Current Topics in Media Computing and HCI

Supporting Exploratory Workflows

Krishna Subramanian
Media Computing Group
RWTH Aachen University

Summer Semester 2018

http://hci.rwth-aachen.de/cthci
Data analysis, machine learning, 3D modeling, etc., have an exploratory workflow (i.e., highly iterative and non-linear [1, 7]).
Exploratory workflow

- Data analysis, machine learning, 3D modeling, etc., have an exploratory workflow (i.e., highly iterative and non-linear [1, 7]).

Load data → Compute mean, sd → Plot histogram

Histogram of dist
Exploratory workflow

- Data analysis, machine learning, 3D modeling, etc., have an exploratory workflow (i.e., highly iterative and non-linear [1, 7]).

Load data → Compute mean, sd → Plot histogram → Remove outlier

Compute mean, sd → Plot histogram

Histogram of data
Existing tools for exploratory processes

• Several approaches
• GUIs
• Visual programming
• Text-based programming

Weka (Machine Learning)

JMP (Data Analysis)

predictiveanalyticstoday.com

igcgs.org
Existing tools for exploratory processes

• Several approaches
  • GUIs
  • Visual programming
  • Text-based programming
Existing tools for exploratory processes

- Several approaches
- GUIs
- Visual programming
- Text-based programming
Text-based programming environments

- E.g., Matlab, Python, and R.
- Linear workflow for a non-linear process => mismatch!
- What are the implications of this mismatch?
Problems in data analysis [3]

- Literate programming tool: Jupyter
- Intersperse code with narratives

![Jupyter Notebook screenshot](dataquest.io)
• Semi-structured interviews with 21 data scientists

• **Grounded Theory Methodology (GTM)**
  
  • A theory of the process is formed over observations (interviews, questionnaires, …)
  
  • Theories are formed over time with the researcher trying to *disprove* her theories with each participant.
  
  • Study ends when the analyst hits the *saturation point* i.e., new participants do not add to the theories.
Findings: Use cases

• Jupyter notebooks were used in the following different ways

  • **Scratchpads**: debug a piece of code, test an example, etc; sometimes at the beginning, sometimes at the end of the notebook; some code cells were moved to the production pipeline.

  • **Production pipeline**: code is ‘formalized’ by moving it into a larger file.

  • **Sharing with others/self**: analysts put in extra effort to make the notebooks readable.
Findings: Iteration behavior

- **Organization**: analysts either kept similar code together or wrote code in chronological order.
- **Clean up**: analysts cleaned up code after one logical step.
- **Copy-and-paste programming**: analysts constantly copy-pasted code and executed them with different arguments.
We collected 40 R scripts from OSF (Open Science Framework) and researchers at the local university.

• ~70% lines of code were clones.

• Barely any functions were used.

• We observed analysts \((n=5)\) perform data analysis using RStudio or Jupyter.
  • Over-reliance on interactive consoles.
  • Issues with code navigation.
Solutions
Variolite: Supporting Exploratory Programming by Data Scientists

Mary Beth Kery, Amber Horvath, Brad Myers
Carnegie Mellon University 2016

Paper: [2]
Video: https://www.youtube.com/watch?v=rBrvT7vGleM
Design as Exploration:  
Creating Interface Alternatives through Parallel Authoring and Runtime Tuning

Björn Hartmann, Loren Yu, Abel Allison, Yeonsoo Yang, Scott R. Klemmer  
Stanford HCI Group

Paper: [10]  
Video: https://youtu.be/X0nSaFG-mKI
• View diffs of code changes and corresponding output

• Allows users to link code to output

• General idea: ‘Speculative reprogramming’ [6]
Other relevant work

- Implications of exploratory workflows in statistical machine learning: [7].
- Interface for comparing multiple alternative documents: [8].
- Supporting simultaneous development of alternative solutions to ill-defined problems: [9].

Study #1.
References 1 of 2


References 2 of 2


