



Understanding Your Software Development

(through git repository data mining and code analysis)

Boyana R. Norris (she/her)
University of Oregon

Software Productivity and Sustainability track @ Argonne Training Program on Extreme-Scale Computing summer school

Contributors: Stephen Fickas (UO), Bosco Ndemeye (UO), Boyana R. Norris (UO), Jason Prideaux (UO)

<https://github.com/CAT-SDK/GremCat>



See slide 2 for
license details

License, Citation and Acknowledgements

License and Citation

- This work is licensed under a [Creative Commons Attribution 4.0 International License](#) (CC BY 4.0). 
- **The requested citation the overall tutorial is: David E. Bernholdt, Anshu Dubey, Todd Gamblin, Jared O'Neal, and Boyana R. Norris, Software Productivity and Sustainability track, in Argonne Training Program on Extreme-Scale Computing, St. Charles, Illinois, 2022. DOI: [10.6084/m9.figshare.20416215](https://doi.org/10.6084/m9.figshare.20416215).**
- Individual modules may be cited as *Speaker, Module Title*, in Better Scientific Software tutorial, ISC, 2022 ...

Acknowledgements

- This work was supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research (ASCR), and by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.
- This work was performed in part at the Argonne National Laboratory, which is managed by UChicago Argonne, LLC for the U.S. Department of Energy under Contract No. DE-AC02-06CH11357.
- This work was performed in part at the Lawrence Livermore National Laboratory, which is managed by Lawrence Livermore National Security, LLC for the U.S. Department of Energy under Contract No. DE-AC52-07NA27344.
- This work was performed in part at the Los Alamos National Laboratory, which is managed by Triad National Security, LLC for the U.S. Department of Energy under Contract No.89233218CNA000001
- This work was performed in part at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.
- This work was performed in part at Sandia National Laboratories. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
- This work was performed in part at University of Oregon through a subcontract with Argonne National Laboratory.

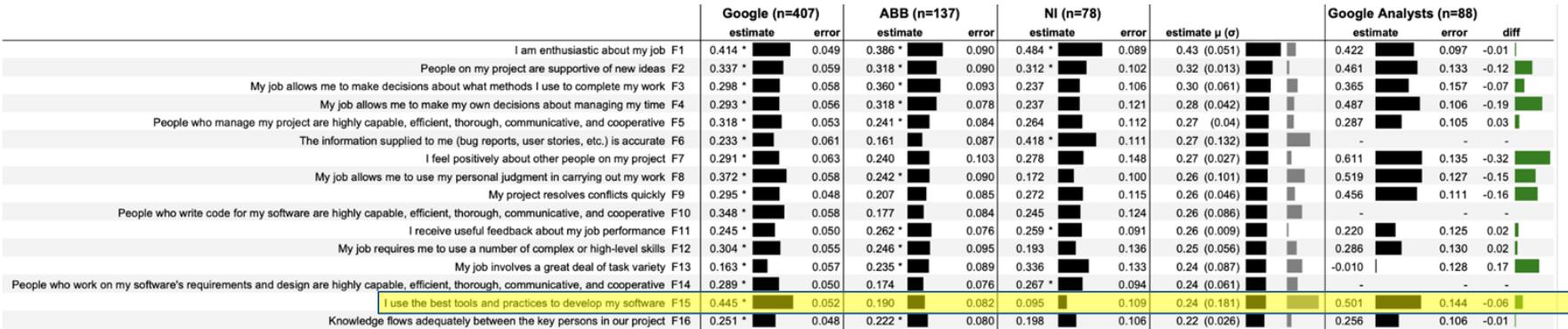
Motivation

- Software development productivity is affected by many factors, many of which are difficult or impossible to measure (accurately).
- So far in this track, the focus was on concepts and best practices based on research, experience and expert opinion.
- Here, we focus on *supporting* best practices through software **tools** that can help understand how existing or new software practices may affect these observables.
- Tools are meant to be used -- so we will show some simple ways in which *anyone* can begin exploring their own project's data without significant up-front investment of time and effort.

Motivation

(Am I just making things up?)

Murphy-Hill, E., Ciera Jaspan, Caitlin Sadowski, D. Shepherd, Michael Phillips, C. Winter, Andrea Knight, Edward K. Smith and M. Jorde. "What Predicts Software Developers' Productivity?" IEEE Transactions on Software Engineering 47 (2021): 582-594.



I use the best tools and practices to develop my software F15

0.445 * 0.052

Top factor at Google!



Fig. 4: 48 factors' correlation with developers' and analysts' self-rated productivity at three companies.



Top three (on average):

- Job enthusiasm (F1)
- Peer support for new ideas (F2)
- Useful feedback about job performance (F11)

Other software-related factors

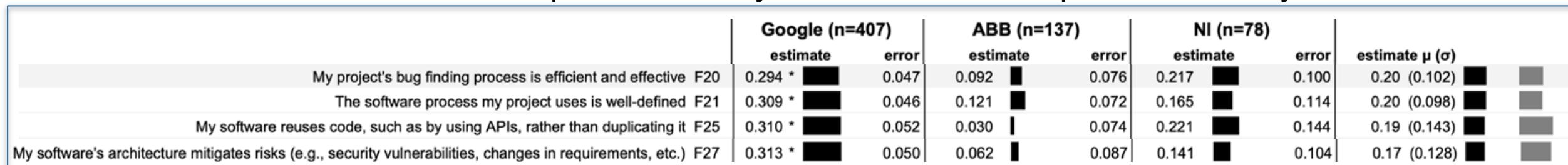
Murphy-Hill, E., Ciera Jaspan, Caitlin Sadowski, D. Shepherd, Michael Phillips, C. Winter, Andrea Knight, Edward K. Smith and M. Jorde. "What Predicts Software Developers' Productivity?" IEEE Transactions on Software Engineering 47 (2021): 582-594.

A few other positively rated software-related factors (in decreasing order of average scores):

- + My project's bug finding process is efficient and effective.
- + The software process my project uses is well defined.
- + My software reuses code, such as by using APIs, rather than duplicating it.
- + My software's architecture mitigates risks (e.g., security vulnerability, changes in requirements, etc.).

Negative impact on productivity:

- My software requires extensive data storage.
- My software is extremely complex.
- My software's platform (e.g. development environment, software stack, hardware stack) changes rapidly.
- Extensive documentation is required to use my software at different points in its lifecycle.



Murphy-Hill et al. study conclusions

“A notable outcome of the ranking is that the top 10 productivity factors are non-technical. This is somewhat surprising, given that most software engineering research tends to focus on technical aspects of software engineering, in our estimation.

Thus, a vigorous refocusing on human factors may yield substantially more industry impact for the software engineering research community. For instance, answering the following questions may be especially fruitful:

- What makes software developers **enthusiastic** about their job? What accounts for differences in levels of enthusiasm between developers? What interventions can increase enthusiasm? This work can extend existing work on developer happiness [24] and motivation [25].
- What kinds of new ideas are commonly expressed in software development practice? What actions influence developers’ feelings of support for those ideas? What interventions can increase **support for new ideas**, while maintaining current **commitments**?
- What kinds of job feedback do software engineers receive, and what makes it **useful**? What kinds of feedback is not useful? What **interventions** can increase the regularity and usefulness of feedback?” [emphasizes mine]

What we are trying to do

Motivation and enthusiasm are ephemeral things that cannot be quantified easily.

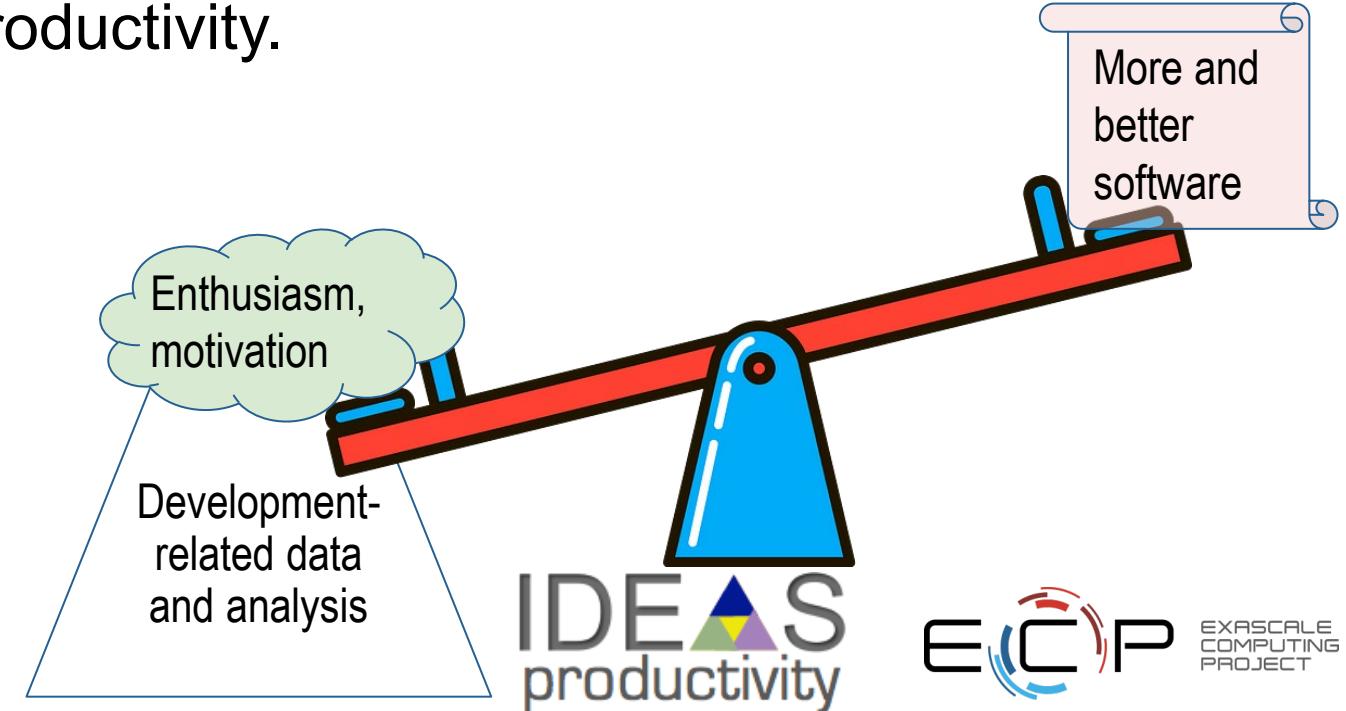
So, we are betting that using **some data** is better than not looking at any data, as we make decisions that are aimed at improving software development productivity and code quality.

This work is part of two DOE **ECP** projects:



IDEAS-ECP: foster and promote SE practices for better software
xSDK: Extreme-scale Scientific Software Development Kit

On the other hand, we have many artifacts and associated metadata that are related (if indirectly) to productivity.



What this module is about

- We introduce a **flexible**,
efficient, and
usable software framework for **acquiring**,
storing,
manipulating, and
visualizing development-related data.
- We demonstrate a few of its capabilities here; we continue to add new patterns and tools.
 - Contributions welcome! github.com/CAT-SDK/GremCat/

ECP projects that may be present in examples in this module: Spack, LAMMPS, PETSc, Nek5000 E3SM, QMCPACK, QDPXX, LATTE NAMD, fast-export, Enzo, TAU2, xpress-apex, LATTE, NWChem, FLASH, Gingko.

Part I: Mining development metadata

Part II: Analyzing code

Part I: Mining your development data

Development data:

- Git metadata: commits, forks, branches, developers
- Issues and associated discussions
- Pull requests (github, gitlab) and associated discussions
- Mailing list archives

Goal: Analyze available data to help *formulate* and *answer* questions about development processes and their impact on productivity and code quality.

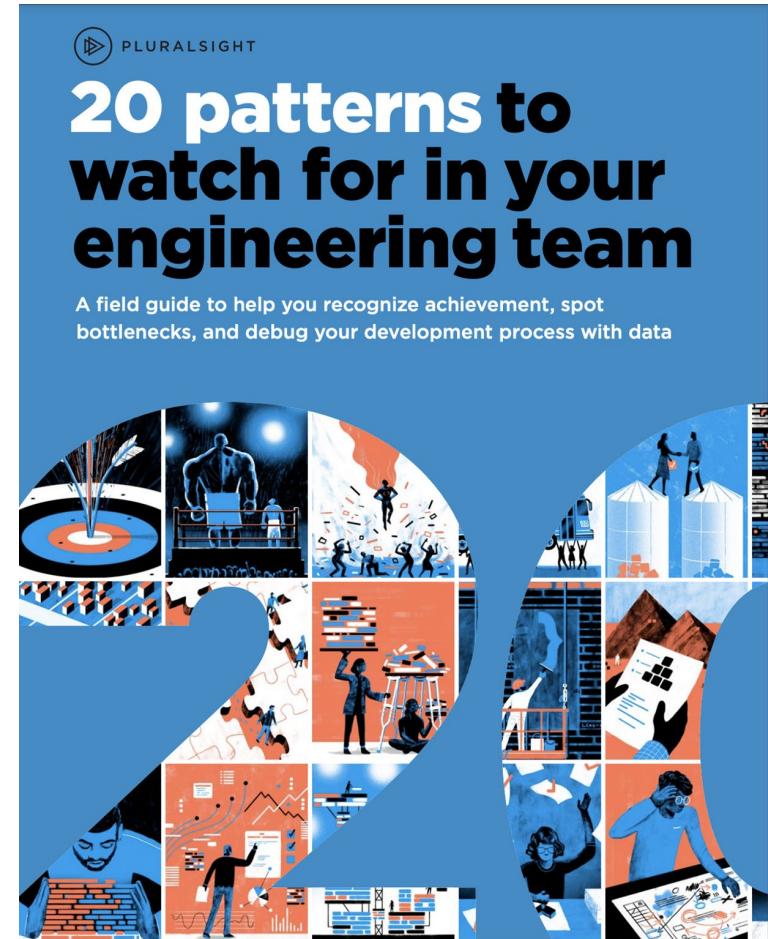
Questions that can be answered (in part) with IDEAS data analysis tools

- If I adopt practice **X**, how will metric **Y** be affected?
- How active is the developer community? (git, issues, PRs, emails)
- What parts of the code base could benefit from review or refactoring? (git, issues)
- What is the project's reliance on individual developers? (git)
- How are developers' contributions split among different categories? (git, manual labeling required)
- How engaged are the user and developer communities? (PRs, issues, mailing lists)
- What are some hot topics of discussion? (issues, mailing lists)
- How and on what do developers collaborate? (git, issues, mailing lists)
-

Common patterns with known implications

Pluralsight book (2019)¹:

“20 patterns is a collection of work patterns we’ve observed in working with hundreds of software teams. Our hope is that you’ll use this field guide to get a better feel for how the team works, and to recognize achievement, spot bottlenecks, and debug your development process with data.”

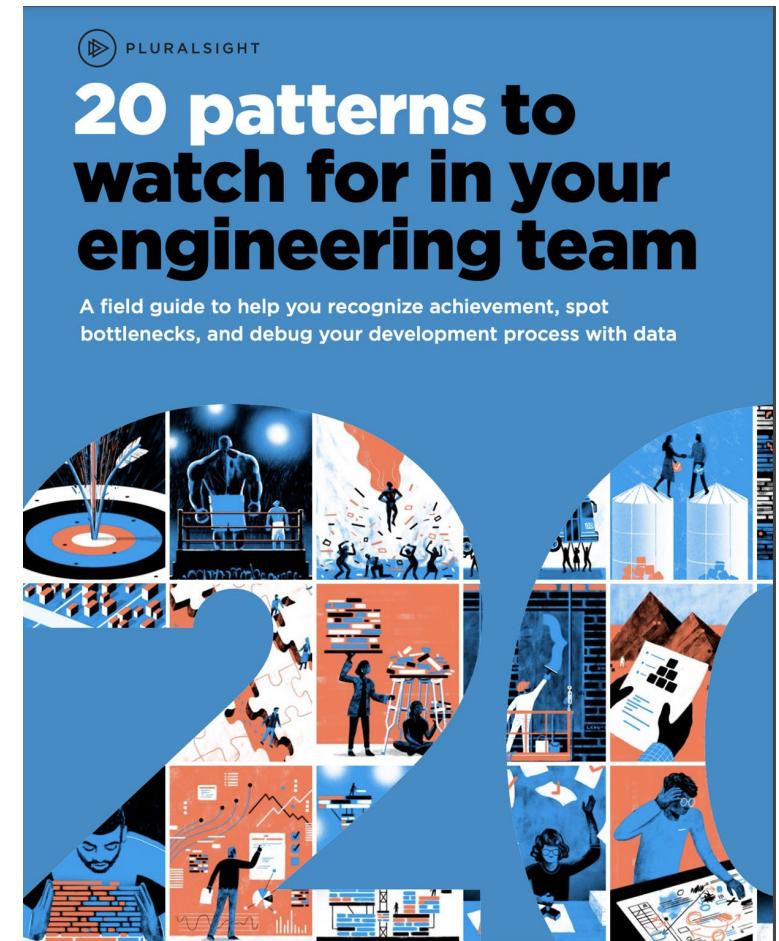


¹https://www.pluralsight.com/content/dam/pluralsight2/landing-pages/offers/flow/pdf/Pluralsight_20Patterns_ebook.pdf

Common patterns with known implications

Starting with these pattern descriptions, we can:

- Identify patterns that are relevant to HPC (open-source) software development.
- Characterize each pattern using data from revision control systems and developer communications.
- Inform decisions of the effects of adopting new SE practices.



¹https://www.pluralsight.com/content/dam/pluralsight2/landing-pages/offers/flow/pdf/Pluralsight_20Patterns_ebook.pdf

Example metrics

Metric	Description
Monthly bug fix rate	Computes cumulative count of bugs closed each month, starting from the beginning of the project. A higher value is not necessarily good, trends are more informative than individual values.
Monthly feature request rate	Number of new feature requests made by users. A higher value may indicate popularity/importance of the feature or missing functionality in the current software version.
Correlation of the number of issues with project age	Gives a good insight into the life cycle of the project by mapping the trend of issues raised over the lifetime of the project.
Commits, derived metrics	Characterizes some aspects of developer activity.
Number of issues	Project-related communications can be used to indicate community involvement and rates at which issues are resolved. For example, fast-growing projects can have significantly more activity in issues than more mature ones.
Issue categories	Identifies the types of issues that are reported in the repository. This information will be useful to derive correlation with other metrics.
Number of followers and watchers	Reports the code maturity and popularity. In addition, the number of followers can be correlated with the time it takes to fix reported issues.
Mailing list metrics	Relates to average time in discussion, most popular topic of interest, product activity based on type of emails, etc.
Number of contributors	Size of development community. When analyzed over time, we can estimate project turnover and identify different types of contributors.
Code complexity	Compute changes to code complexity metrics, e.g., cumulative size or cyclomatic complexity, over time.

Example pattern: Domain champion

The Domain Champion is an expert in a particular area of the codebase. They know nearly everything there is to know about their domain: every class, every method, every algorithm and pattern.

In truth, they probably wrote most of it, and in some cases rewrote the same sections of code multiple times.

The Domain Champion isn't just "the engineer who knows credit card processing"; it's all they ever work on. It's their whole day, every day.

Some degree of job specialization is essential and often motivating. But even within specialized roles there can be 'too much of one thing.' Managers must balance enabling a team member to unilaterally own the expertise, and encouraging breadth of experience.

How to recognize it

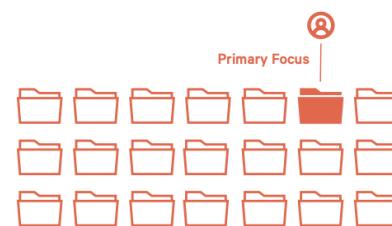
Domain Champions will always work in the same area of code. They'll also rewrite their code over and over, and you'll see it in churn and legacy refactoring metrics as they perfect it.

Domain Champions are deeply familiar with one particular domain. As a result, they'll typically submit their work in small, frequent commits and will show a sustained above average *Impact*.

Because no one else knows more than the Domain Champion, there's usually very little actionable feedback that

others can provide in the review process. As a result, Domain Champions will typically show low *Receptiveness* in incorporating feedback from reviews.

Domain Champions will seldom, if ever, appear blocked. Short-term, it's a highly productive pattern. But it's often not sustainable and can lead to stagnation, which of course can lead to attrition.



What to do

Assign tickets that focus on other areas of the codebase.

Of course, some engineers would prefer to stay where they are. It can be very enjoyable to do a task you're good at. And, it can be uncomfortable to take on work that requires information or skills you have less practice with. But effective managers will strive to challenge their team.

20 patterns to
watch for in your
engineering team

"The Domain Champion is an expert in a particular area of the codebase. They know nearly everything there is to know about their domain: every class, every method, every algorithm and pattern."

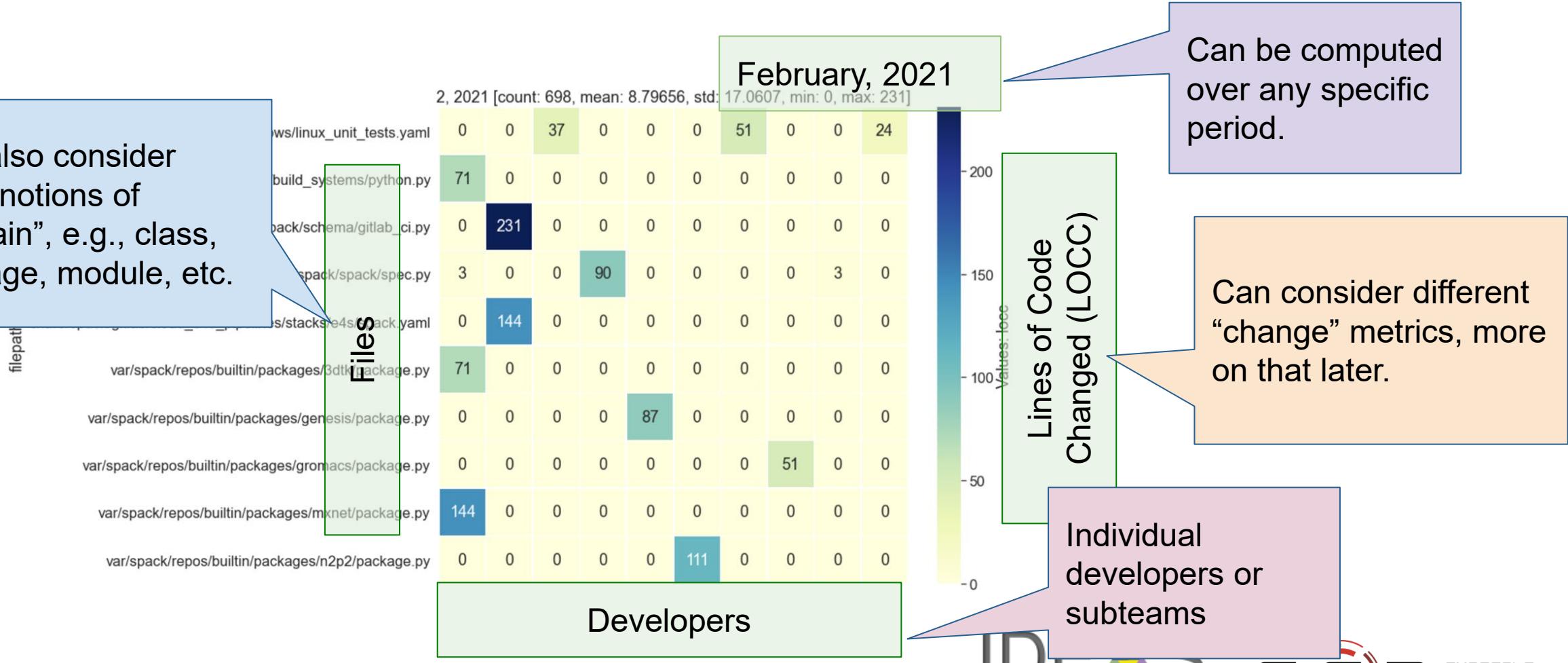
Why do we care?

- Can lead to great productivity
- Code quality implications
- Turnover effects on code -- what happens to the domain when the domain champion leaves?

Domain champion pattern: How do we detect it?

“The Domain Champion is an expert in a particular area of the codebase.”

Can also consider other notions of “domain”, e.g., class, package, module, etc.



Domain Champion Pattern: What, if anything, should we do?

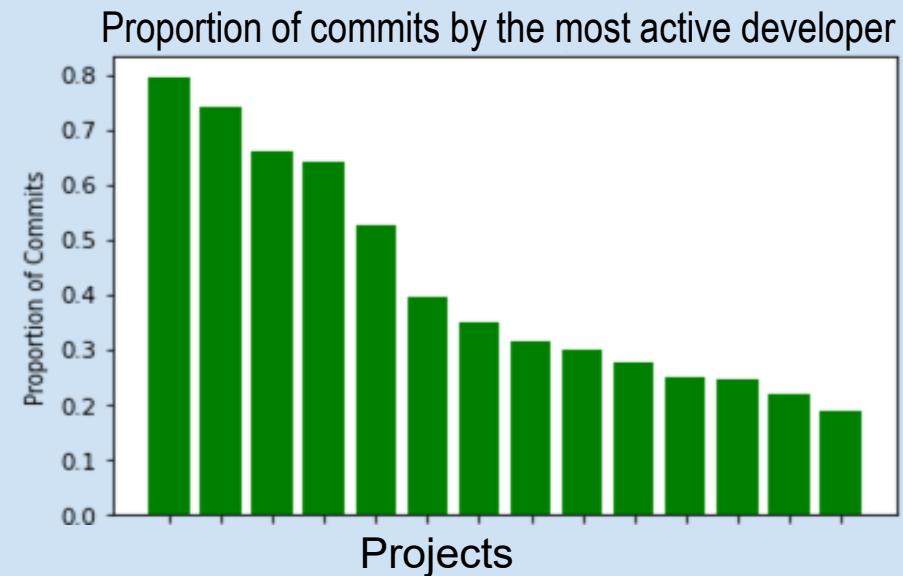
“The Domain Champion is an expert in a particular area of the codebase.”

- + Highly productive pattern in the short term.
- There's usually very little actionable feedback that others can provide in code reviews.
- Potentially not sustainable → can lead to stagnation.

😊 Possible actions:

- Assign the DC tickets for other areas of the code.
- Make an effort to involve others (e.g., new developers) in work in that domain.

One possible extension: Project Champion



Another example pattern: Unusually high churn

Churn is a natural and healthy part of the development process and varies from project to project. However, Unusually High Churn is often an early indicator that a team or a person may be struggling with an assignment.

In benchmarking the code contribution patterns of over 85,000 software engineers, Pluralsight's data science team identified that Code Churn levels frequently run between 13-30% of all code committed (i.e., 70-87% Efficiency), while a typical team can expect to operate in the neighborhood of 25% Code Churn (75% Efficiency).

Testing, reworking, and exploring various solutions is expected, and these levels will vary between people, types of projects, and stage in the software lifecycle. Given the variance, becoming familiar with your team's 'normal' levels is necessary to identify when something is off.

Unusually high churn levels aren't a problem in themselves. More likely, there are outside factors causing the problem.

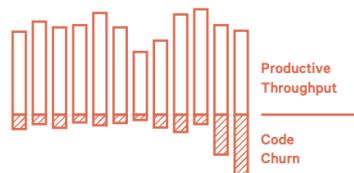
An unusually high level of churn can be indicative of one of three behaviors:

- **Perfectionism:** When an engineers' standards of "good enough" are not aligned with the company's standard of "good enough." Engineers keep going back into the code to rewrite it because they think it can and should be better but may not add much to the actual functionality of the code.

- **They're struggling with the problem at hand.** This situation manifests differently than with Hoarding the Code (pattern #2), because in this case, the engineer initially thought they had correctly solved the problem, perhaps even sent it off for review, and then discovered it needed to be rewritten. Not just touched up. Rewritten.
- Or, most commonly, **issues concerning external stakeholders.** We see this with unclear or ambiguous specs, late arriving requirements, or mid-sprint updates to the deliverables.

How to recognize it

This pattern is characterized by **high levels of churn in the back of the sprint** or project. Watch for churn rates that climb significantly above the engineer's historical average (see the *Snapshot* and *Spot Check* reports), pairing that information with where they are in a project.



20 patterns to
watch for in your
engineering team

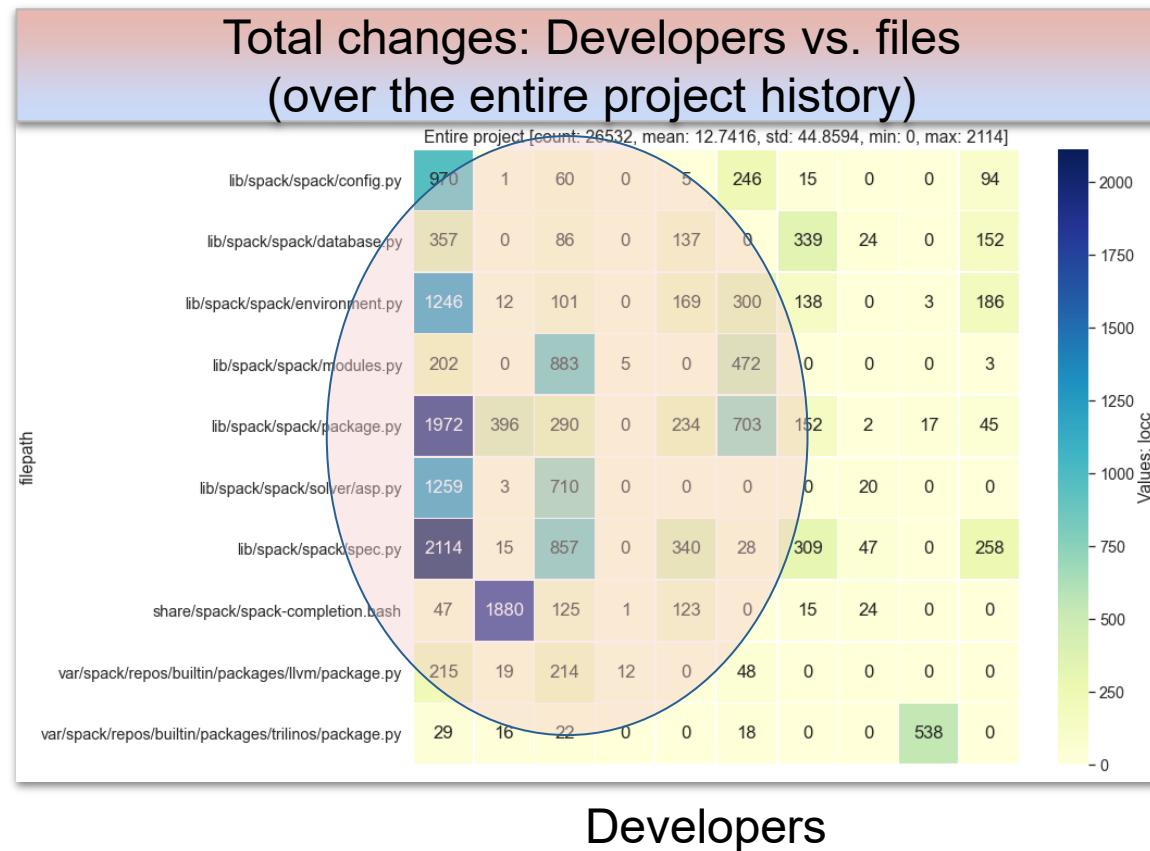
(Code churn: code that is rewritten or deleted shortly after being written)

What files (components) have unusually large or frequent changes? How many developers are involved?

- + could be a sign of normal productive development
- may indicate need for more developer resources
- high conflict potential

Unusually high churn pattern: How do we detect it?

- Decide on the *granularity*. Some possibilities: modules, classes, files, functions.
 - Define the *actors* -- groups of people (e.g., sub-teams), individual developers; doesn't have to be people, it can also be milestones or other project entities.
 - Choose the *time period*.
 - Choose the *churn* metric. Some examples: lines of code, cos (and other) difference between code versions, number of PRs, commits, number of files.

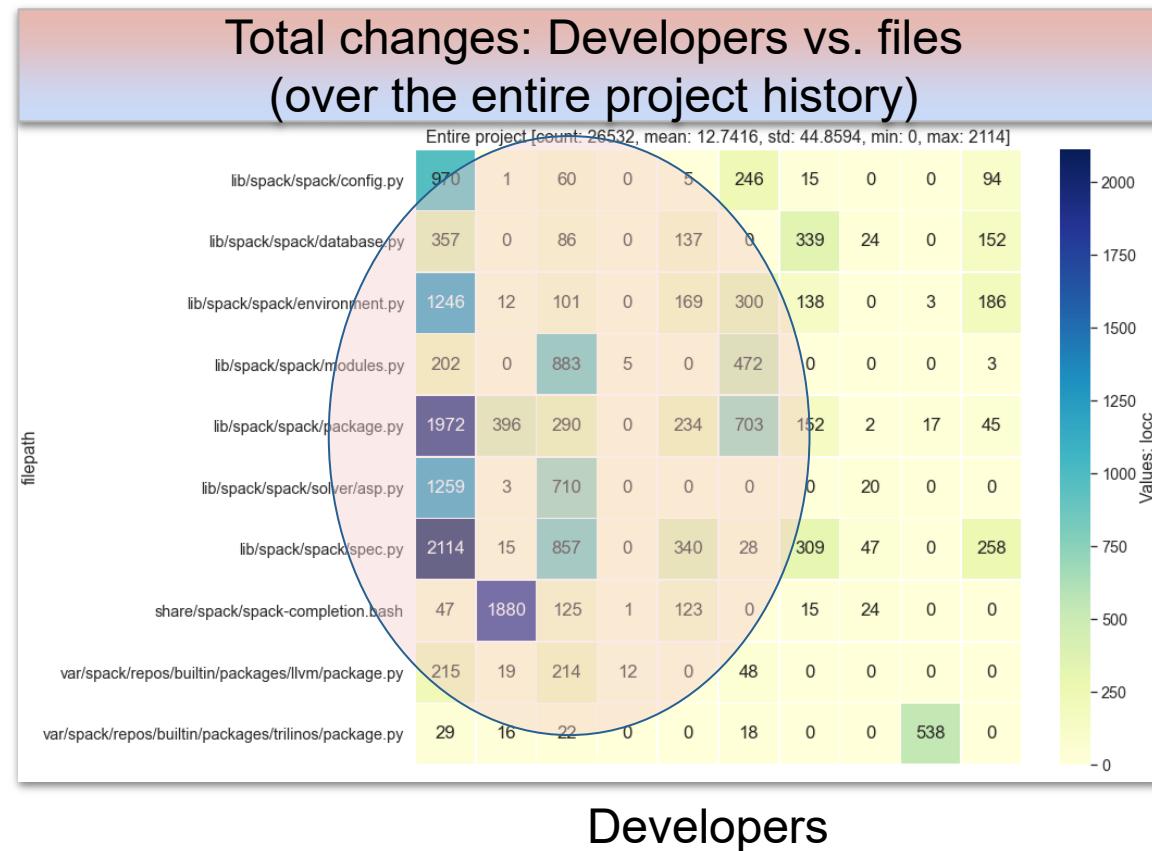


Unusually high churn pattern: What to do?

Unusually large amount of changes to single files or components may lead to development inefficiencies.

😊 Possible actions:

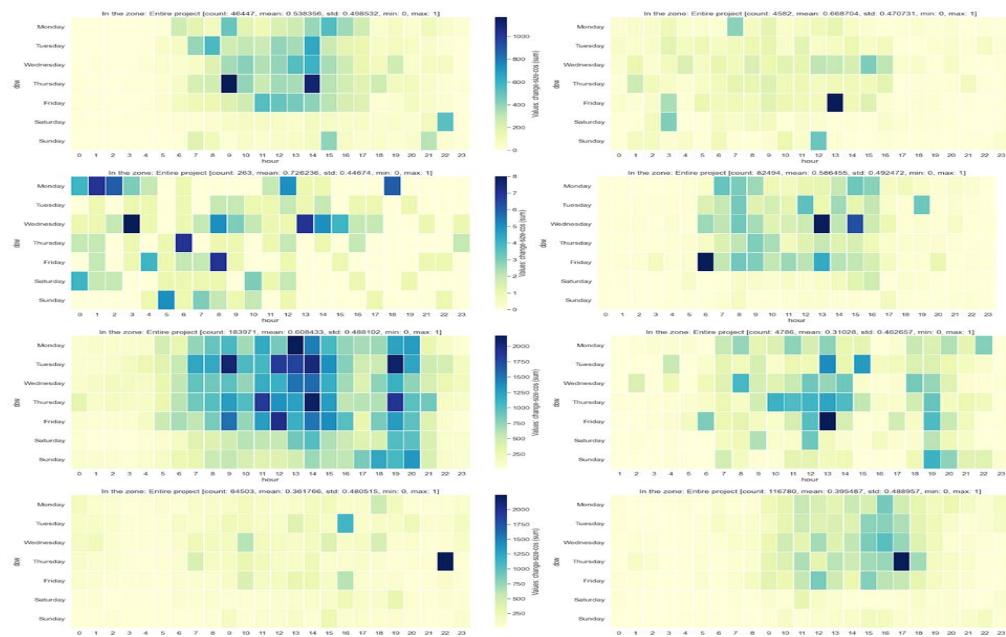
- Consider refactoring the high-churn project components
 - Consider involving more developers in high-churn areas that are dominated by a single person



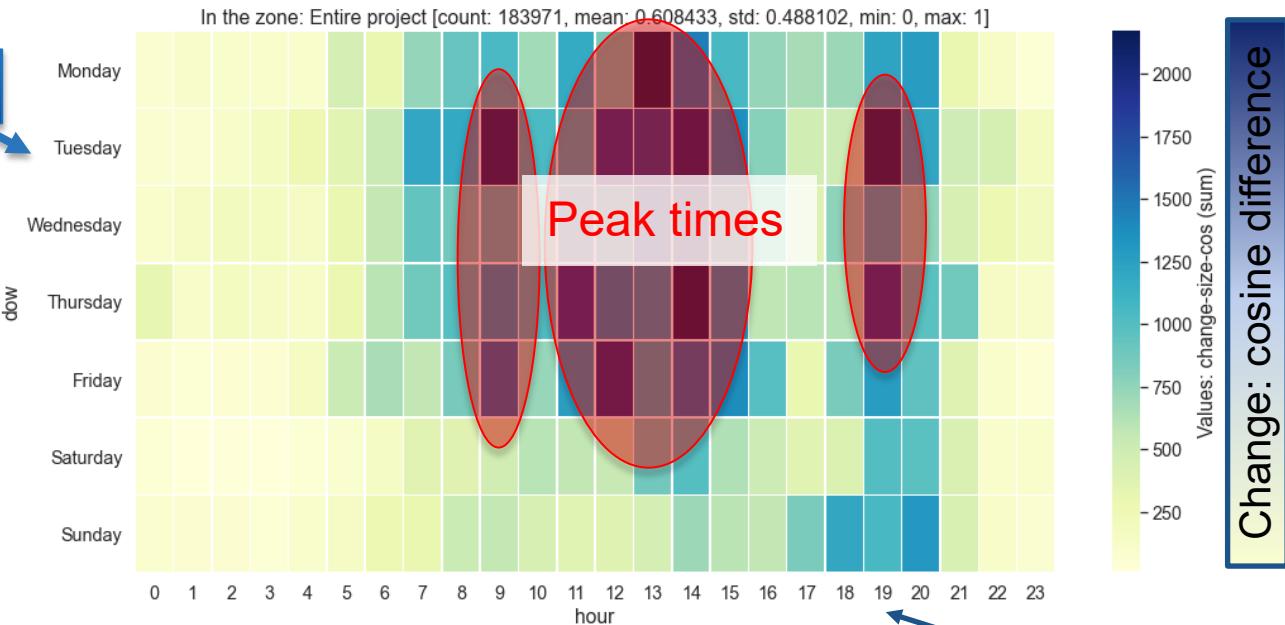
Pattern: In the zone

When are top contributors most productive?

- + Consistent high productivity during certain times of day
- Burnout, work-life balance



Weekday



Hour

How to compute?

1. Choose **productivity** metric (LOCC, text difference, # commits, # PRs, etc.)
2. Choose **time period**
3. Choose visualization

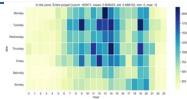
In the zone pattern: What to do?

😊 Possible actions:

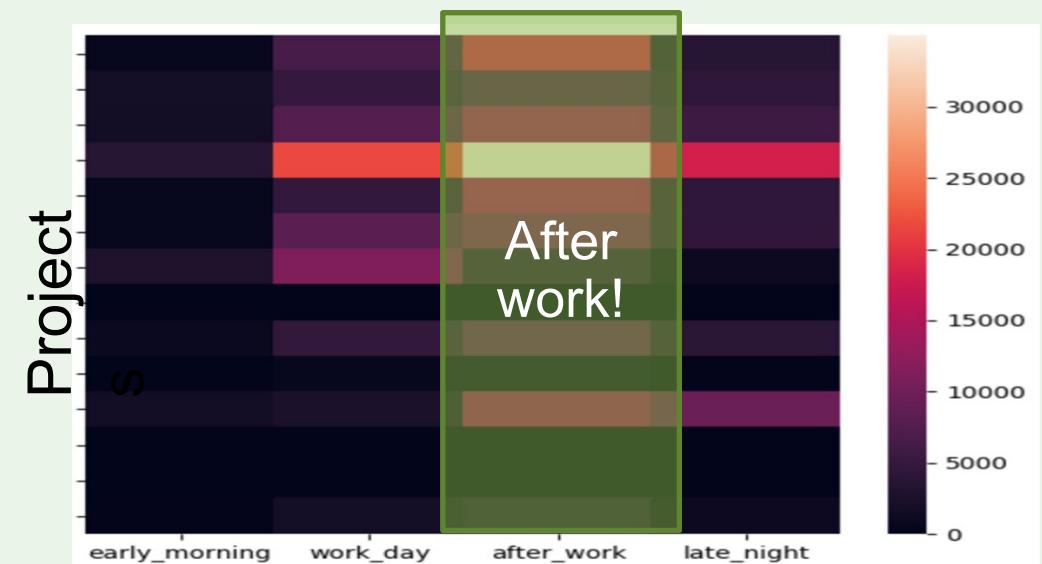
- Acknowledge the consistently high-performing developers.
- Recognize and acknowledge positive change in non-top developers (e.g., junior or new contributors).
- Consider the timing of and number of meetings during developers' most productive times.

When are top contributors most productive?

- + Consistent high productivity during certain times of day
- Burnout, work-life balance



Bonus: Ability to quickly “see” what people are doing on average. Below: The average of over a dozen ECP projects’ most active time periods.

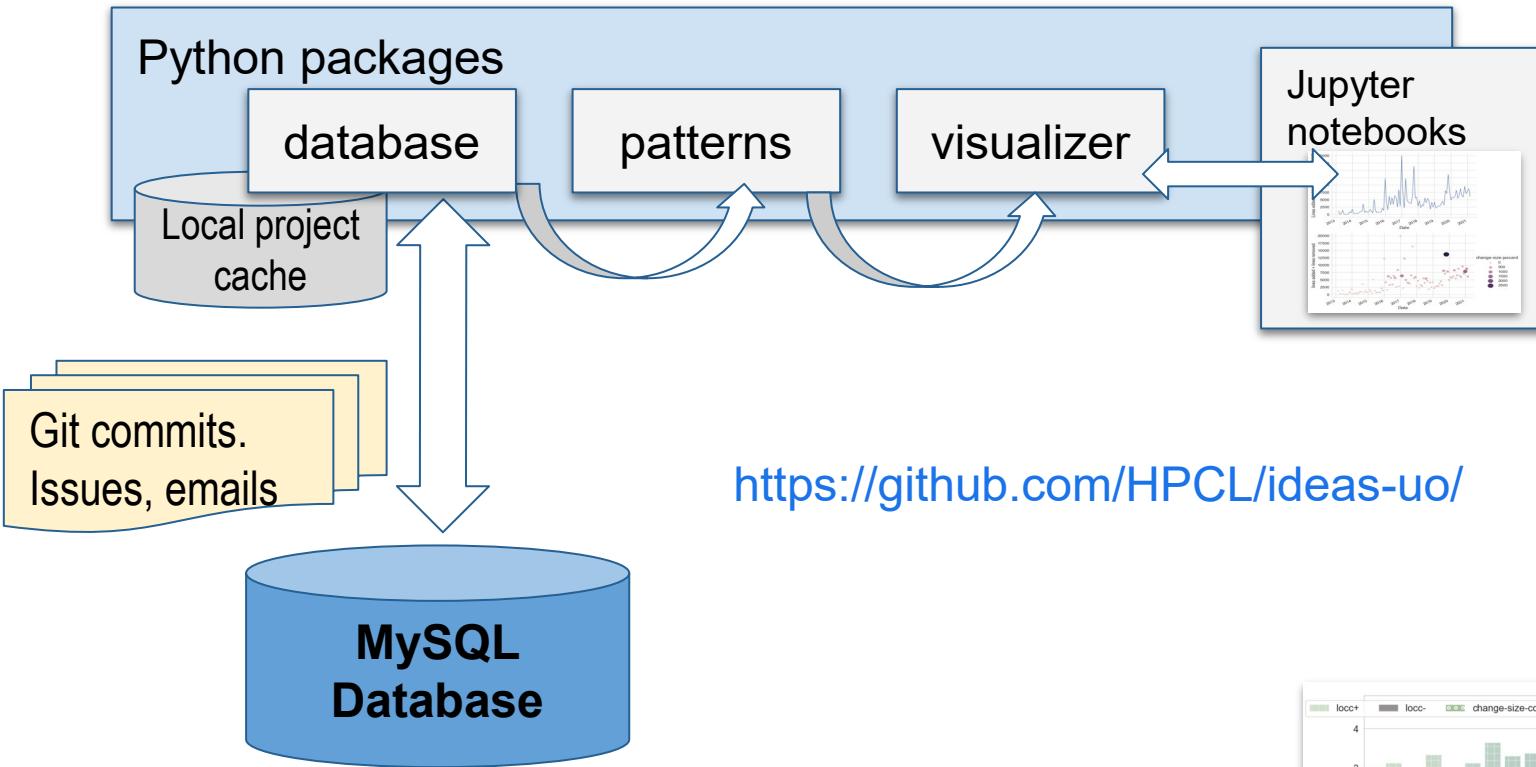


Technical Details and Examples

<https://bit.ly/DevPatterns>



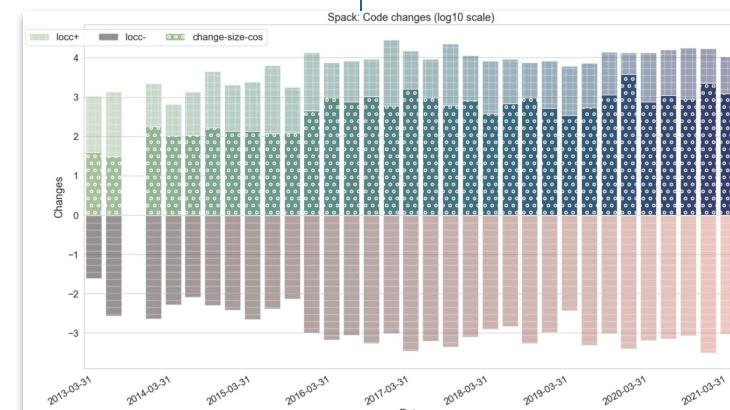
Implementation



Some currently available projects (more are being added constantly): [LAMMPS](#), [Spack](#), [PETSc](#), [Nek5000](#), [NWChem](#), [E3SM](#), [QPMCPACK](#), [qdpxx](#), [NWChem](#), [TAU](#),... Initial import can take up to 36 hours for some larger projects, but subsequent updates are fast and automated.

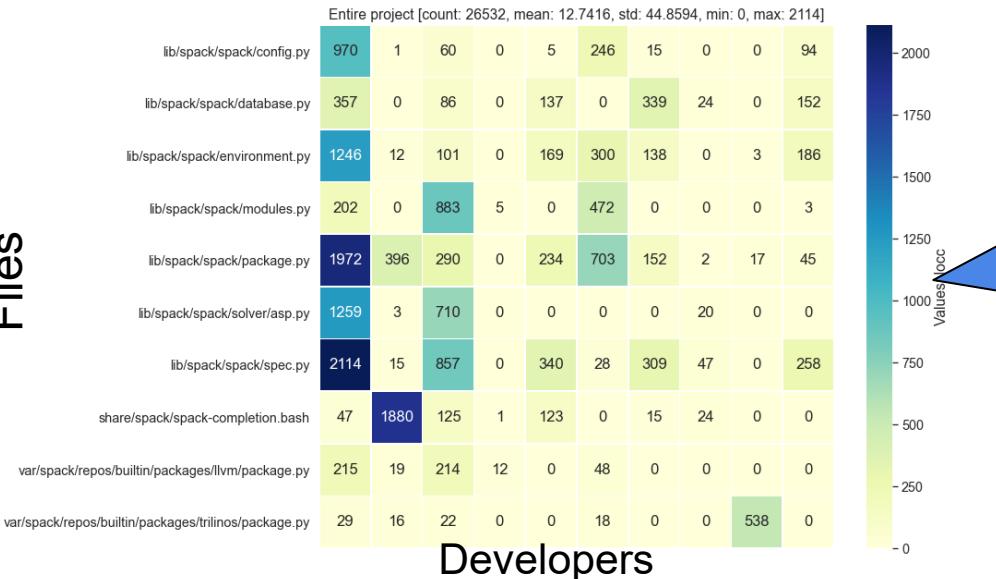
Example analysis workflow

```
from patterns.visualizer import Visualizer  
  
vis =  
Visualizer(project_name='spack')  
  
vis.get_data()  
INFO: Loaded local cached copy of  
spack data.  
INFO: Done computing averages.  
64909 commits (code only)  
  
df =  
vis.plot_overall_project_locc(time_  
range=None, log=True)
```



Impact of different “change” estimates

Files



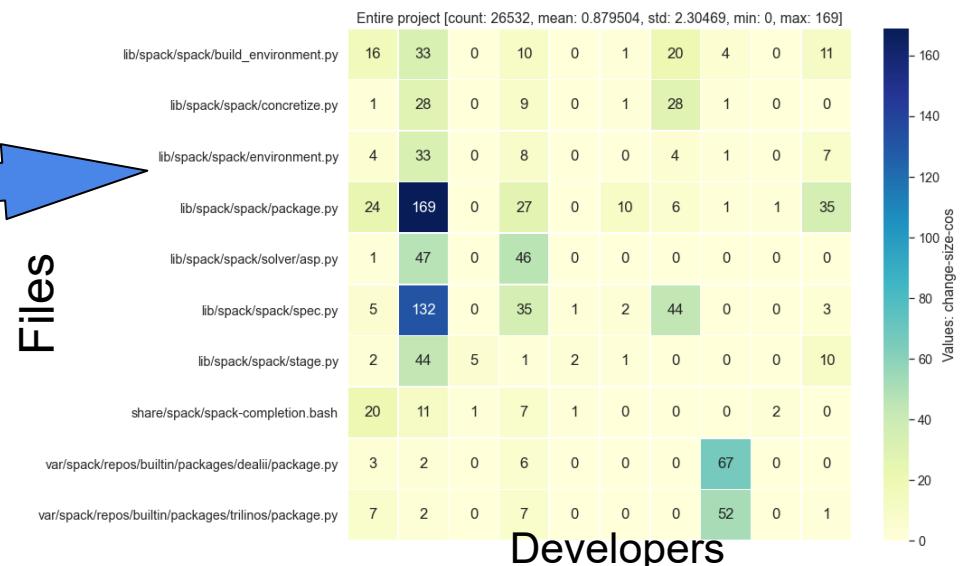
Simple line counts in git commit diffs:

- Patterns such as ‘---+++’ are counted as *edited* lines, e.g., 3 in this example
- Unmatched ‘-’ and ‘+’ lines counted as lines *deleted* and *added*, respectively
- LOCC = edited + deleted + added

Files

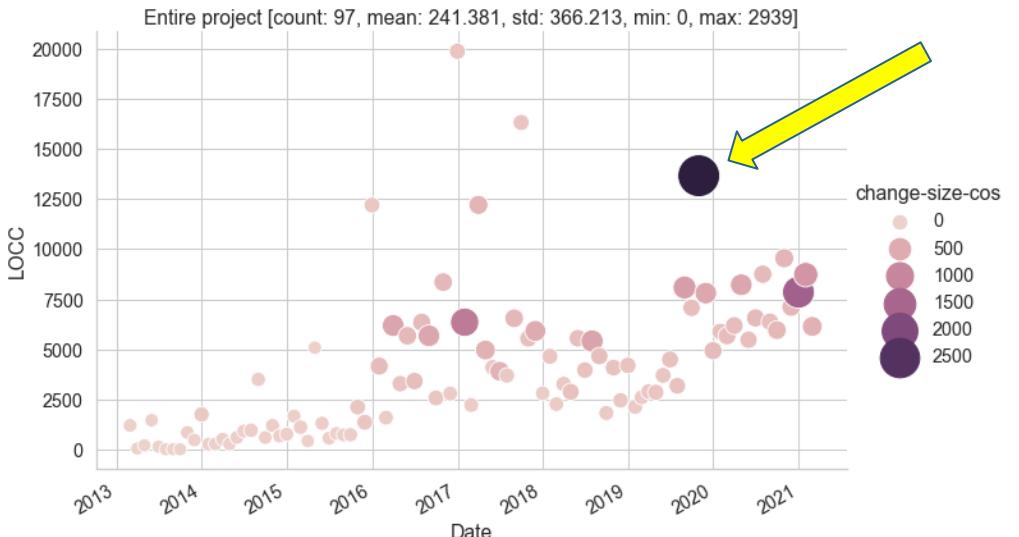
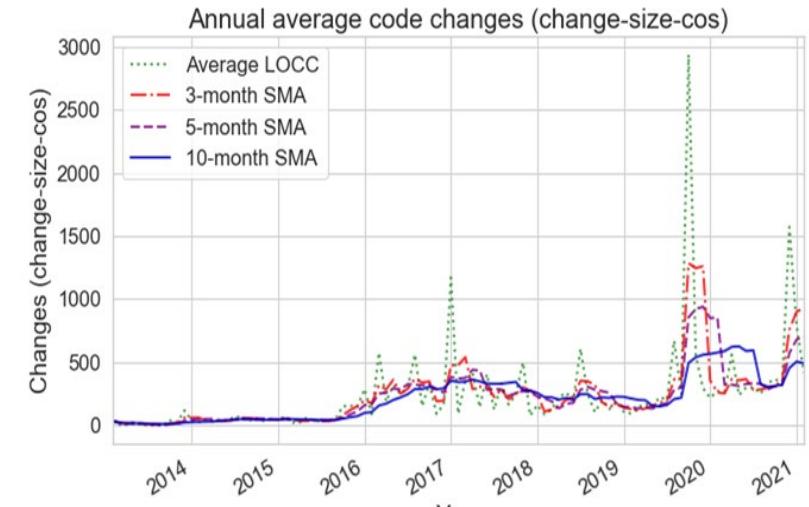
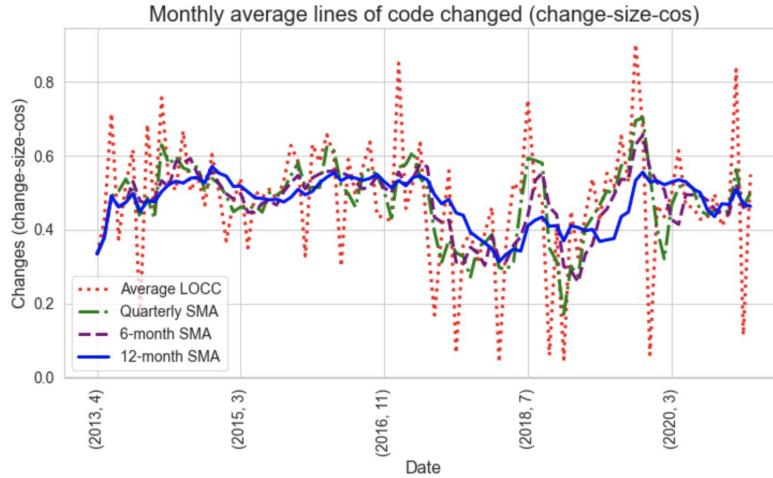
More “intelligent” estimate of the magnitude of changes:

- Collect the *old* and *new* strings corresponding to each commit’s diffs
- Apply text distance metrics (based on the `textdistance` Python package; ~30 methods)
- E.g., change-size-cos is the cos distance between the vector embeddings of *old* and *new*



Impact of different “change” estimates (cont.)

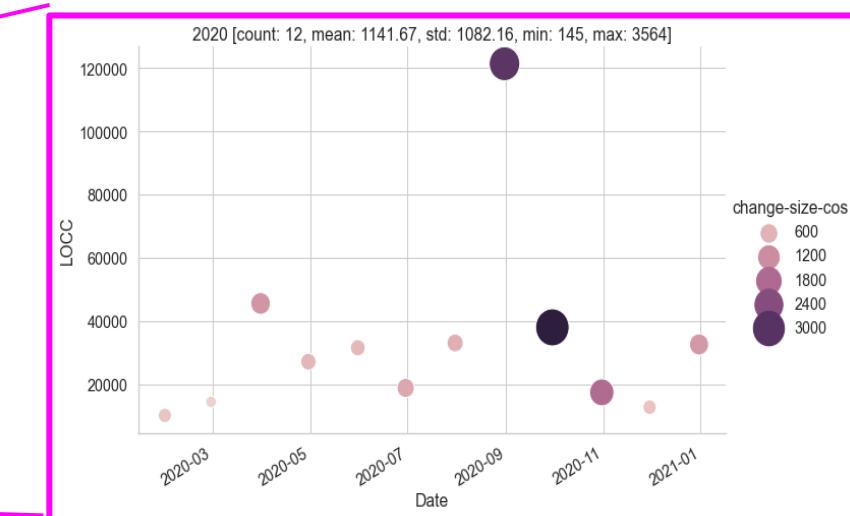
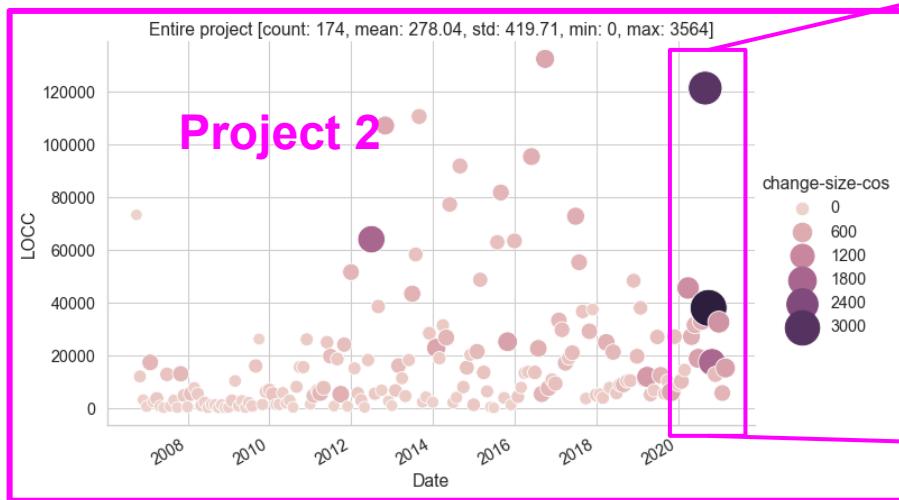
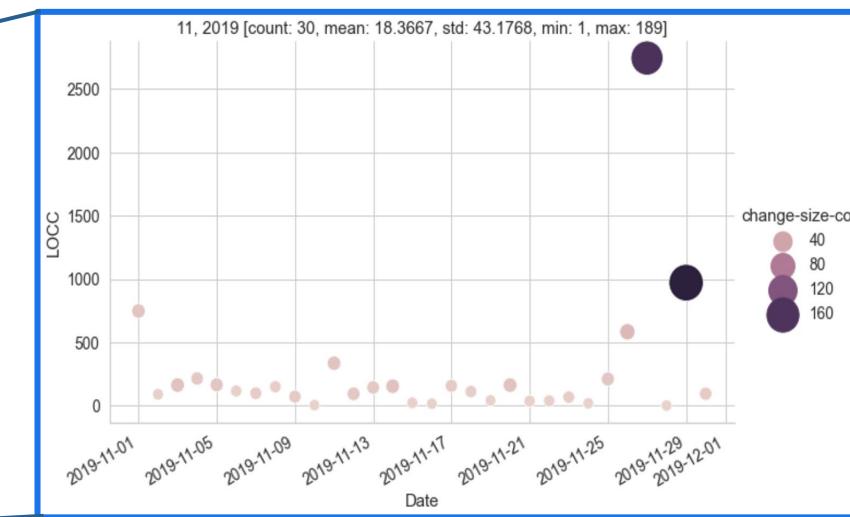
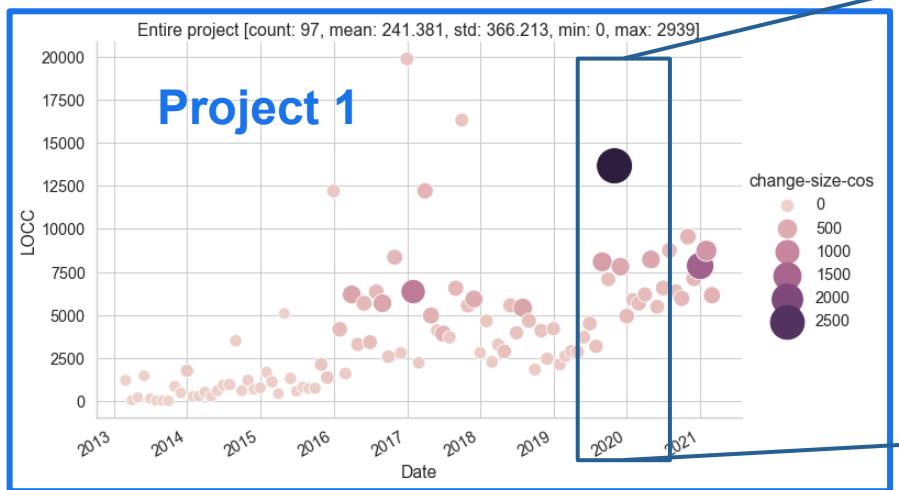
- Time-series git data is messy
 - Moving averages help see trends for any time period



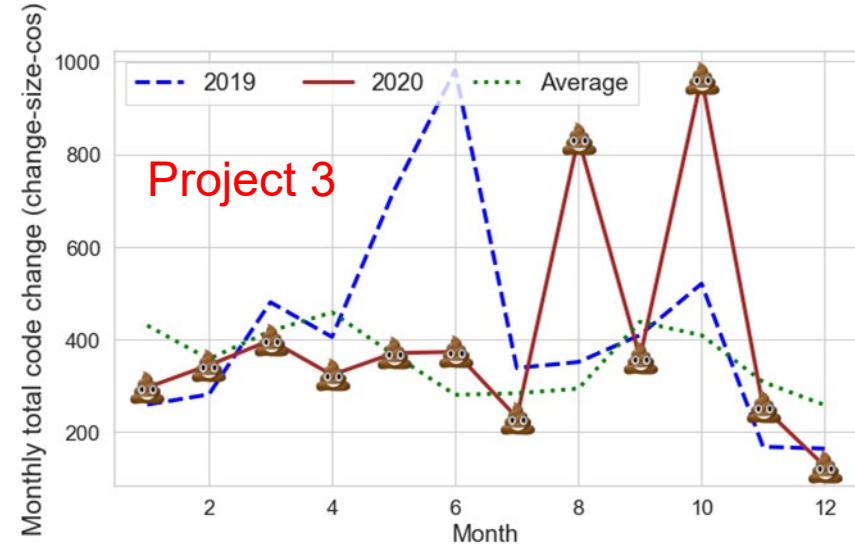
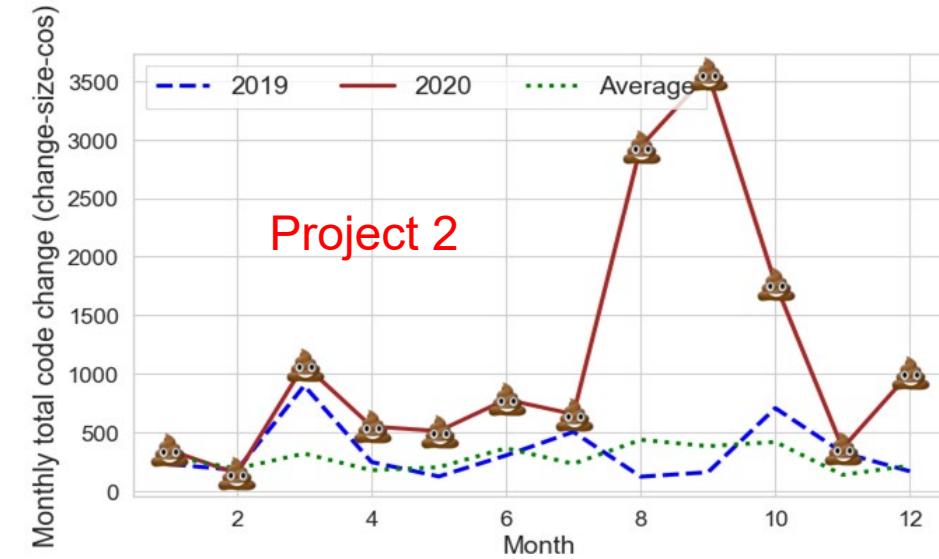
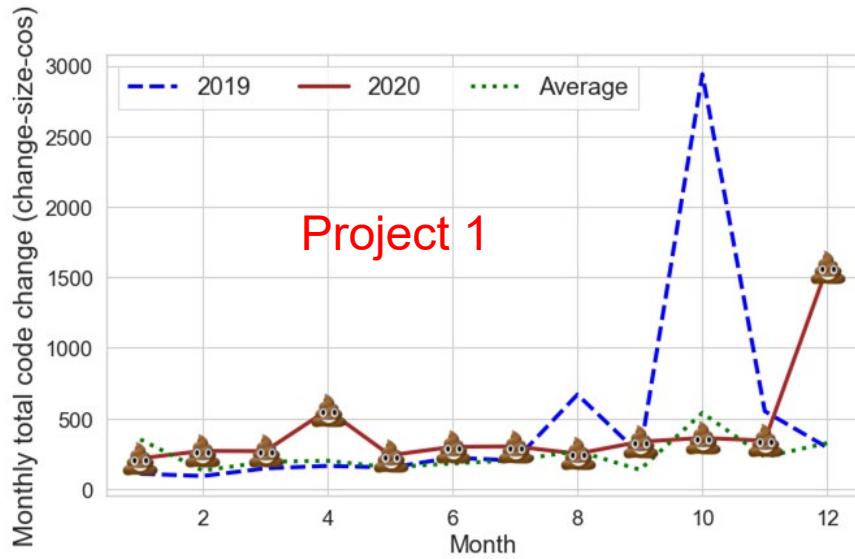
- Combining different “change size” metrics

- Example on right: simple LOCC totals and cos distance

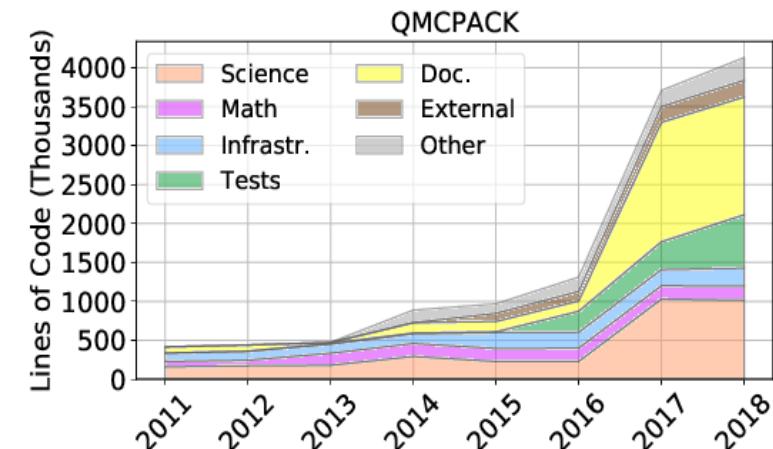
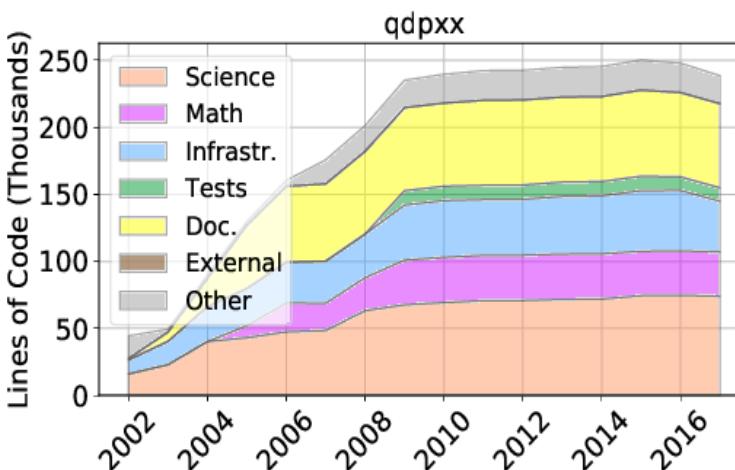
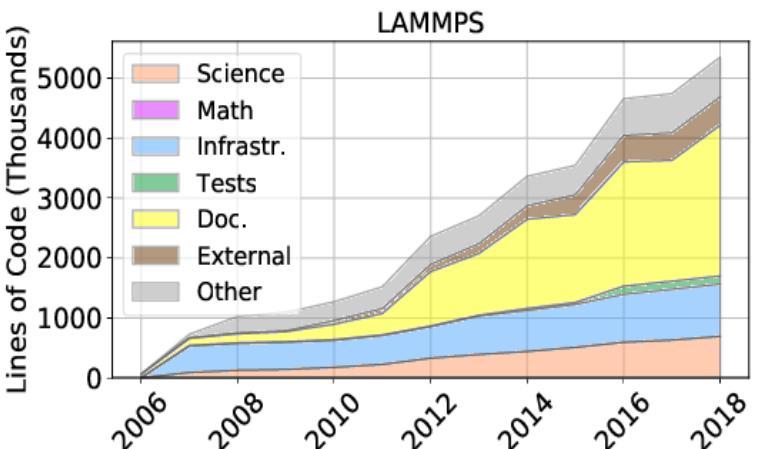
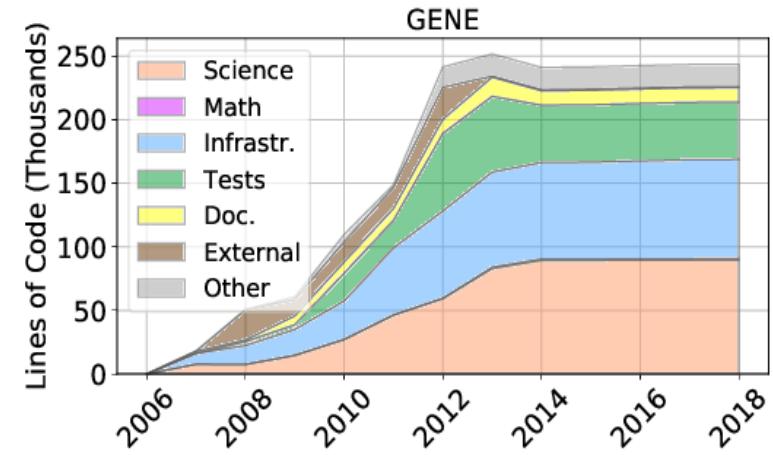
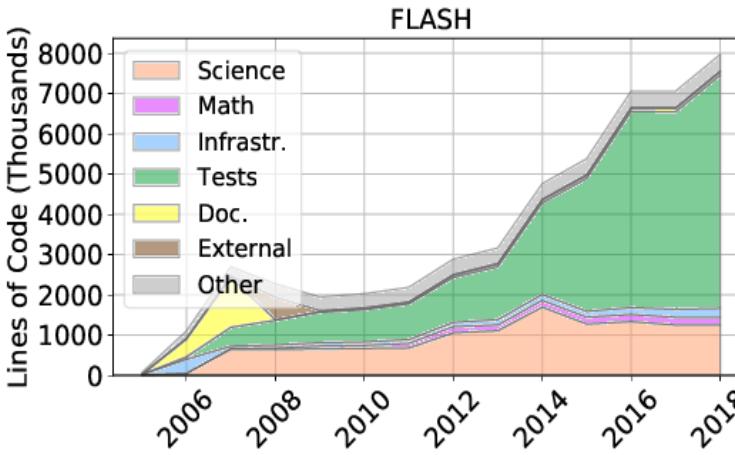
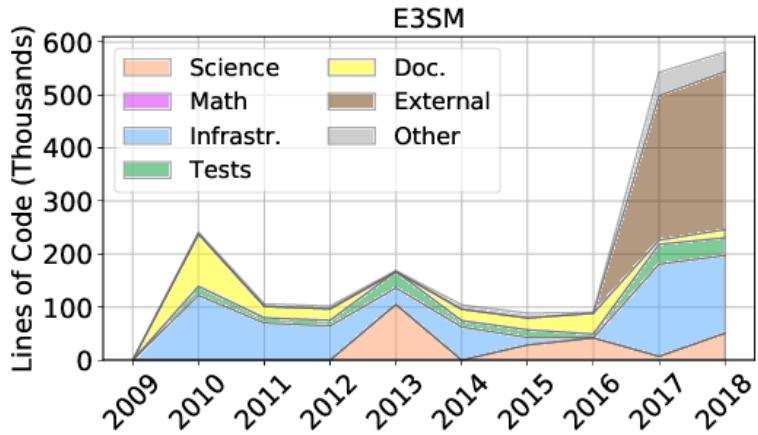
Significant events



How do projects weather interesting times?



Where is development effort going?



Grannan, A., Sood, K., Norris, B., & Dubey, A. (2020). Understanding the landscape of scientific software used on high-performance computing platforms. *The International Journal of High Performance Computing Applications*. <https://doi.org/10.1177/1094342019899451>

Example Tool: MeerCAT

Using git data to improve developer processes and efficiency

A Draft PR is created to merge feature branch into main branch.

The MeerCat PRA triggers and analyzes file changes and reports back.

More traditional Code Quality and testing tools trigger and report results.

A screenshot of a GitHub pull request page for a file named 'arithmetic.py'. The PR is labeled as 'Draft' and wants to merge 1 commit from the 'file_explorer_test' branch into the 'main' branch. A comment from 'uomeercat' is shown, edited by 'jp prideaux'. The comment is titled 'MeerCat report:' and contains the following text:

- The one file in the PR has problems with its docstring. I can help fix it.
- The changes to the file will likely cause issues with existing test cases. I can identify those cases and suggest changes.
- I can help add useful labels to the PR.
- I found the following people who may wish to join the discussion before the PR is accepted:
 - @jp prideaux - past committer and expert on the file.
 - @Juan-Pablo-Flores - reviewer of past accepted PRs including the file.
 - @fickas - major committer to files in the same directory.

Please see my Pull-Request Assistant for more details.

A screenshot of a GitHub pull request page for a file named 'arithmetic.py'. The PR is labeled as 'Draft' and wants to merge 1 commit from the 'file_explorer_test' branch into the 'main' branch. The status bar at the bottom says 'Add more commits by pushing to the file_explorer_test branch on flickas/anl_test_repo.' Below the status bar, there is a summary of CI checks and code scanning results:

- Some checks were not successful**
2 successful and 1 failing checks
 - CodeQL / Analyze (python) (pull_request)** Successful in 1m — Analyze (python) [Details](#)
 - ci/circleci: build-and-test** — Your tests failed on CircleCI [Details](#)
 - Code scanning results / CodeQL** Successful in 4s — No new or fixed alerts [Details](#)
- This pull request is still a work in progress**
Draft pull requests cannot be merged. [Ready for review](#)

Merge pull request [▼](#) You can also [open this in GitHub Desktop](#) or view [command line instructions](#).

The MeerCat PRA provides the following analysis:

1. Are required docstrings in place and actually match the code?
 2. If changes to files will cause existing test cases to become misaligned.
 3. Potentially useful labels to add to the PR for other tools to use.
 4. Suggests other people to add to the discussion given their prior roles related to files in this PR.

The link will take the user to the MeerCat PRA site for further refinement of the PR (see next slide).

! Draft **Update arithmetic.py #24**
fickas wants to merge 1 commit into [main](#) from [file_explorer_test](#)

 uomeercat commented 19 hours ago • edited by jprideaux

[Collaborator](#)  

MeerCat report:

- The one file in the PR has problems with its docstring. I can help fix it.
- The changes to the file will likely cause issues with existing test cases. I can identify those cases and suggest changes.
- I can help add useful labels to the PR.
- I found the following people who may wish to join the discussion before the PR is accepted:
 - @jprideaux - past committer and expert on the file.
 - @Juan-Pablo-Flores - reviewer of past accepted PRs including the file.
 - @fickas - major committer to files in the same directory.

[Please see my Pull-Request Assistant for more details.](#)

Add more commits by pushing to the [file_explorer_test](#) branch on [fickas/anl_test_repo](#).

 **Some checks were not successful** [Hide all checks](#)

2 successful and 1 failing checks

  CodeQL / Analyze (python) (pull_request)	Successful in 1m — Analyze (python) Details
  ci/circleci: build-and-test — Your tests failed on CircleCI	Details
  Code scanning results / CodeQL	Successful in 4s — No new or fixed alerts Details

 **This pull request is still a work in progress** [Ready for review](#)

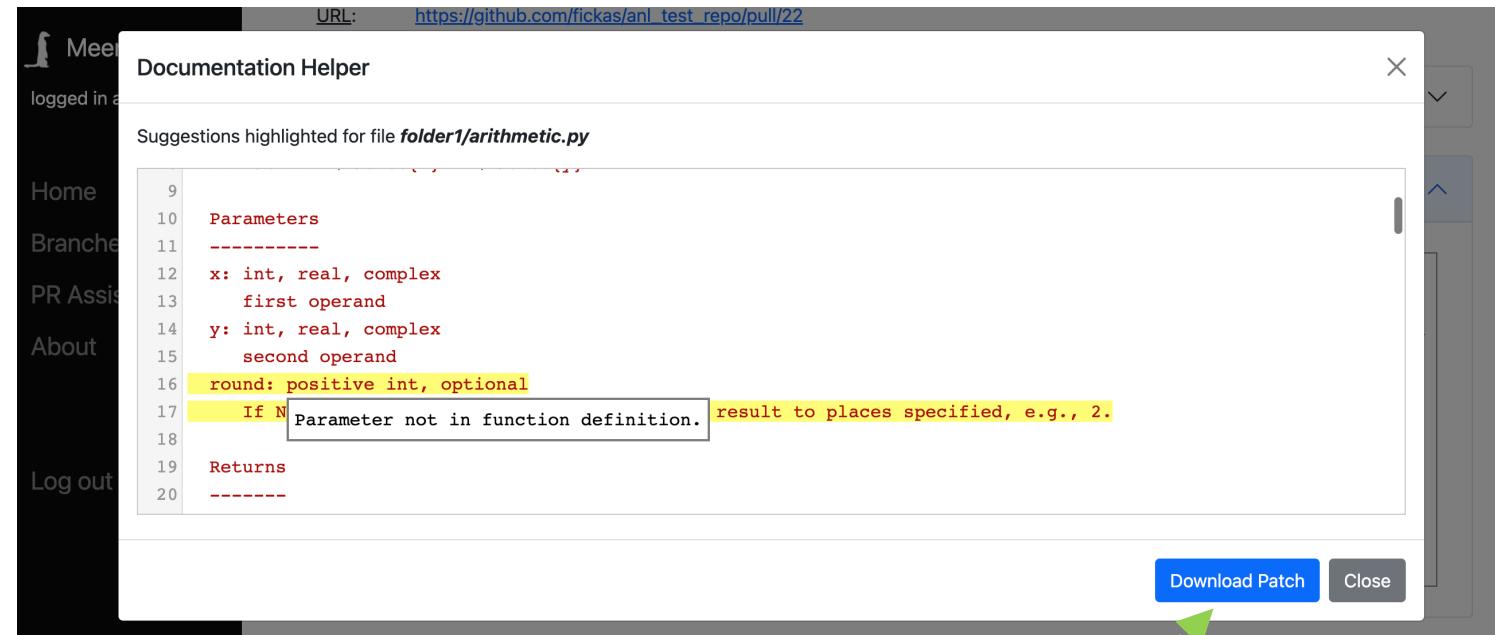
Draft pull requests cannot be merged.

[Merge pull request](#)  You can also [open this in GitHub Desktop](#) or view [command line instructions](#).

Once the user is at the PRA site, several aids are available.

After analysis, the PRA finds that one commit has removed a parameter from a function `sub`, but there is been no change to the documentation, i.e., code and documentation are misaligned.

Here the PRA editor is highlighting the Numpy docstring code that needs to be changed and why.



The user can download changes made to a Git Patch file and then easily merge in with existing PR.

The PRA can also search the repo for test files that reference the now changed ***sub*** function.

It discovers one such file (now misaligned) and suggests changes in the editor. Once the user removes the now invalid cases, the PRA will add the test file (and commits) to the existing Pull Request through the Patch mechanism.



Suggestions highlighted for file *folder1/test_arithmetic.py*

```
1 from arithmetic import sub
2
3 #Unit tests for sub function
4 def test_sub():
5     assert sub(2,1)==1
6     assert sub(4,2)==2
7     assert sub(.3333, .1111, round=4)==.2222
8     assert Test cases no longer valid. .22
```

Download Patch Close

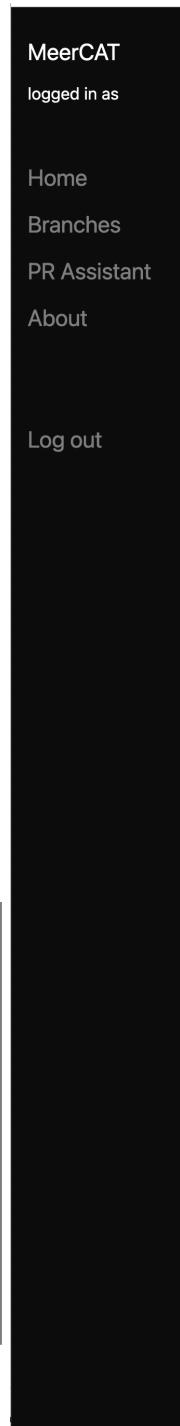
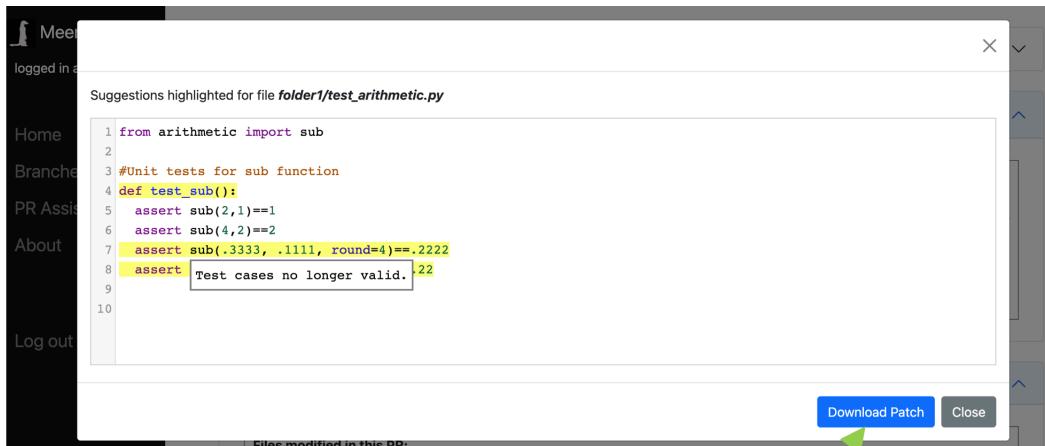
The user can download changes made to a Git Patch file and then easily merge in with existing PR.

MeerCAT: File Explorer

1. MeerCat does analysis and leaves comment in GitHub PR
2. If user clicks link, she is taken to the PRA tool (previous slides).
3. From the PRA, click on a specific file to get to the file explorer.

Benefits:

- + Problems detected early, don't have to wait for CI failure
- + Better overall documentation, testing



File Explorer

File name: folder1/arithmetic.py - [Go to file on GitHub](#)

Functions/Subroutines defined in file:

Signature	Doc String	Doc String Comments	Calling functions (test in red)
sub(x, y):	Yes	ARGUMENTS match	Go to folder1/arithmetic.py to find list sub on GitHub Go to folder1/test_arithmetic.py to find test sub on GitHub
list_sub(list1:list, list2: list) -> list:		No docstring found	
check_fum():		No docstring found	
mult(x, y):	Yes	ARGUMENTS match	

Developers:

Author	Total number of commits	Total lines changed	Date of last commit	Link to last commit
fickas - fickas@cs.uoregon.edu	11	182	Nov. 3, 2021, 11:24 a.m.	Go to commit on GitHub
Jason Prideaux - jprideau@cs.uoregon.edu	1	2	May 6, 2022, 10:31 a.m.	Go to commit on GitHub
Juan-Pablo-Flores - jpfloresd.97@gmail.com	2	96	June 6, 2022, 3:22 p.m.	Go to commit on GitHub

Change History

[Go to file blame on GitHub](#)

Included in Pull Requests:

PR #	PR URL	PR Issue link	Notes
12	Go to PR on GitHub	[", "]	TBD
13	Go to PR on GitHub	[", "]	TBD
14	Go to PR on GitHub	[", "]	TBD
16	Go to PR on GitHub	[", "]	TBD
17	Go to PR on GitHub	[", "]	TBD
22	Go to PR on GitHub	[", "]	TBD

Related to Issues:

In progress.

Interested Parties:

Part II: Analyzing Code



Defects that code analyses can catch

- ⌚ **Security:** Buffer overruns, improperly validated input.
- ⌚ **Memory safety:** Null dereference, uninitialized data.
- ⌚ **Resource leaks:** Memory, OS resources.
- ⌚ **API Protocols:** improper use of APIs, incomplete/incorrect implementations
- ⌚ **Exceptions:** Arithmetic/library/user-defined
- ⌚ **Encapsulation:** Accessing internal data, calling private functions.
- ⌚ **Data races:** Two threads access the same data without synchronization

Key idea: check compliance with (mostly) simple, mechanical design rules.

Standard: **ISO/IEC 5055:2021(E)**: Information technology — Software measurement — Software quality measurement — Automated source code quality measures



General-purpose tools for code checking (bugs, style)

➤ C/C++

- Run a bunch of general analyses with **scan-check**
(wrapper around `clang --analyze`, which uses the static analyzer below)
 - Minimally invasive, not very customizable
 - Works great with CMake and Autoconf builds
- Clang static analyzer component: *extensible* analysis framework for bug finding
 - Can do more complex analyses (path-sensitive, inter-procedural analysis based on a symbolic execution technique)
 - Requires more compiler knowledge to extend
- Clang-tidy: *extensible* (libTooling-based) framework for diagnosing typical programming errors or style issues
 - Checking and enforcing of simple coding conventions
 - Modular, provides API for implementing new checks
 - Relatively easy to integrate into Cmake

➤ Fortran

- Flang (compiler front-end to LLVM)
- Fortran-linter (limited)

The screenshot shows a browser window displaying a code analysis report from the HYPRE project. The code being analyzed is:

```
1083 }
1084
1085 #undef __FUNC__
1086 #define __FUNC__ "Mat_dhTranspose"
1087 void Mat_dhTranspose(Mat_dh A, Mat_dh *Bout)
1088 {
1089     START_FUNC_DH
1090     Mat_dh B;
1091
1092     if (np_dh > 1) { SET_V_ERROR("only for sequential"); }
1093
1094     Mat_dhCreate(&B); CHECK_V_ERROR;
1095     *Bout = B;
1096
1097     B->m = B->n = A->m;
1098     mat_dh_transpose_private(A->m, A->rp, &B->rp, A->cval, &B->cval,
1099                             A->aval, &B->aval); CHECK_V_ERROR;
1100 }
```

Annotations highlight several issues:

- 1 'B' declared without an initial value →
- 2 ← Assuming 'np_dh' is <= 1 →
- 3 ← Taking false branch →
- 4 ← Calling 'Mat_dhCreate' →
- 9 ← Returning from 'Mat_dhCreate' →
- 10 ← Assuming 'errFlag_dh' is false →
- 11 ← Taking false branch →
- 12 ← Assigned value is garbage or undefined →

Example development workflow that considers code quality

Example “make commit” workflow (easy in C/C++, and hopefully possible soon for Fortran):

- `clang-format` passes and reformats the code
- `clang-tidy` passes and enforces coding conventions
- `clang static analyzer` compiles debug and production builds (check errors)
- Project-specific analysis for debug and production build (check errors)
- debug/production builds get compiled and unit tests launched (check errors)
- production build + unit tests run under valgrind (check errors)
- production build gets compiled and unit test launched (check errors)
- production build with `--coverage` gets compiled and unit test launched against llvm-cov (write unit-test coverage stats)

Our goals and approach

Make it easy(-ish) to define and apply static and dynamic program analysis techniques to identify quality-related problems in HPC codes.

How? Two parts:

- A. By integrating general **static** and **dynamic** program analyses into the HPC software development process: mainly through documentation and examples.
- B. By creating easy interfaces to custom analyses, with examples.

Why?

- Abstraction
 - Elide details of a specific implementation.
 - Capture semantically relevant details; ignore the rest.
- Programs as data
 - Programs are just trees/graphs!
 - ...and we have lots of ways to analyze trees/graphs



Static program analysis is...

Ensure everything is checked the same way.

Examples:

- clang-tidy
- Clang static analyzer

Systematic examination of an abstraction of program state space.

Only track “important” things...

Applies to all possible executions.

Dynamic program analysis is...

Instrumented code only.

Examples:

- Valgrind
- Clang/LLVM
sanitizers (better!)

Partial examination of an abstraction of
a **single** execution path at **runtime**.

Can capture
information not
available statically.

Applies to specific
executions; can miss
errors.

Example: Using scan-build with HYPRE¹

```
hypre/src/cmbuild$ scan-build cmake ..  
hypre/src/cmbuild$ scan-build make
```



```
week4 — ssh -AY apollo — 95x20  
[ 99%] Building C object CMakeFiles/HYPRE.dir/sstruct_ls/sys_pfmg_setup_interp.c.o  
[ 99%] Building C object CMakeFiles/HYPRE.dir/sstruct_ls/sys_pfmg_setup_rap.c.o  
[ 99%] Building C object CMakeFiles/HYPRE.dir/sstruct_ls/sys_pfmg_solve.c.o  
/home/users/norris/test/hypre/src/sstruct_ls/sys_pfmg_solve.c:157:25: warning: The left operand  
of '>' is a garbage value [core.UndefinedBinaryOperatorResult]  
    if (b_dot_b > 0)  
        ~~~~~~ ^  
/home/users/norris/test/hypre/src/sstruct_ls/sys_pfmg_solve.c:168:22: warning: The right operan  
d of '/' is a garbage value [core.UndefinedBinaryOperatorResult]  
    if ((r_dot_r/b_dot_b < eps) && (i > 0))  
        ~~~~~~  
2 warnings generated.  
[ 99%] Building C object CMakeFiles/HYPRE.dir/sstruct_ls/sys_semi_interp.c.o  
[ 99%] Building C object CMakeFiles/HYPRE.dir/sstruct_ls/sys_semi_restrict.c.o  
[100%] Linking C static library libHYPRE.a  
[100%] Built target HYPRE  
scan-build: Analysis run complete.  
scan-build: 1136 bugs found.  
scan-build: Run 'scan-view /tmp/scan-build-2021-05-27-073157-25436-1' to examine bug reports.  
norris@apollo:~/test/hypre/src/cmbuild$
```

¹ [HYPRE: Scalable Linear Solvers and Multigrid Methods](https://github.com/hypre-space/hypre). <https://github.com/hypre-space/hypre>

Example: hypre (cont.)

```
hypre/src/cmbuild$ scan-view /tmp/scan-build-2021-05-27-073157-25436-1
```

The screenshot shows a browser window titled 'cmbuild - scan-build results' at the URL '127.0.0.1:8181'. The page displays the following information:

User: norris@apollo
Working Directory: /home/users/norris/test/hypre/src/cmbuild
Command Line: make
Clang Version: clang version 13.0.0 (https://github.com/llvm/llvm-project.git b7911e80d6926f9280ceb23d4e86e25c29370904)
Date: Thu May 27 07:31:57 2021

Bug Summary

Bug Type	Quantity	Display?
All Bugs	1136	<input checked="" type="checkbox"/>
Logic error		
Array subscript is undefined	3	<input checked="" type="checkbox"/>
Assigned value is garbage or undefined	33	<input checked="" type="checkbox"/>
Branch condition evaluates to a garbage value	9	<input checked="" type="checkbox"/>
Dereference of null pointer	378	<input checked="" type="checkbox"/>
Dereference of undefined pointer value	131	<input checked="" type="checkbox"/>
Division by zero	2	<input checked="" type="checkbox"/>
Garbage return value	2	<input checked="" type="checkbox"/>
Result of operation is garbage or undefined	66	<input checked="" type="checkbox"/>
Uninitialized argument value	56	<input checked="" type="checkbox"/>
Unused code		
Dead assignment	387	<input checked="" type="checkbox"/>
Dead increment	10	<input checked="" type="checkbox"/>
Dead initialization	57	<input checked="" type="checkbox"/>
Dead nested assignment	2	<input checked="" type="checkbox"/>

details

Eek! Too much information, can synthesize a more actionable report.

The screenshot shows a terminal window with the command 'hypre/src/cmbuild\$ scan-view /tmp/scan-build-2021-05-27-073157-25436-1' in red. The terminal output shows a code editor with annotated lines of C code. The annotations are numbered 1 through 12 and highlight various issues:

```
1083 }
1084
1085 #undef __FUNC__
1086 #define __FUNC__ "Mat_dhTranspose"
1087 void Mat_dhTranspose(Mat_dh A, Mat_dh *Bout)
1088 {
1089     START_FUNC_DH
1090     Mat_dh B;
1091
1092     1 'B' declared without an initial value →
1093
1094     if (np_dh > 1) { SET_V_ERROR("only for sequential"); }
1095
1096     2 ← Assuming 'np_dh' is <= 1 →
1097
1098     3 ← Taking false branch →
1099
1100     Mat_dhCreate(&B); CHECK_V_ERROR;
1101
1102     4 ← Calling 'Mat_dhCreate' →
1103
1104     9 ← Returning from 'Mat_dhCreate' →
1105
1106     10 ← Assuming 'errFlag_dh' is false →
1107
1108     11 ← Taking false branch →
1109
1110     *Bout = B;
1111
1112     12 ← Assigned value is garbage or undefined →
1113
1114     B->m = B->n = A->m;
1115     mat_dh_transpose_private(A->m, A->rp, &B->rp, A->cval, &B->cval,
1116                             A->aval, &B->aval); CHECK_V_ERROR;
1117
1118     END_FUNC_DH
1119 }
```

What about project-specific requirements?

Do you need to be a compiler expert to implement new program checks?

Thankfully -- no!



Implementation approach: Part B

Develop custom static and dynamic checking based on project-specific requirements.

- C/C++
 - Static: use and extend existing APIs (Clang static analyzer, clang-tidy); implement custom AST traversals and matchers (more details next)
 - Dynamic: use Clang sanitizer APIs; python for simplicity and easy of extensions by HPC software developers
- Fortran
 - Static: need to write new Flang-based checkers *only* for things that Fortran developers actually care about
 - Dynamic: do we need anything?

Example: Checking for violations of PETSc developer rules

From PETSc Style and Usage Guide: <https://petsc.org/release/developers/style/>

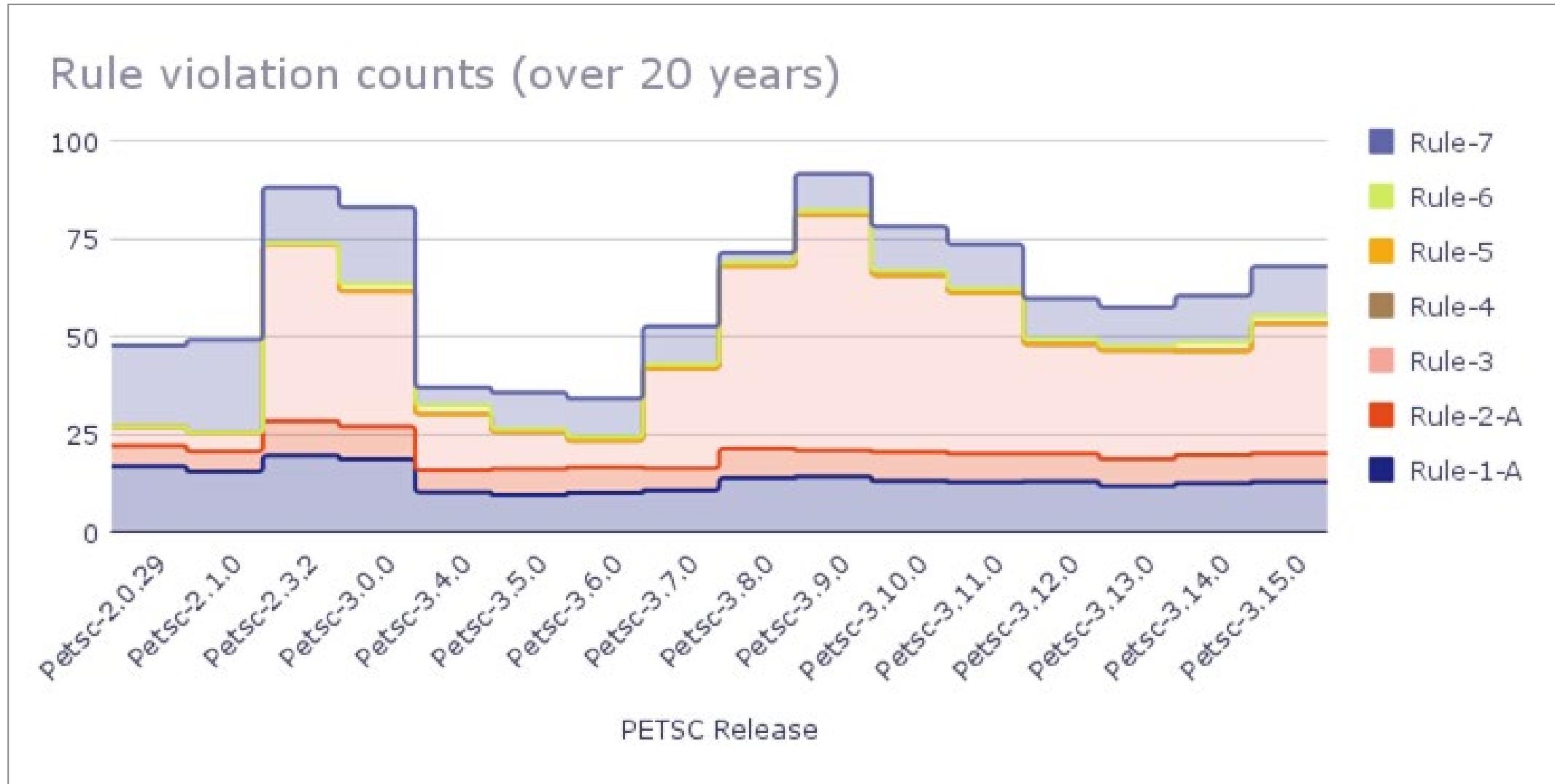
Examples of project-specific rule violations (PETSc 3.14)

PETSc Rule	PETSc Construct	Description	path	Line	Column
Rule-1	Function definition in the library	PetscErrorCode PETSCMAP1(VecScatterBeginMPI3Node)(VecScatter ctx, Vec xin, Vec yin, InsertMode addv, ScatterMode mode)	~/petsc-3.14.3/src/vec/vscat/impls/mpi3/vpscat.h	249	16
Rule-2	Macro in the library	#define mpi_reduce_scatter PETSC_MPI_REDUCE_SCATTER	~/petsc-3.14.3/include/petsc/mpiuni/mpiunifdef.h	118	2
Rule-3	Function declaration in the library	PETSC_INTERN PetscErrorCode MatFactorFactorizeSchurComplement_Private(Mat);	~/petsc-3.14.3/include/petsc/private/matimpl.h	494	29
	Function declaration in the library	PETSC_EXTERN PetscErrorCode MatFactorFactorizeSchurComplement(Mat);	~/petsc-3.14.3/include/petscmat.h	1245	29
Rule-4	Function definition in the library	PETSC_EXTERN PetscErrorCode DMDAVecGetArray(DM, Vec, void *)	~/petsc-3.14.3/include/petscdmda.h	113	29
	Function call in the application	ierr = VecGetArray(y, yy)	~/petsc-3.14.3/include/petscvec.h	545	10
Rule-5	Function call in the library	ierr = PetscFEPushforwardGradient(fe, fegeom, 1, interpolantGrad);	~/petsc-3.14.3/include/petsc/private/petscfeimpl.h	332	10
Rule-6	If in the library	if (p == 0) return node;	~/petsc-3.14.3/src/dm/impls/plex/gmshlex.h	231	3
Rule-7	Macro in the library	#ifndef PETSC4PY_COMPAT_MUMPS_H	~/petsc-3.14.3/src/binding/petsc4py/src/include/compat/mumps.h	1	1

Rules from: <https://petsc.org/release/developers/style/>



Example results for a subset of the PETSc rules



Capabilities summary

Type of Data/Analysis	Database	Examples	Repository Location
Git data: commits, changes (lines, files, etc.)	✓	✓	github.com/CAT-SDK/GremCat/
Github and Gilab issues and associated metadata	✓	✓	github.com/CAT-SDK/GremCat/
Code quality checkers (dynamic & static)	✗	✓	<ul style="list-style-type: none">github.com/HPCL/code-analysis (dynamic)github.com/HPCL/llvm-project/tree/xsdk-uo/clang-tools-extra/clang-tidy/petsc (static)
Mailing lists	✓	✓	Not publicly available yet, contact norris@cs.uoregon.edu

Summary

- We introduced a **flexible, efficient, and usable** software framework for **acquiring, storing, manipulating, and visualizing** development-related data.
- We demonstrated a few of its capabilities here; a growing number of analyses and tools are continuously being developed.
 - **Contributions and/or requests welcome!** <https://github.com/CAT-SDK/GremCat>
- Acknowledgments: DOE ECP IDEAS Productivity Project
 - Carter Perkins, Bosco Ndemeye, Stephen Fickas, University of Oregon
 - Armando Acosta and Kanika Sood, California State University, Fullerton
 - Anshu Dubey and Lois Curfman McInnes, Argonne National Laboratory

Thank you!

ECP projects that may be present in examples in this presentation: Spack, LAMMPS, PETSc, Nek5000
E3SM, QMCPACK, QDPXX, LATTE NAMD, HYPRE, fast-export, Enzo, TAU2, xpress-apex, LATTE, NWChem

