Goldfish Bowl Panel: Software Development Analytics

Tim Menzies
West Virginia University
Morgantown, WV, USA
tim@menzies.us

Thomas Zimmermann
Microsoft Research
Redmond, WA, USA
tzimmer@microsoft.com

***Abstract*—Gaming companies now routinely apply data mining to their user data in order to plan the next release of their software. We predict that such *software development analytics* will become commonplace, in the near future. For example, as large software systems migrate to the cloud, they are divided and sold as dozens of smaller apps; when shopping inside the cloud, users are free to mix and match their apps from multiple vendors (e.g. Google Docs' word processor with Zoho's slide manager); to extend, or even retain, market share cloud vendors must mine their user data in order to understand what features best attract their clients. This panel will address the open issues with analytics. Issues addressed will include the following. What is the potential for software development analytics? What are the strengths and weaknesses of the current generation of analytics tools? How best can we mature those tools?**

***Keywords*—analytics; empirical software engineering; mining software repositories; industry;**

I. Topic and Relevance

In software development we have lots of data. For example check-ins, work items, and test executions are recorded in software repositories such as CVS, Subversion, GIT, and Bugzilla. Telemetry data reflects how customers experience software, which includes application and feature usage and exposes reliability. The sheer amount is impressive: For the 10K projects monitored by the web-page http://CIA.vc every 17 seconds a commit takes place. The open source platform SourceForge.Net hosts over 300K projects, and according to Github.com 1M people host 2.9M GIT repositories. The bug database of the Mozilla Firefox projects now contains almost 700K reports according to Ohloh.Net (all October 2011).

Analytics takes data and turns it into insight to inform better development decisions [1]. While analytics is commonly used in business, notably in marketing to better reach and understand customers, application in software development has been somewhat limited. Fortunately, in the past years there have been great efforts that turned software development data into insight. For example, Bird et al. [2] showed empirically that distributed development had negligible impact on the quality of Windows Vista because the development process utilized mitigated the risks. The Office team at Microsoft extensively uses telemetry data collected from customers to inform design decisions by answering questions like *"how frequently is a command used?"* and *"how many files contain this feature?"* [3][4]. This is not just a phenomenon at Microsoft, many researchers and companies have recognized the potential of analytics: The game development company Zynga extensively leverages data to guide the development of their online games [5]. Furthermore over the past few years several researchers have proposed analytics for software development [6] [7].

Similar to today's society and businesses [8], software development is at the crossroads to become more data-driven. With Web services and the Cloud the amount of data will explode, but also the opportunities to gain insight. To make the right decisions during this transition it is important for us to better understand the data and analytics needs. This goldfish panel is a first step into this direction.

II. Goals of the Panel

The goals of the goldfish panel are to raise awareness of software development analytics in the SE community and to discuss several topics, including but not limited to:

- **Data collection.** We need to rethink how we collect data. Often traditional GQM approaches do not work in industry because collecting new data takes too long. What decisions can be made on available data?
- **Data quality.** To make decisions based on data, the quality of the data has to be high [9]. How can we achieve high data quality at little cost?
- **Data privacy.** Data can be very dangerous when used inappropriately. How can we ensure proper use of data?
- **Understanding data needs.** What are the data and information needs of developers and managers for data-driven decision making?
- **User experience.** What is the best ways to surface and interact with software development data and analysis?
- **Education.** How can we prepare software engineers for data analysis and data-driven decision making?

III. Format of the Panel

The panel will be organized as an open fishbowl conversation, which works as follows:

> *"Four to five chairs are arranged in an inner circle. This is the fishbowl. The remaining chairs are arranged in concentric circles outside the fishbowl. A few participants are selected to fill the fishbowl, while the rest of the group sits on the chairs outside the fishbowl. In an open fishbowl, one chair is left empty. [...] The moderator introduces the topic and the participants start discussing the topic. The audience outside the fishbowl listens in on the discussion. In an open fishbowl, any member of the audience can, at any time, occupy the empty chair and join the fishbowl. When this happens, an existing member of the fish-*

bowl must voluntarily leave the fishbowl and free a chair. The discussion continues with participants frequently entering and leaving the fishbowl. [...]"

via: http://en.wikipedia.org/wiki/Fishbowl_(conversation)

IV. Initial Panel Membership

For the initial panel, the following researchers have confirmed their attendance at the time of this writing.

- **Lionel Briand** is a Professor and FNR PEARL chair at the University of Luxembourg's Interdisciplinary Centre for Security, Reliability and Trust. His research interests include model-driven development, testing and verification, search-based software engineering, and empirical software engineering with a strong focus on industry-driven research and innovation. He was awarded the IEEE Computer Society Harlan Mills award for his work on model-based testing and verification.
- **Dieter Rombach** is the Head of the Research Group for Software Engineering (AGSE) at University of Kaiserlautern and the Head of the Fraunhofer Institute for Experimental Software Engineering. As one of the pioneers of experimental software engineering research, he transferred many techniques, methods, and tools into the industrial practice.
- **Pete Rotella** and **Sunita Chulani** are Senior Advisory analysts at Cisco Systems focused on analyzing, modeling, tooling and reporting all sources of software engineering data including in-process metrics, field data, customer satisfaction data, revenue and bookings data. Pete Rotella has a Masters from Duke University and Sunita Chulani has a PhD in Software Engineering from University of Southern California. Combined they have more than 6 decades of experience in software engineering analysis.
- **Dongmei Zhang** is Research Manager of the Software Analytics group at Microsoft Research Asia. She has vast experience with successful technology transfer to many product groups inside Microsoft and how practitioners take actions on the insights produced by analytics solutions [10].

In addition, we will personally invite several researchers as follow-up members of the panel.

V. Social Media

For documentation of the outcomes of the panel, please visit the software development analytics blog:

http://analytics12.blogspot.com

VI. About the Organizers

Tim Menzies has been working on advanced modeling and artificial intelligence since 1986. He received his PhD from the University of New South Wales, Sydney, Australia and is the author of 200+ refereed papers. A former research chair for NASA, Dr. Menzies is now an associate professor at the West Virginia University's Lane Department of Computer Science and Electrical Engineering. He currently serves as the chair of the steering committee of the PROMISE conference on repeatable software engineering experiments and will be PC co-chair for IEEE ASE 2012. For more information, visit his web page at http://menzies.us.

Thomas Zimmermann is a researcher at Microsoft Research and an adjunct professor at the University of Calgary His research focuses on systematic mining of software repositories to conduct empirical studies and to build new tools. He co-organized a working session on Myths in Software Engineering, the DEFECTS workshops (2008 and 2009) as well as the RSSE workshops on recommender systems (2008 and 2010). He served as PC co-chair for the MSR conference on mining software repositories (MSR 2010 and 2011). For more information, visit http://thomas-zimmermann.com.

References

[1] T. Davenport, J. Harris, and R. Morison. *Analytics at Work.* Harvard Business School Publishing Corporation, Boston, MA, 2010.

[2] Christian Bird, Nachiappan Nagappan, Premkumar Devanbu, Harald Gall, and Brendan Murphy. *Does distributed development affect software quality? An empirical case study of Windows Vista.* In ICSE '09: Proceedings of the 31st International Conference on Software Engineering, pp. 518-528, 2009.

[3] T. Briggs. *How does usage data improve the office user experience?* http://blogs.technet.com/b/office2010/archive/2010/02/09/how-does-usage-data-improve-the-office-user-experience.aspx, Feb 2010.

[4] P. Koss-Nobel. *Data driven engineering: Tracking usage to make decisions.* http://blogs.technet.com/b/office2010/archive/2009/11/03/data-driven-engineering-tracking-usage-to-make-decisions.aspx, Nov 2009.

[5] Ken Rudin. Actionable Analytics at Zynga: *Leveraging Big Data to Make Online Games More Fun and Social.* http://tdwi.org/videos/2010/08/actionable-analytics-at-zynga-leveraging-big-data-to-make-online-games-more-fun-and-social.aspx

[6] R. P. Buse and T. Zimmermann. *Analytics for software development.* In Proceedings of the FSE/SDP Workshop on the Future of Software Engineering Research, November 2010.

[7] A. Marcus and T. Menzies. *Software is Data Too.* In Proceedings of the FSE/SDP Workshop on the Future of Software Engineering Research, November 2010.

[8] T. May. *The New Know: Innovation Powered by Analytics.* Wiley, 2009.

[9] Christian Bird, Adrian Bachmann, Eirik Aune, John Duffy, Abraham Bernstein, and Premkumar Devanbu. *Fair and Balanced? Bias in Bug-Fix Datasets.* In Proceedings of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Amsterdam, Netherlands,2009.

[10] Dongmei Zhang, Yingnong Dang, Jian-Guang Lou, Shi Han, Haidong Zhang, and Tao Xie. Software Analytics as a Learning Case in Practice: Approaches and Experiences. In Proceedings of International Workshop on Machine Learning Technologies in Software Engineering (MALETS 2011), Lawrence, Kansas, 2011.