkey. In a short writeup, provide a link to the questionnaire, discuss your selection of open versus closed ended questions, any issues that arose in the wording or presentation of questions, what population you would sample, and how you would invite participants.

6. **Intro Quantitative Analysis**

a. **Method**

- (Ch 10) **Analysis and Interpretation**. from C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012
- (Ch 6) **Statistical Methods and Measurement**. from F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering. Springer 2008 (similar content as the Wohlin chapter but slightly different presentation; read one or the other)
- (Ch 6) **Hypothesis Testing.** from MacKenzie. Human-Computer Interaction. Elsevier 2013

b. **Example**

- Filippova, A., Trainer, E., & Herbsleb, J. D. (2017). From diversity by numbers to diversity as process: supporting inclusiveness in software development teams with brainstorming. In Proceedings of the 39th International Conference on Software Engineering (pp. 152-163). IEEE. [focus on the quantitative analysis of survey responses]
- Vasilescu, B., Filkov, V., & Serebrenik, A. (2015). Perceptions of diversity on GitHub: A user survey. In Proceedings of the Eighth International Workshop on Cooperative and Human Aspects of Software Engineering (pp. 50-56). IEEE. [focus on the quantitative analysis of survey responses]
- Kaptein, M., & Robertson, J. (2012, May). Rethinking statistical analysis methods for CHI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114). ACM. [focus on the threats to validity and ways to mitigate]

**c. Assignment**

In this assignment, you will practice basic quantitative data analysis methods. We will mine the 2017 Open Source Survey results from Zenodo: https://zenodo.org/record/806811 (https://zenodo.org/record/806811#.W7PD0y_Mz1L)

Formulate two research questions about participating in open source development, motivate them in 1-2 paragraphs with a few citations to relevant literature, and answer them using a quantitative analysis of data, e.g., based on ANOVA or multiple linear regression. Go beyond the basic frequency counts from R. Stuart Geiger's paper and focus your research questions on correlations, regressions, or descriptive breakouts between subgroups.

7. Experiment

**a. Method**

(Ch 1,2&8) **Experiments and Generalized Causal Inference, Statistical Conclusion Validity and Internal Validity, Randomized Experiments.** Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference: Wadsworth Cengage learning

**b. Example**

- Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Single versus double blind reviewing at WSDM 2017. arXiv preprint arXiv:1702.00502.

**c. Assignment**

Design an experiment based on your previous analysis of the GitHub open source survey results. The experiment should allow you to either test (some of) the same hypotheses you explored statistically, or an interesting or suggestive finding emerging out of that analysis.

Prepare a very short report describing your experiment design, including:

- your experimental hypothesis,
- a description of the experimental and control groups,
- your experimental procedure,

- acquiring or preparing any materials you need,
- the type of participants you require and how you would recruit them,
- how you would analyze the data,
- analysis of the threats to validity

**8. Quasi-experimental Design & Linear Regression**

**a. Method**

- (Ch 2) **The Simple Regression Model** Woolridge, J. M. (2003). Introductory econometrics: A modern approach. Thomson, Mason. [skim]
- (Ch 1&2) **General Aspects of Fitting Regression Models.** F.E. Harrell, Jr., Regression Modeling Strategies, Springer Series in Statistics, Chapters 1&2 - Regression general aspects: [Chapter 1: skim] [Chapter 2: read 2.1--2.3, 2.7]

**b. Example**

- Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A. L. (2016). Quantifying the evolution of individual scientific impact. Science, 354(6312), aaf5239.
- Lim, S. (2009). How and why do college students use Wikipedia? Journal of the Association for Information Science and Technology, 60(11), 2189-2202.
- Bird, C., Nagappan, N., Devanbu, P., Gall, H., & Murphy, B. (2009). Does distributed development affect software quality? An empirical case study of Windows Vista. Communications of the ACM, 52(8), 85-93.

**c. Assignment**

We are going to use the Aminer data set (https://www.aminer.org/aminernetwork), which contains information on papers, paper citations, authors, and author collaborations for more than 1.7 million computer science authors, to conduct quantitative analysis.

Please Formulate one research question about academic publishing / scientific impact, motivate it in 1-2 paragraphs with 3+ citations to relevant literature, and answer it using a quantitative analysis.

**9. Time Series Analysis**

**a. Method**

- Cowpertwait, P. S., & Metcalfe, A. V. (2009). Introductory time series with R. Springer Science & Business Media.
- (Ch 10.) **Basic Regression Analysis with Time Series Data**. Woolridge, J. M. (2003). Introductory econometrics: A modern approach. Thomson, Mason.

**b. Example**

- Kenmei, B., Antoniol, G., & Di Penta, M. (2008). Trend analysis and issue prediction in large-scale open source systems. In Software Maintenance and Reengineering, 2008. CSMR 2008. 12th European Conference on (pp. 73-82). IEEE.
- Trockman. A., Zhou, S., Kästner, C., & Vasilescu. B. (2017). Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem.

**c. Assignment**

Use the same AMiner data set from the previous assignment to answer a research question of your choosing with an interrupted time-series design. For example, you can

use the affiliations data to answer: Do researchers publish increasingly more papers after they join / leave UofT, compared to before? (in this case the "intervention" is joining / leaving UofT) Be creative!

**10. Mixed-Methods**

a. **Method**

- (Ch 10) Creswell, John W., and J. David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.
- Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. MIS quarterly, 37(1), 21-54.

b. **Example**

- Greiler, M., Deursen, A. V., & Storey, M. A. (2012). Test confessions: a study of testing practices for plug-in systems. In Proceedings of the 34th International Conference on Software Engineering (pp. 244-254). IEEE Press.
- Trockman. A., Zhou, S., Kästner, C., & Vasilescu. B. (2017). Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem.

c. **Assignment**

How would you turn your quantitative studies from last weeks (either regression or time series analysis) into a mixed-methods design? Pick one of the studies. Write 1-2 paragraphs describing the mixed-methods design. Which method(s) would you add? How would you combine them? Which specific threats to validity of the quantitative study would this mixed-methods approach help reduce?

**11. Text Mining**

**a. Method**

- (Ch 1) Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing (Vol. 999). Cambridge: MIT press.
- (Ch 3, 6) **Analyzing text in software projects, Latent Dirichlet Allocation.** Bird, C., Menzies, T., & Zimmermann, T. (Eds.). (2015). The Art and Science of Analyzing Software Data.

**b. Example**

- Asuncion, H. U., Asuncion, A. U., & Taylor, R. N. (2010, May). Software traceability with topic modeling. In Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1 (pp. 95-104). ACM.
- Wang, Y. C., Kraut, R., & Levine, J. M. (2012, February). To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 833-842). ACM.
- Tan, C., & Lee, L. (2015, May). All who wander: On the prevalence and characteristics of multi-community engagement. In Proceedings of the 24th International Conference on World Wide Web (pp. 1056-1066).

**c. Assignment**

Add a research question to either your previous regression or time series studies and use LDA to answer it (or use LDA as an additional method in a mixed-methods design). Ideally use the LDA to calculate some feature(s) and use this feature in a multiple regression model.

**12. Social Network Analysis**

a. Method

- (Ch 1,2 &9) Graph Theory, Community. “The New Science of Networks" by Albert-László Barabási. Cambridge University Press, 2016

b. Example

- Bird, C., Pattison, D., D'Souza, R., Filkov, V., & Devanbu, P. (2008). Latent social structure in open source projects. In Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering (pp. 24-35). ACM.
- Backstrom, L., & Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on Facebook. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 831-841). ACM.

c. Assignment

Explore a research question of your choice on the AMiner collaboration data set, such that your methods include (1) social network analysis, and (2) multiple regression.