“Operational Data are not Experimental Data”

Audris Mockus
Department of Electrical Engineering and Computer Science
University of Tennesse, Knoxville
audris@utk.edu

[2014-06-24]

Defining Features of Software Repository Data

Not experiment data

Defining Features of Software Repository Data

`Not experiment data`

Definition (Operational Data (OD))

Digital traces produced in the regular course of work or play (i.e., data generated or managed by operational support (OS) tools)

- no carefully designed measurement system

Why Study OD?

- Prevalent
  - Massive data from software development
  - Increasingly used in practice
- Treacherous - unlike experimental data
  - Multiple contexts
  - Missing events
  - Incorrect, filtered, or tampered with
- Continuously changing
  - OS systems and practices are evolving
  - New OS tools are being introduced in SE and beyond

Summary

- Defining features of OD
  - No two events have the same context
    - Observables represent a mix of platonic concepts
  - Not everything is observed
  - Data may be incorrect
- ML/Statistics/Databases/... assume experiment data
- How to engineer ODS?
  - Understand practices of using operational systems
  - Establish Data Laws
    - Use other sources, experiment, ...
  - Use Data Laws to
    - Recover the context, correct data, impute missing

Audris Mockus. Engineering big data solutions. In ICSE'14 FOSE, 2014

PostDoc Opening at UTK:

Who: An incurably curious person deeply interested in understanding the world through the observations recorded as data of every size or shape. Passion for hacking the data analysis to describe, understand, model, and present complex and dynamic interrelationships, and discover insidious data quality problems. An uncompromising striving to obtain reproducible and practically relevant results.
What: You will develop techniques to explore, understand, and model various phenomena based on very large operational data from software and related domains to shape the future of this rapidly evolving domain. You will collaborate with a multidisciplinary team of engineers, qualitative and quantitative scientists on a wide range of problems of practical significance. This position will bring analytical rigor and statistical methods to the challenges of understanding the accuracy, completeness, and relevance of data, and how it reflect people's behavior.
Requirements:

- PhD preferred in statistics, applied mathematics, operation research, computer science or related field;
- Substantial real-world experience, especially in areas of data analysis.
- Familiarity statistical software (R, S-Plus, or similar).
- Familiarity with machine learning and/or experimental design principles.
- Proficiency with databases and scripting or programming languages (particularly Python or Java).
- Ability to draw real-world conclusions from data and recommend actions.
- Demonstrated willingness to both teach others and learn new techniques.