

Program Analysis

1. Introduction

Kihong Heo

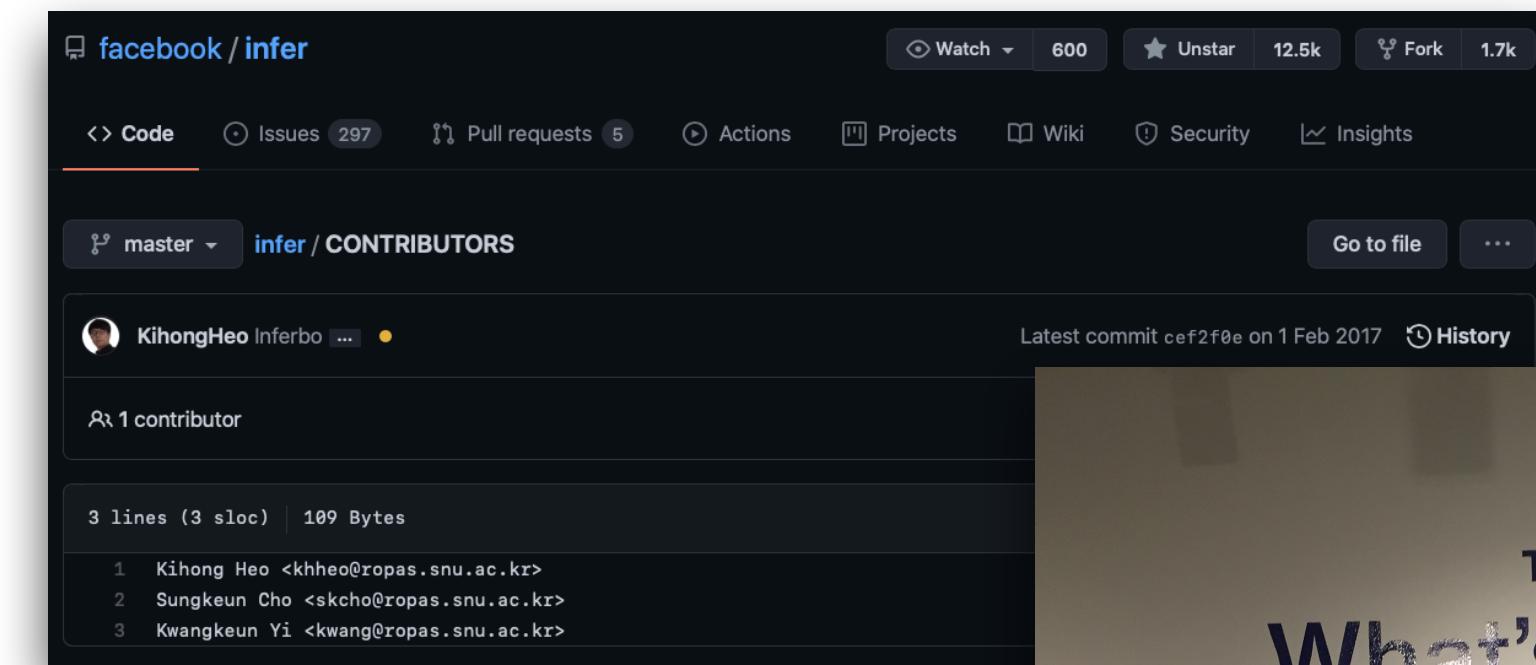


About Me

- Instructor: Kihong Heo (허기홍, kihong.heo@kaist.ac.kr)
- KAIST CS / GSIS / Programming Systems Lab.
- Homepage: <https://kihongheo.kaist.ac.kr> / <https://prosys.kaist.ac.kr>
- Office: N5 2321
- Office Hours: after each class (by appointment)

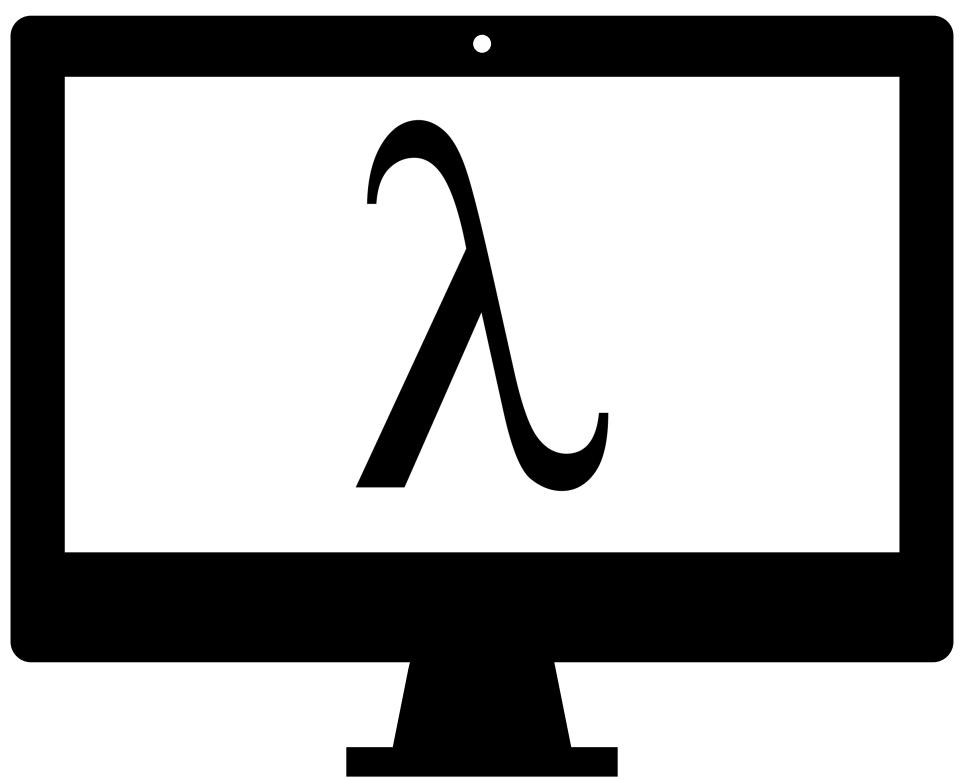
My Research

- Goal: solid PL theories \Leftrightarrow powerful programming systems
- Keywords: programming language, program analysis, SW engineering, SW security
- Good memories:



*<https://research.fb.com/blog/2017/02/inferbo-infer-based-buffer-overrun-analyzer/>

My Research

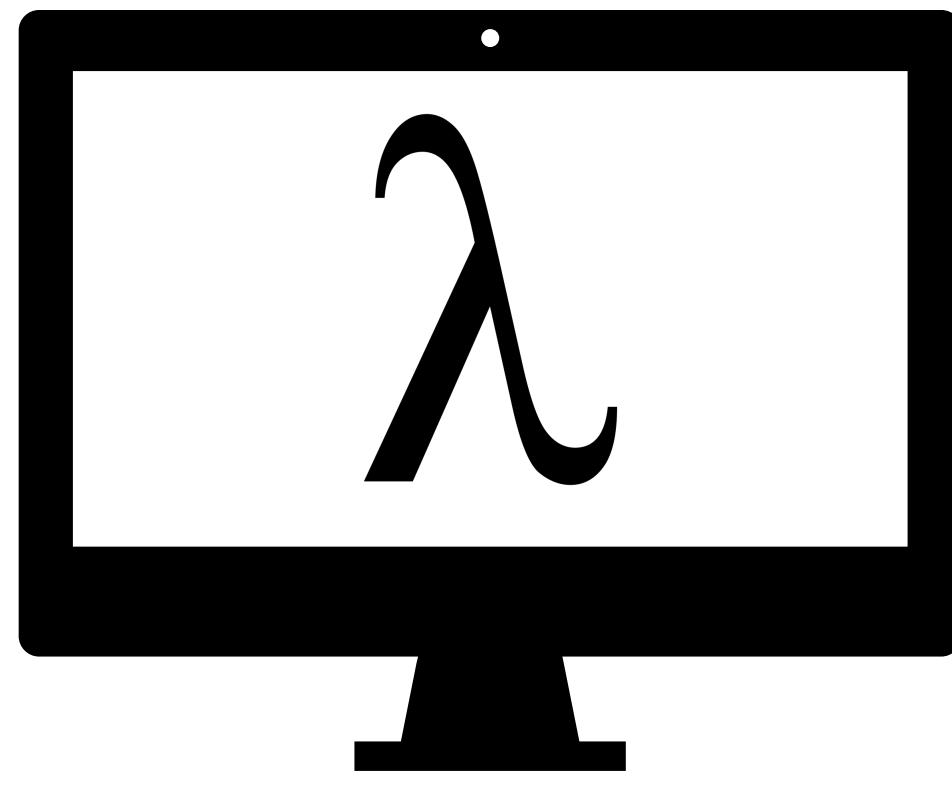


**Next-generation
Programming Systems**

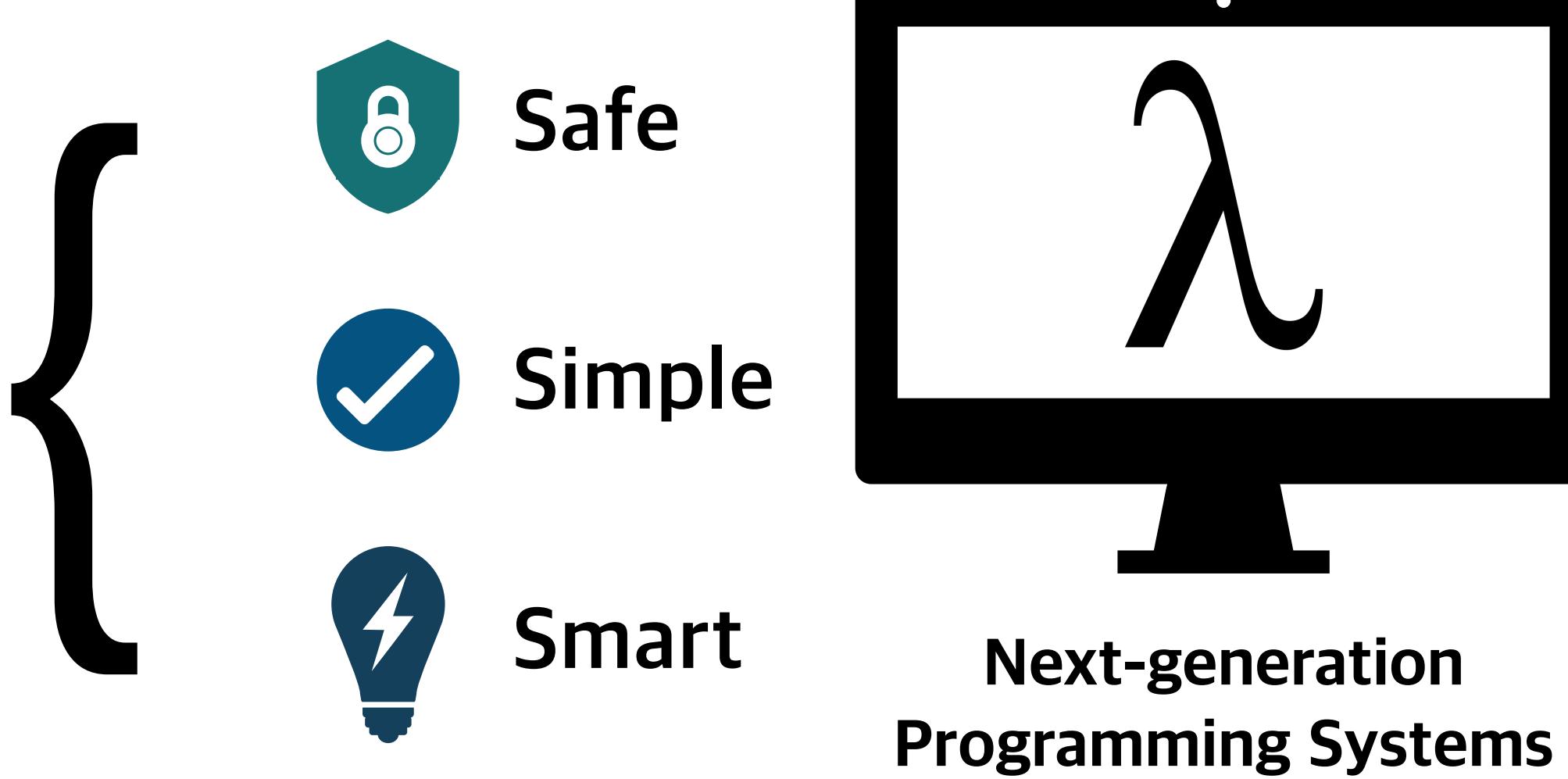
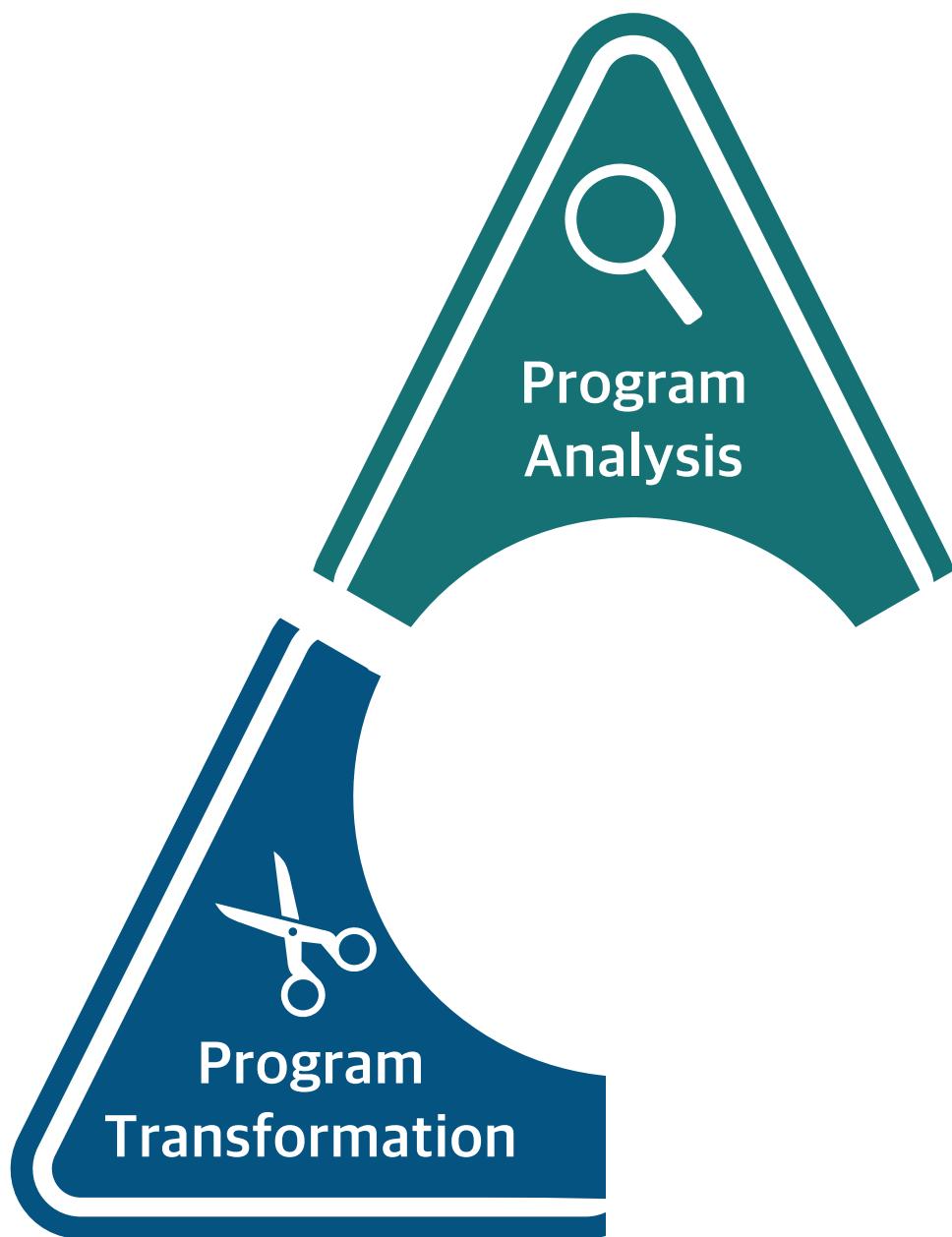
My Research



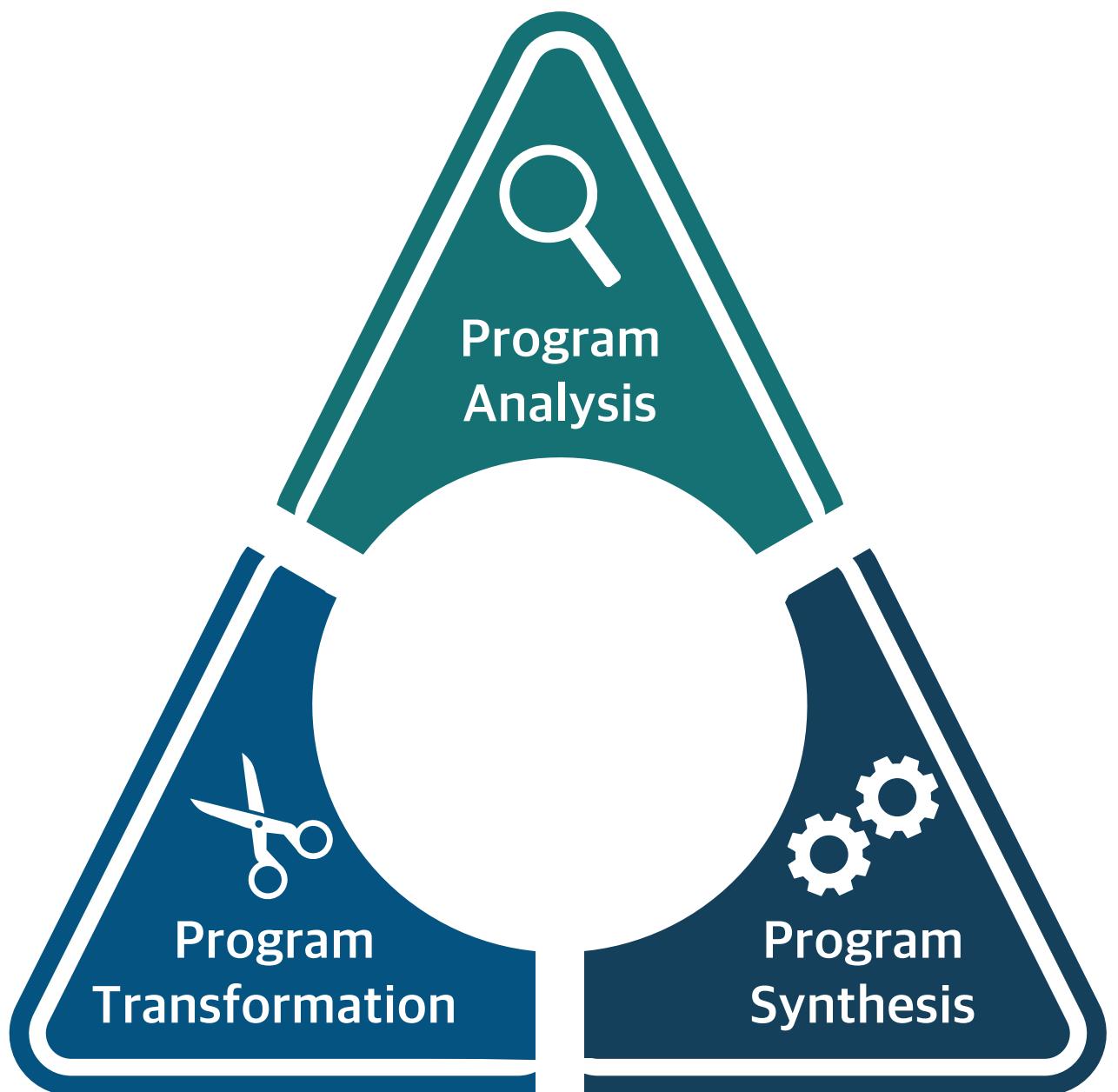
{

**Safe****Simple****Smart****Next-generation
Programming Systems**

My Research



My Research



{



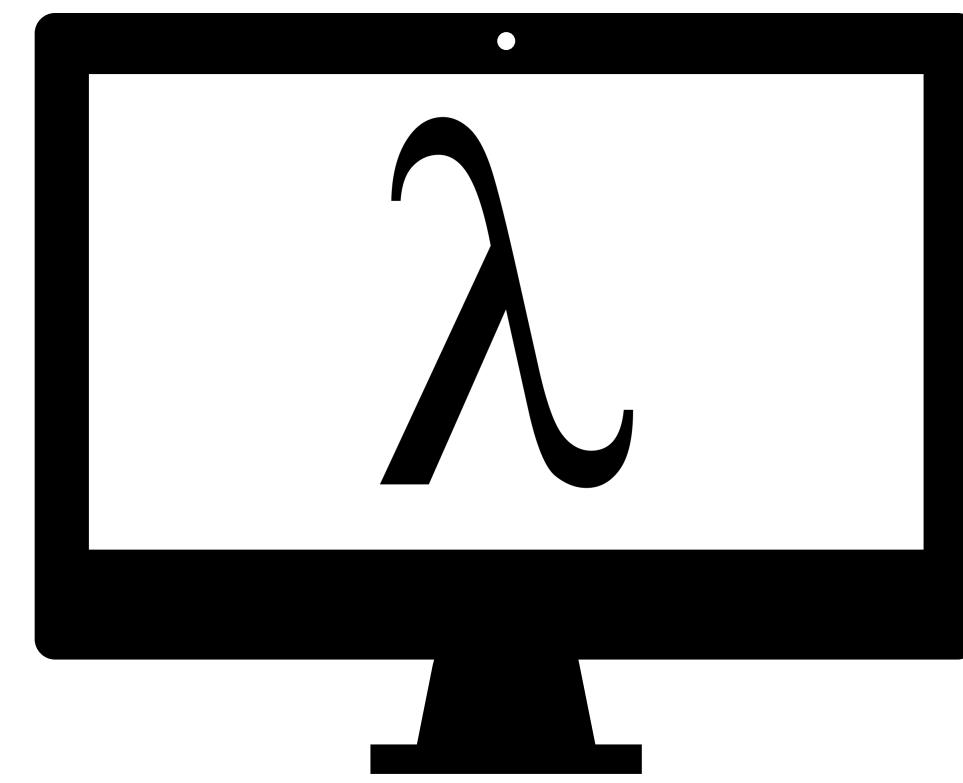
Safe



Simple

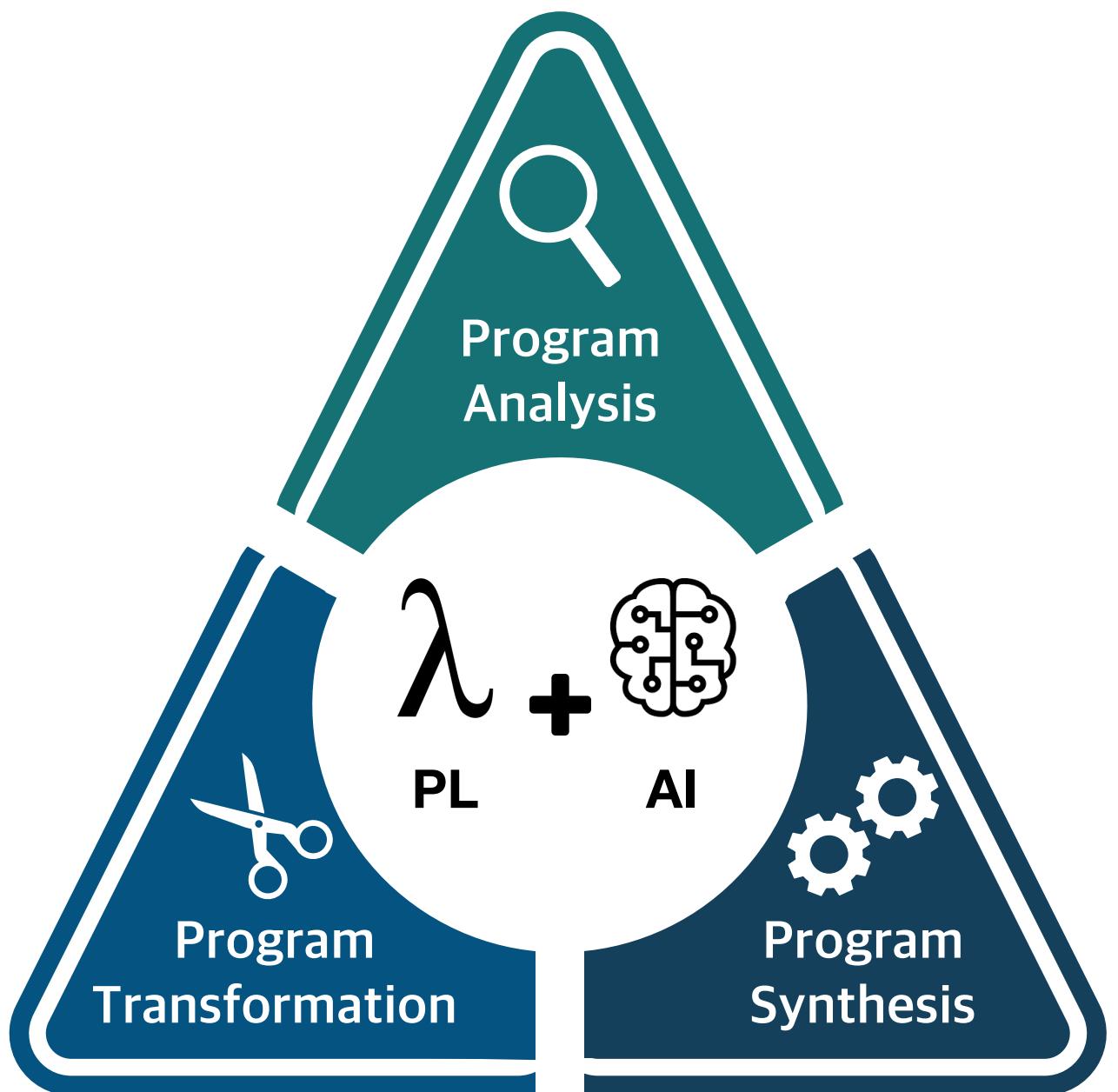


Smart



Next-generation
Programming Systems

My Research



{



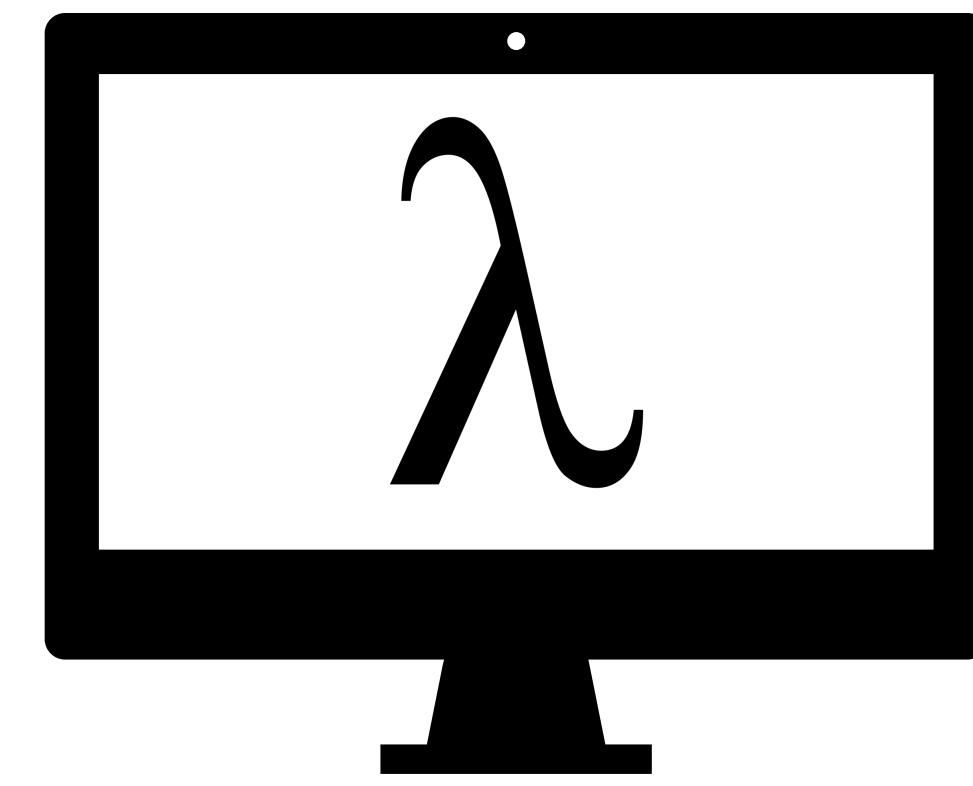
Safe



Simple



Smart



Next-generation
Programming Systems

Course Information

- Course Website: <https://github.com/prosyslab-classroom/cs524-program-analysis>
- Q&A Board: <https://github.com/prosyslab-classroom/cs524-program-analysis/issues>
- TAs (mailing list: cs524.ta@prosys.kr)
 - Tae Eun Kim (김태은)
 - Geon Park (박건)
- Textbook:
 - Lecture slides will be provided
 - Xavier Rival and Kwangkeun Yi,
Introduction to Static Analysis: an Abstract Interpretation Perspective, MIT Press, 2020

Grading

- Homework: 50%
- Final exam: 40%
- Participation: 10%
 - Active participation including questions or discussions (online or offline)
- DO NOT allow for P/NR grading
- NOTE: **nonnegotiable!**
 - DO NOT send an email for a negotiation (cheating)

Important Notice (1): Academic Integrity

- DO NOT share the course contents (e.g., assignments or exams) with others
 - Esp., Github public repository, chegg.com, etc
- DO NOT discuss the details of solutions with others
- DO NOT plagiarize
 - We keep all the submissions from previous years!
- Any integrity violation: at **LEAST F**
- If you have questions: QnA board > TAs > instructor

전산학부 명예규정 2022 봄학기 / School of Computing Honor Code 2022 Spring

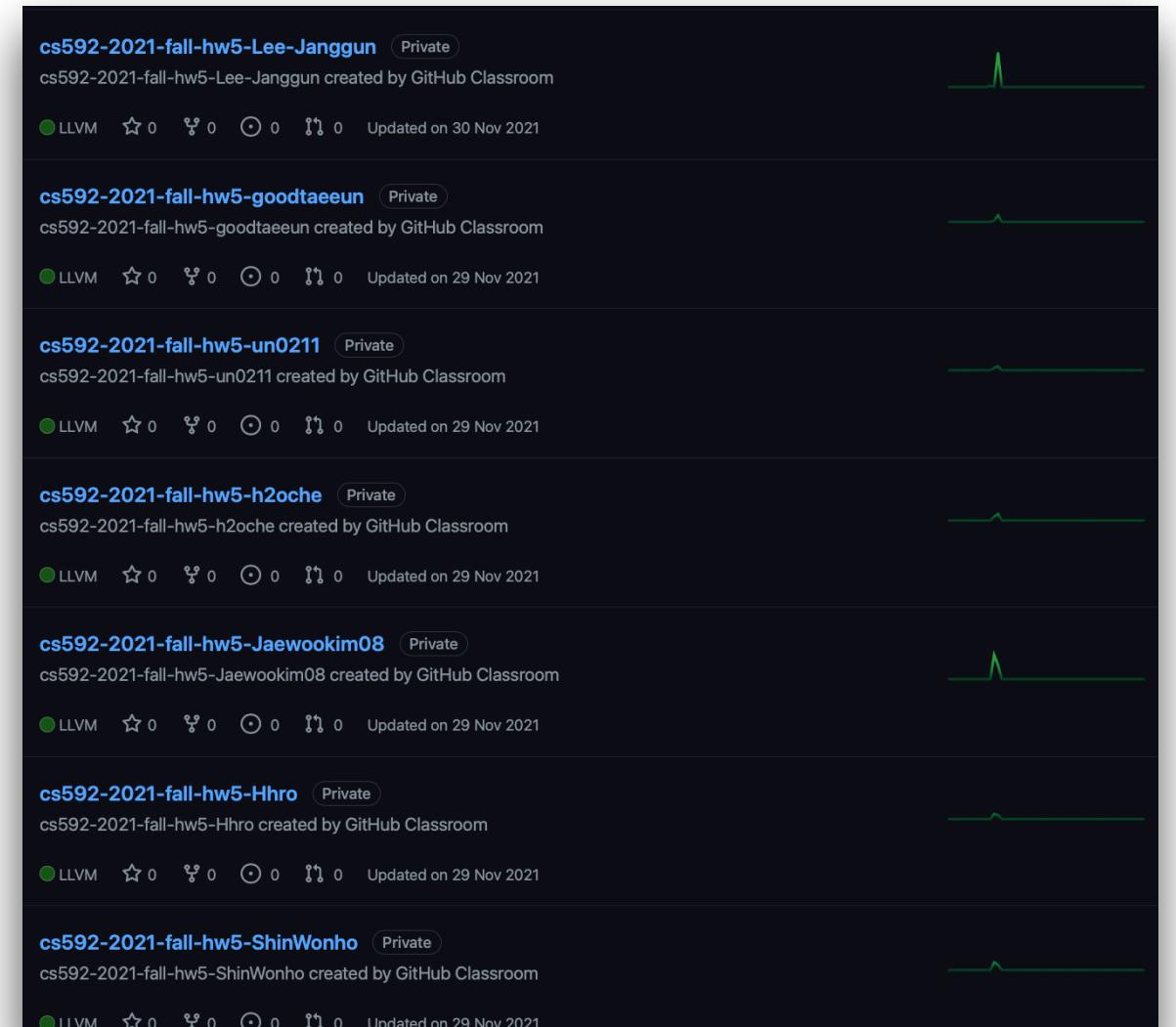
카이스트 전산학부가 운영하는 모든 수업에 참여하는 학생은 개인의 명예와 타인의 권리를 함께 존중하며 성실성과 정직성을 지키기 위하여 최선을 다합니다. 모든 시험 및 과제를 작성에 있어 허위되거나 다른 형태의 도움도 받지 않습니다. 다음의 행위들은 학생의 성실성과 정직성을 위반하는 것으로 간주됩니다:

- 본인 이외의 사람/기관이 작성한 답안지, 숙제, 프로그램 소스 코드, 보고서 등을 참고 및 이용하는 행위
- 시험 및 과제들과 관련해 [chegg.com](https://www.chegg.com)과 같이 정답을 공유하는 온라인 서비스를 이용하는 행위
- GitHub 등의 코드 저장소에 본인의 과제 답안을 공개하거나, 타인이 공개된 답안을 참고 및 이용하는 행위
- 다른 학생이 본인이 작성한 답안지, 숙제, 프로그램 소스 코드, 보고서 등을 참고하도록 용인하는 행위
- 다른 학생이 작성한 결과물을 자신의 것인 양 제출하는 행위
- 다른 학생을 대신해 시험을 치루는 행위
- 개인이 수행하도록 되어있는 take-home 시험이나 과제를 작성에 있어 허락 없이 공동 작업을 하거나 부적절한 도움을 받는 행위
- 표절, 길이와 무관하게 적절한 인용이나 언급 없이 타인의 창작물(참고서적, 문헌, 온라인상의 자료)을 무단으로 사용하는 행위

규정 위반 여부의 판단과 처벌 수위는 교과목 담당 교수에 의해 결정됩니다. 모든 부정행위는 전산학부 학사주임 교수 및 학부장님께 보고되어, 적발시 전산학부 내부적으로 아래와 같은 제재를 받습니다.

- 항후 2학기 동안 모든 포상 및 장학금 수여/추천 대상에서 제외함
- (주진공이 전산학부가 아닌 경우) 항후 2학기 동안 전산학부로의 전과 금지

위반의 심각성에 따라 학교 전체의 상별위원회에 회부될 수 있습니다. 카이스트 학사규정이 적용하는 칭계의 범위는 아래 첨부된 학생 징계 일정 기준을 참고하세요 (학생 핸드북 한글판 67페이지, 다운로드는 https://portal.kaist.ac.kr/ennotice/student_notice/11614934649338)

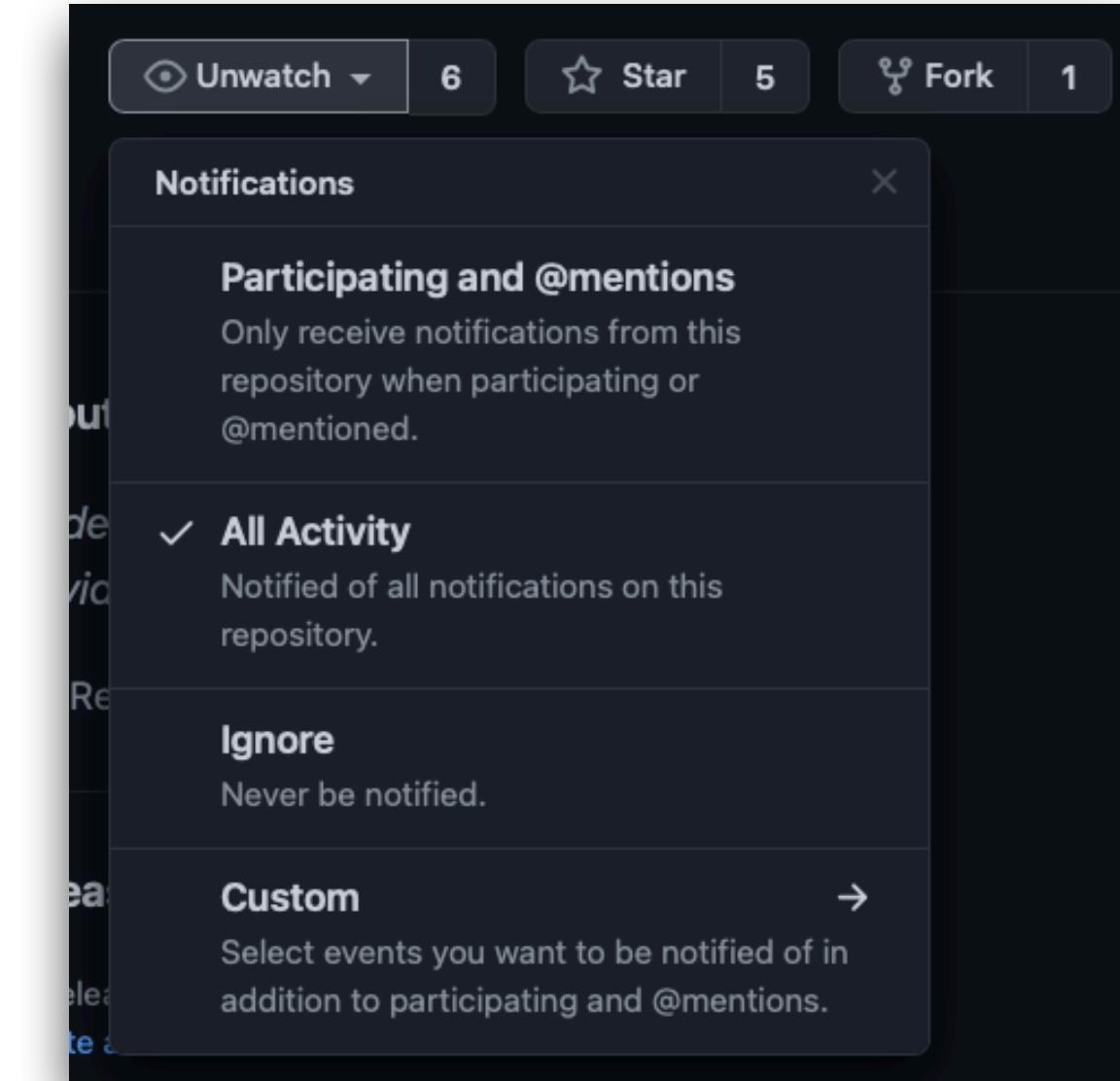


Important Notice (2): In-class

- Language: English (default), Korean (supplementary)
- Attendance: always (default), absence (if necessary)
 - No quantified attendance score
 - “What is **essential** is **invisible** to the eye” - The Little Prince
 - I expect you to be here, as you expect me to be here!
- Questions & discussion (either in Korean or in English): highly encouraged

Important Notice (3): Out-of-class

- All Q&A and public notices: Github issue board
 - “Watch” all notifications
- Private notices (grading, etc): KLMS
- Questions are always welcome except for
 - Too detailed ones (TAs are not debuggers!)
 - Directly related to the solutions
- Actively discuss with your classmates



A long time ago
in a galaxy far, far away....

**SOFTWARE
BUGS**

Software Bugs: A Persistent Problem

- A long time ago, far far away



The Patriot Missile (1991)
Floating-point roundoff
28 soldiers died



The Ariane-5 Rocket (1996)
Integer Overflow
\$100M



NASA's Mars Climate Orbiter (1999)
Meters-Inches Miscalculation
\$125M

- Unfortunately, it becomes your own problem now

CNN U.S. | World | Politics | Money | Opinion | Health | Entertainment | Tech | Style | Travel | Sports | Video | Live TV | **US**

The 'Heartbleed' security flaw that affects most of the Internet

By Heather Kelly, CNN
Updated 5:11 PM ET, Wed April 9, 2014

A large red heart icon with liquid dripping down from it, symbolizing the 'Heartbleed' bug.

This dangerous Android security bug could let anyone hack your phone camera

By Anthony Spadafina November 23, 2019

Camera app vulnerabilities allow attackers to remotely take photos, record video and spy on users

A smartphone displaying a green binary code pattern over a keyboard background, symbolizing a security vulnerability in mobile devices.

AERIAN MARSHALL / TRANSPORTATION 06:30 2019 07:00 AM

What Boeing's 737 MAX Has to Do With Cars: Software

Investigators believe faulty software contributed to two fatal crashes. A newly discovered fault will likely keep the 737 MAX grounded until the fall.

A Boeing 737 MAX airplane flying through a cloudy sky, symbolizing a major aviation safety issue due to software bugs.

Homeland Security warns that certain heart devices can be hacked



New in Life & Style

Interfaith 4th-graders bond through poetry, art and Steph Curry 2:03 PM

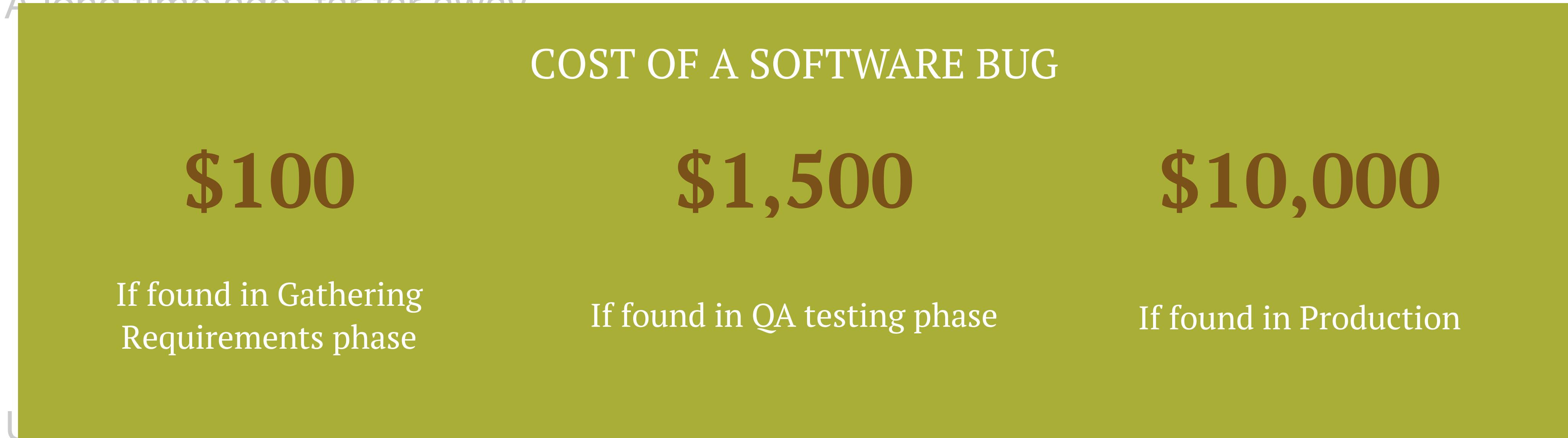
6 ways to celebrate Valentine's Day in Lake Geneva 8:55 AM

Six ways to keep your kids healthy during winter 8:56 AM

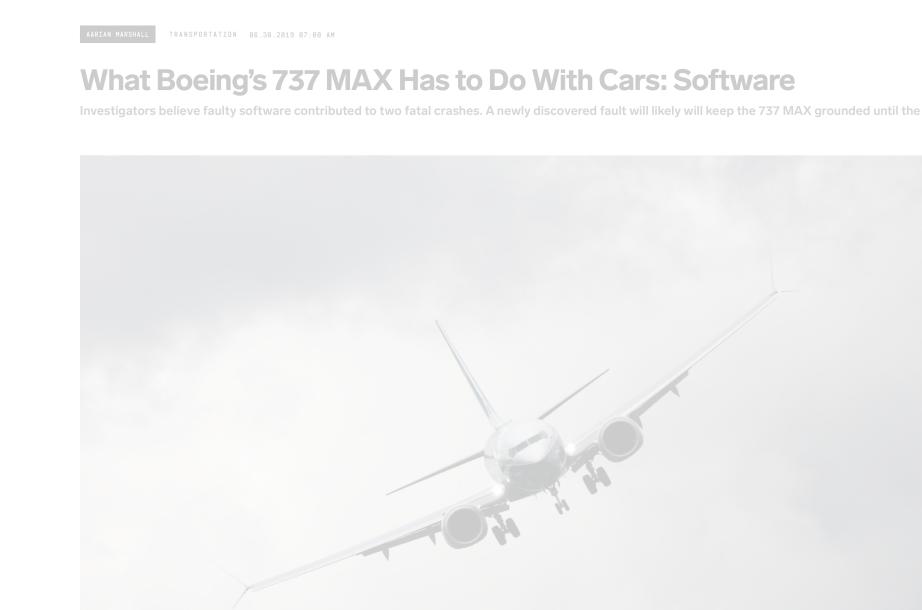
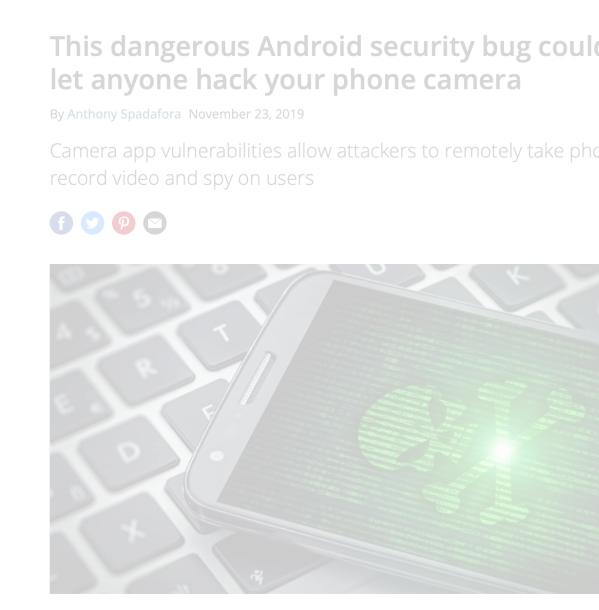
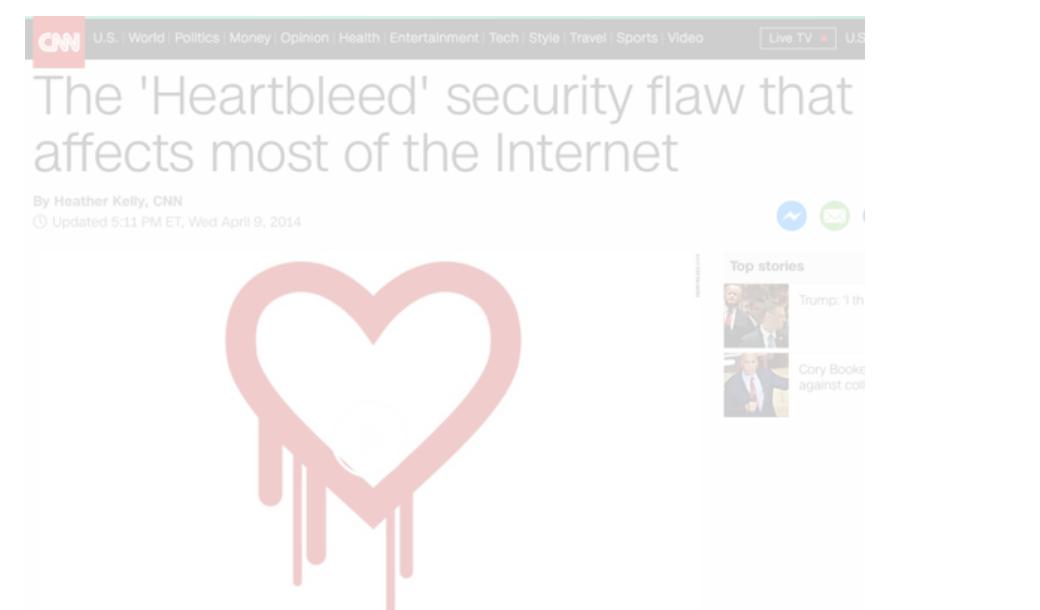
See More

Software Bugs: A Persistent Problem

- A long time ago, far far away...

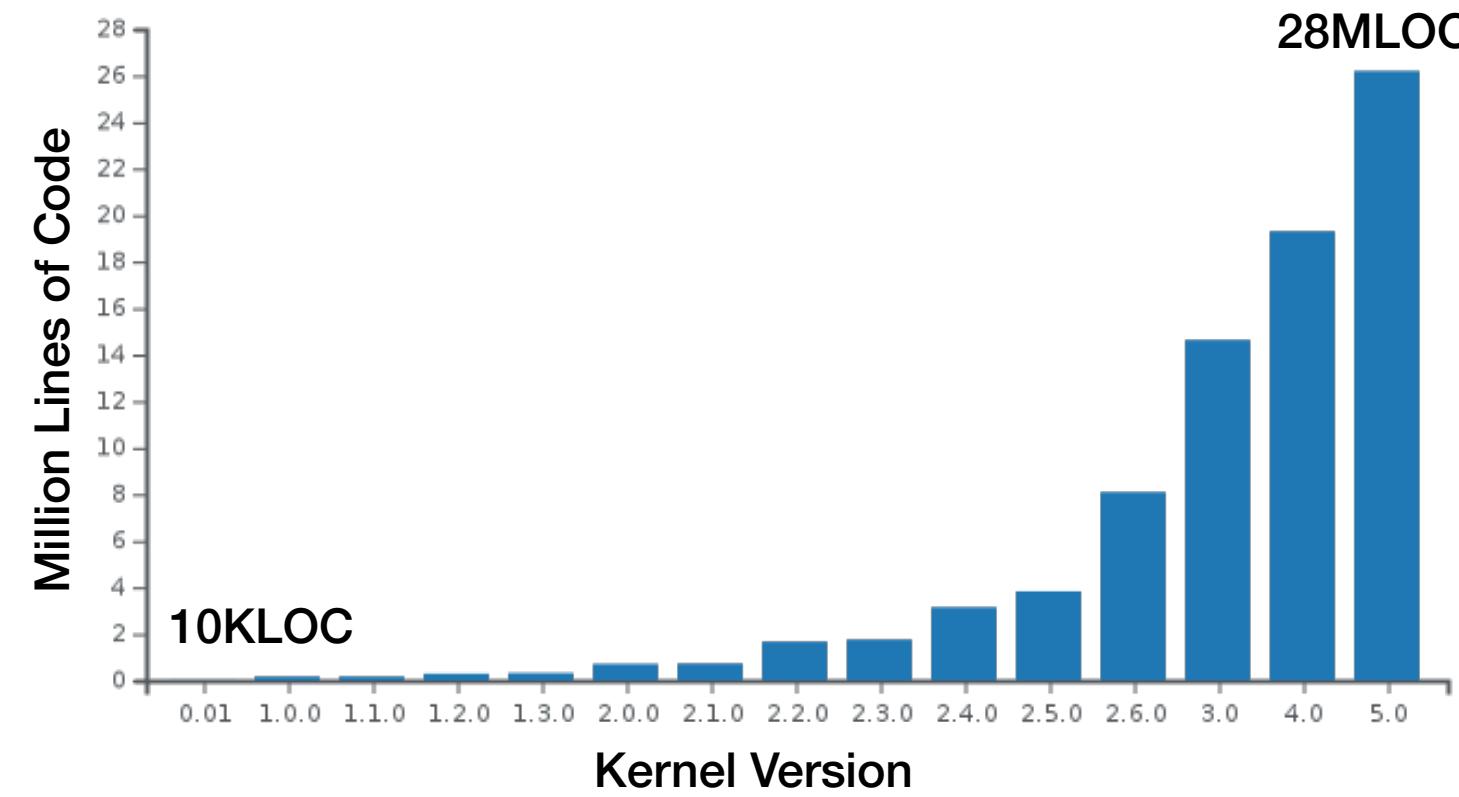


- IBM Systems Sciences Institute, 2015

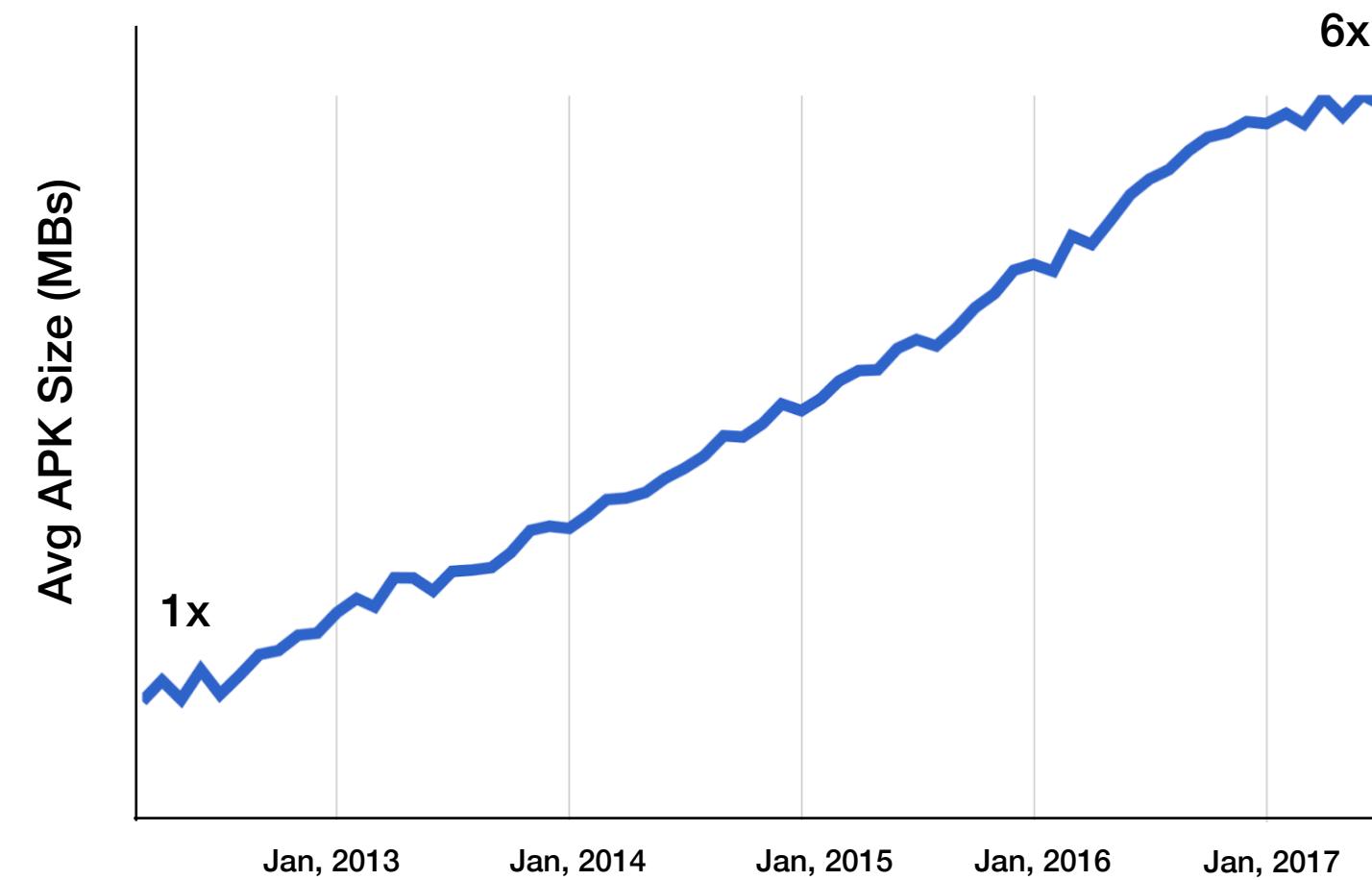


Why Software Still Fails?

Size of Linux Kernel



Avg. Size of Android Apps



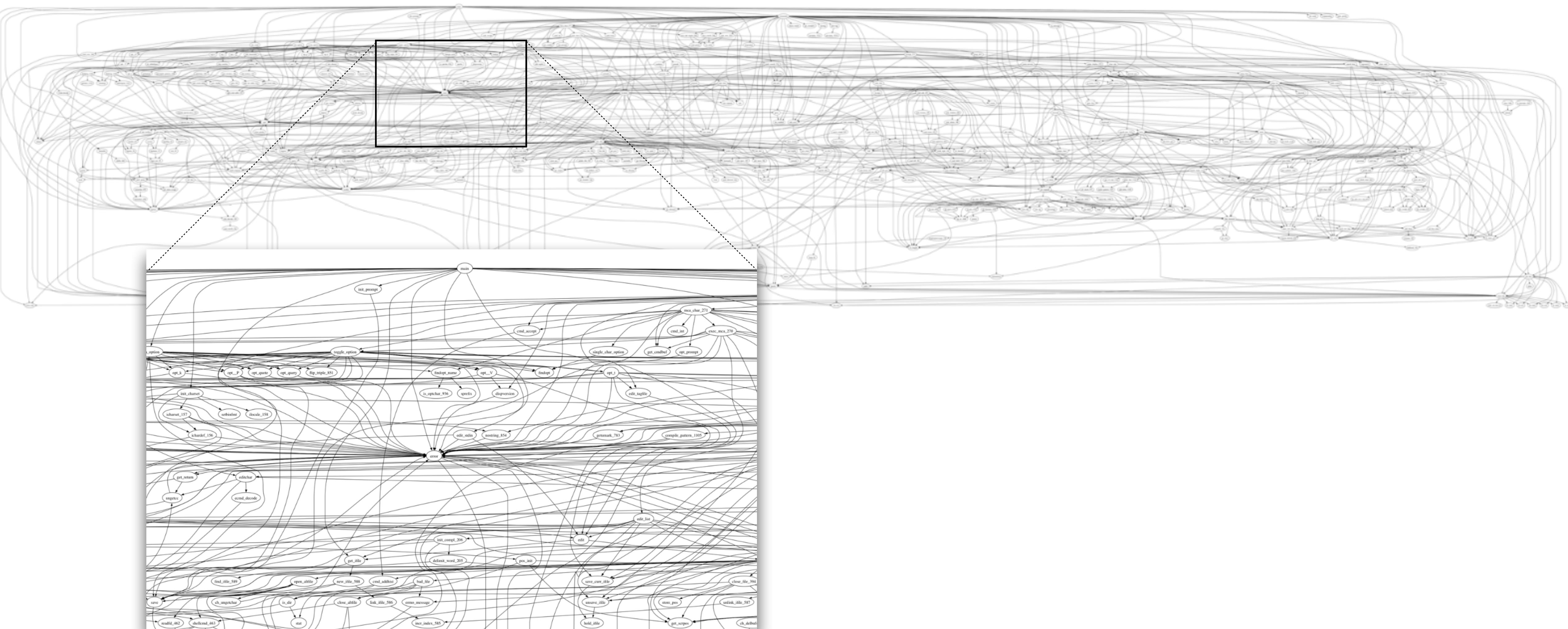
X



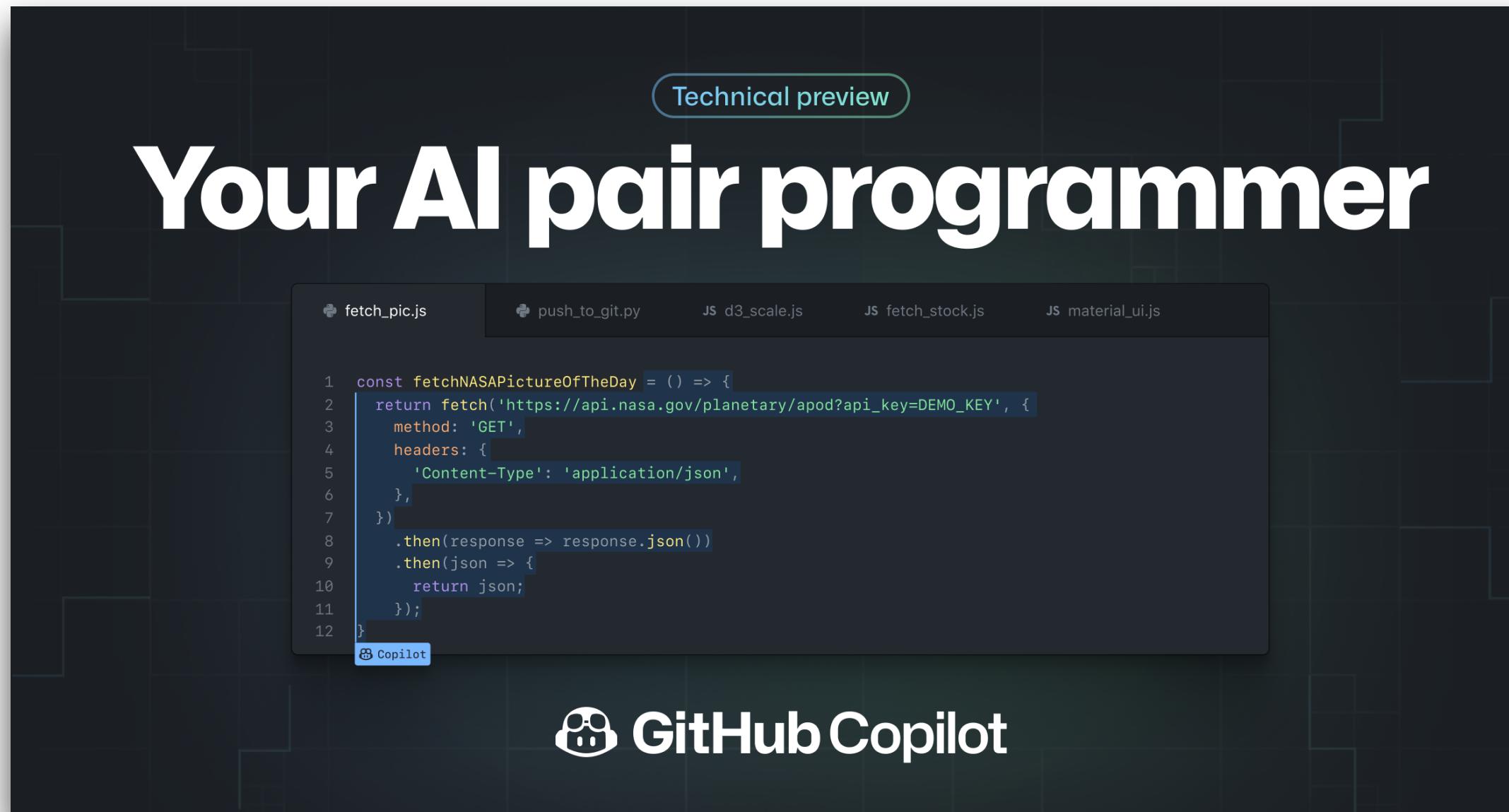
**20M+ New Developers
85M+ New Repositories
227M+ New Pull Requests
in 2022**

Software Complexity

less-382 (23,822 LOC)



The Era of AI



A screenshot of the GitHub Copilot interface showing a Python script named "parse_expenses.py". The code defines a function to parse expense strings into triples (date, value, currency). It uses the datetime module and splits lines by '#'. The "Copilot" button is at the bottom right.

```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12     expenses = []
13     for line in expenses_string.splitlines():
14         if line.startswith("#"):
15             continue
16         date, value, currency = line.split(" ")
17         expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                           float(value),
19                           currency))
20
return expenses
```

The word "Copilot" is visible at the bottom right of the code editor.

Artificial Unintelligence (인공저능)

```
 시작하기 × C+ int check_buffer_overflow() { Untitled-2 ●  
 1 int check_buffer_overflow() {  
 2     int buffer[5];  
     buffer[6] = 0;  
     return 0;  
 3 }
```

```
1 // read an image file  
2 int read_file() {  
3     FILE *fp;  
     int i, j, k;  
     int n;  
     int m;  
     int nc;  
     int nr;  
     int nb;  
     int nc_max;  
     int nr_max;  
     int nb_max;  
     int nc_min;  
     int nr_min;  
     int nb_min;  
     int nc_avg;  
     int nr_avg;  
     int nb_avg;  
     int nc_sum;  
     int nr_sum;  
     int nb_sum;  
     int nc_var;  
     int nr_var;  
     int nb_var;  
     int nc_std;  
     int nr_std;  
     int nb_std;  
     int nc_med;  
     int nr_med;  
     int nb_med;  
     int nc_mode;  
     int nr_mode;  
     int nb_mode;  
     int nc_min_index;  
     int nr_min_index;  
     int nb_min_index;  
     int nc_max_index;  
     int nr_max_index;  
     int nb_max_index;  
     int nc_med_index;  
     int nr_med_index;  
     int nb_med_index;  
     int nc_mode_index;  
     int nr_mode_index;  
     int nb_mode_index;  
     int nc_sum_index;  
     int nr_sum_index;  
     int nb_sum_index;  
     int nc_var_index;  
     int nr_var_index;  
     int nb_var_index;  
     int nc_std_index;  
     int nr_std_index;  
     int nb_std_index;  
     int nc_avg_index;  
     int nr_avg_index;  
     int nb_avg_index;  
     int nc_min_index_index;  
     int nr_min_index_index
```

Why?



SW Bug by Human Developer (1)

- Example: gimp-2.6.7 (CVE-2009-1570)

```
long ToL (char *pbuffer) { return (puffer[0] | puffer[1]<<8 | puffer[2]<<16 | puffer[3]<<24); }

short ToS (char *pbuffer) { return ((short)(puffer[0] | puffer[1]<<8)); }

gint32 ReadBMP (gchar *name, GError **error) {
    if (fread(buffer, Bitmap_File_Head.biSize - 4, fd) != 0)
        FATALP ("BMP: Error reading BMP file header #3");
    Bitmap_Head.biWidth = ToL (&buffer[0x00]); 1
    Bitmap_Head.biBitCnt = ToS (&buffer[0x0A]); 2

    rowbytes = ((Bitmap_Head.biWidth * Bitmap_Head.biBitCnt - 1) / 32) * 4 + 4; 4
    image_ID = ReadImage (rowbytes); 5
    ...
}

gint32 ReadImage (int rowbytes) {
    buffer = malloc(rowbytes); 6 // malloc with overflowed size
    ...
}
```

SW Bug by Human Developer (2)

- Example: sam2p-0.49.4 (CVE-2017-1663)

```
long ToL (char *pbuffer) { return (puffer[0] | puffer[1]<<8 | puffer[2]<<16 | puffer[3]<<24); } 3
short ToS (char *pbuffer) { return ((short)(puffer[0] | puffer[1]<<8)); }
bitmap_type bmp_load_image (FILE* filename) {
    if (fread(buffer, Bitmap_File_Head.biSize - 4, fd) != 0)
        FATALP ("BMP: Error reading BMP file header #3");
    Bitmap_Head.biWidth = ToL (&buffer[0x00]); 1
    Bitmap_Head.biBitCnt = ToS (&buffer[0x0A]); 2
    rowbytes = ((Bitmap_Head.biWidth * Bitmap_Head.biBitCnt - 1) / 32) * 4 + 4; 4
    image.bitmap = ReadImage (rowbytes); 5
    ...
}

unsigned char* ReadImage (int rowbytes) {
    unsigned char *buffer = (unsigned char*) new char[rowbytes]; // malloc with overflowed size 6
    ...
}
```

SW Bug by AI Developer

- Example (feat. Copilot):

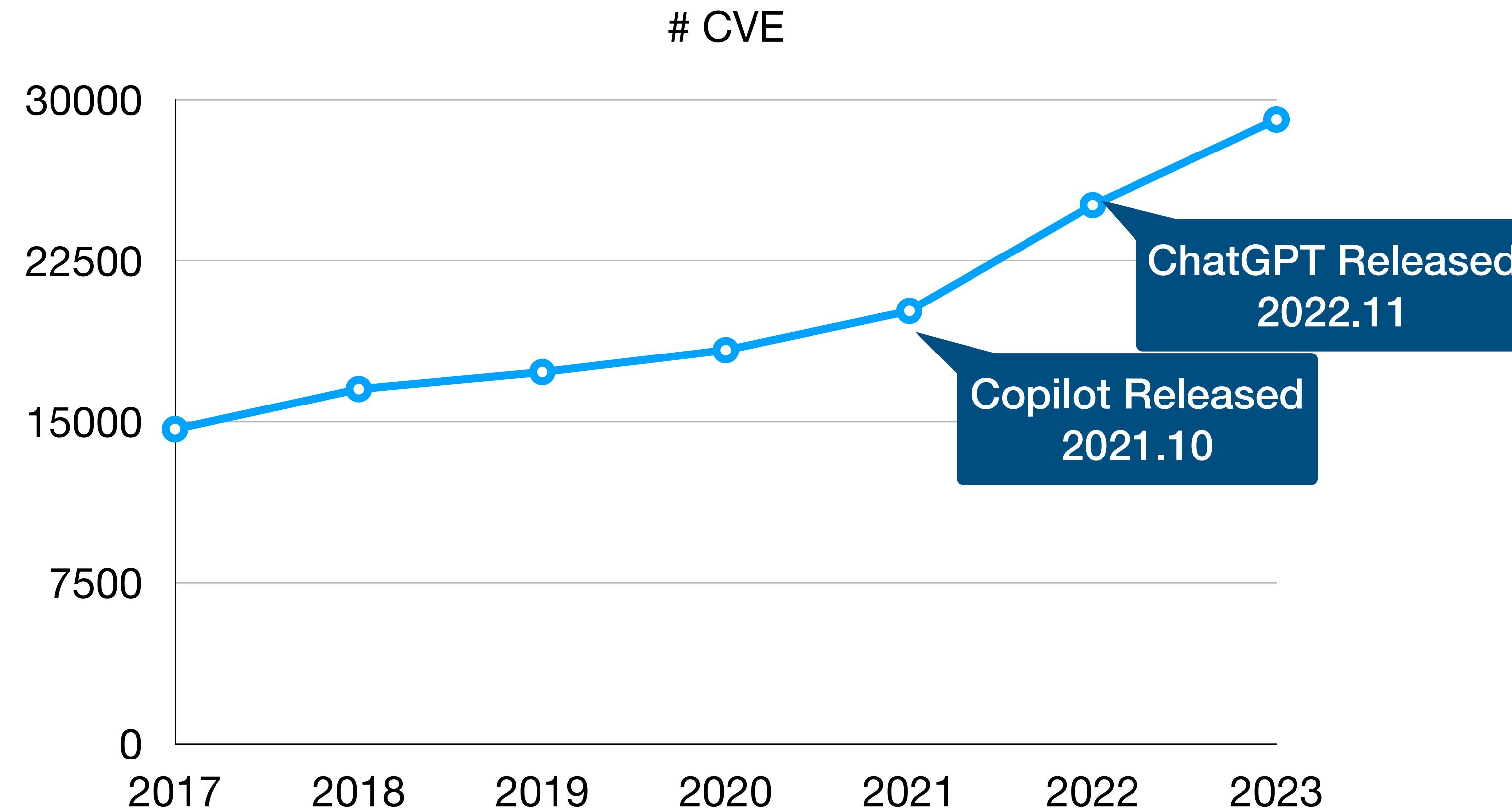
```
int toLong(char *buffer) {
    return (buffer[0]) | (buffer[1] << 8) | (buffer[2] << 16) | (buffer[3] << 24);
}

int f(char *name) {
    int width, height, area;
    char buffer[10];
    FILE *fd = fopen(name, "rb");
    fread(buffer, 10, 1, fd);
    fclose(fd);

    // Copilot, fill in the blank!
    width = toLong(buffer + 18);
    height = toLong(buffer + 22);
    area = width * height;

```

Aftermath?



Course Objectives: Principles

Q: How to **formally estimate** software behavior automatically before its execution?



Course Objectives: Principles

Artifact

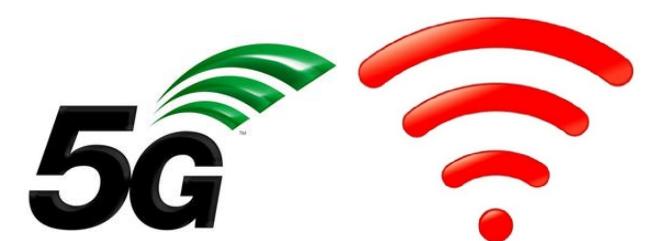


Subject



Principle

$$\vec{F} = m\vec{a}$$



Static Program Analysis

- General methodology to predict software behavior
 - **static**: before execution
 - **automatic**: software is analyzed by software (program analyzer)
 - **systematic**: foundational theory (Abstract Interpretation)
- Applications:
 - bug-finding, verification, code optimization, etc



Success Stories

Domain-specific
Verification



Windows Device Driver
Microsoft

Astrée

Airbus Controller
ENS / AbsInt



Stanford / Synopsys



Facebook



SNU / Fasoo.com



Mathworks



GrammaTech



Semmle / Github



JuliaSoft



GCC



LLVM/Clang

Course Objectives: Practice & Challenge

- Homework: **design & implement** program analyzers
 - 7 (main) + 2 (dummy) assignments
- Programming assignments in OCaml using LLVM & Z3
 - You will write your analyzers in OCaml
 - Your analyzer will analyze LLVM IR code
 - You will utilize Z3 in your analyzers
- Why LLVM? (<https://llvm.org>)
- Why OCaml? (<https://ocaml.org>)
- Why Z3? (<https://github.com/Z3Prover/z3>)



The LLVM Compiler Infrastructure

- The de-facto standard & well-structured compiler toolchain
 - parser, code optimizer, linker, loader, debugger, etc
- A wide variety of frontends: C/C++, Obj-C, Swift, Fortran, etc
 - translated to the LLVM IR (intermediate representation)



Apple's other open secret: the LLVM Compiler

By Prince McLean
Friday, June 20, 2008, 04:10 am PT (07:10 am ET)

SproutCore, profiled earlier this week, isn't the only big news spill out from the top secret WWDC conference due to Apple's embrace of open source sharing. Another future technology featured by the Mac maker last week was LLVM, the Low Level Virtual Machine compiler infrastructure project.

Like SproutCore, LLVM is neither new nor secret, but both have been hiding from attention due to a thick layer of complexity that has obscured their future potential.

Looking for LLVM at WWDC

Again, the trail of breadcrumbs for LLVM starts in the public WWDC schedule. On Tuesday, the session "New Compiler Technology and Future Directions" detailed the following synopsis:

Google Chrome is replacing Microsoft's C++ compiler with Clang

By Muhammad Jarir Kanji · Mar 6, 2018 14:06 EST · HOT!

Alongside bringing better touch support and automatic ad-blocking for 'intrusive' ads to the desktop version of Chrome, Google is also making some changes to its browser under the hood. The company is now starting to build Chrome for Windows using the Clang compiler which it already uses for other platforms like macOS and Linux.

IBM Developer

Power developer portal Blogs Feedback

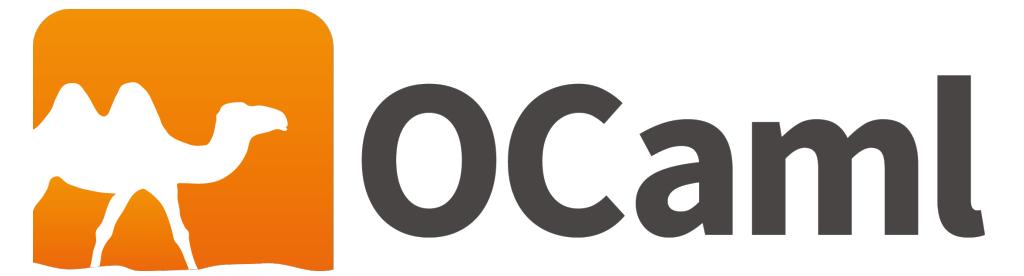
Announcements Compilers

IBM C/C++ and Fortran compilers to adopt LLVM open source infrastructure

SiyuanZhang
Published on February 23, 2020 / Updated on February 26, 2020

The OCaml Language

- Simple, safe, and realistic programming language
 - Strong type system, higher-order functions, etc
- Official OCaml bindings to LLVM and Z3 API supported
- A lot of growing demands from academia and industry
- See the materials on the course webpage



The Z3 Theorem Prover



- State-of-the-art automated theorem prover by Microsoft Research
- Solving satisfiability modulo theory (SMT) problems
 - first-order logic with background theories
(e.g., arithmetic, bit-vectors, arrays, datatypes, uninterpreted functions, etc)

Boolean Satisfiability Problem (SAT)

$$(\neg A \vee B) \wedge (\neg B \vee C) \wedge (A \vee \neg C \vee B)$$

Satisfiable when

$A = \text{false}$

$B = \text{true}$

$C = \text{true}$

Satisfiability modulo theory (SMT)

$$x + 2 = y \implies f \text{read } a, x, 3, y - 2 = f(y - x + 1)$$

Arithmetic

Array

Uninterpreted Functions

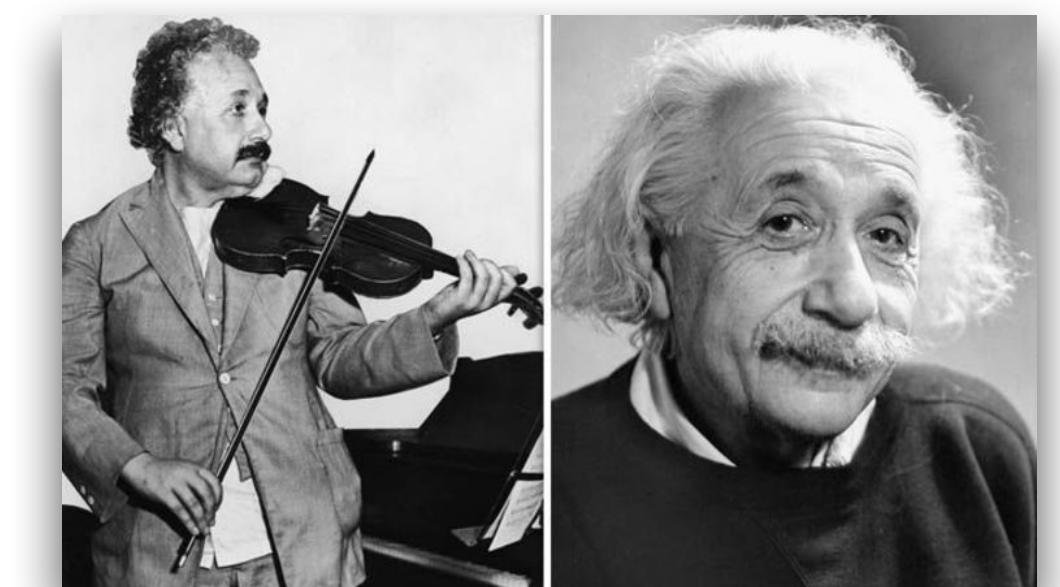
Course Objectives: Active Learning

- The era of AI: NO free lunch and requires more creative thinking
 - Naively understanding & memorizing knowledge becomes obsolete
 - Creativity comes from “pain” (창의력은 노가다에서 나온다)
 - Training GPT-3: 355 yrs on a single GPU (2020)
- Three levels of knowledge

- Understandable knowledge (이해 가능)
- Explainable knowledge (설명 가능)
- Artifiable knowledge (예술로 승화 가능)

by lecture, exam,
programming assignment

by various activities



“The greatest scientists are
artists as well”
- A. Einstein

Activity (1): Writing

- Science is communication
 - Writing: main communication tool
- Roles of KAIST folks
 - World-leading researcher (in English)
 - Inspiring communicator to the community (in Korean)
 - For newcomers, public officials, etc
- Idiot's transition algorithm

```
String.split_on_char ' ' word
|> List.map translate_word
|> String.concat ""
```

← 트윗

국립국어원 @urimal365
[우리말 다듬기] '변이 지방'은 '트랜스 지방'을 다듬은 말입니다.
"이 과자에는 변이 지방이 들어 있지 않다."처럼 다듬은 말 '변이
지방'을 쓸 수 있습니다. #다듬기

오전 9:01 · 2013년 3월 19일 · twtkr

44 리트윗 9 인용한 트윗 6 마음에 들어요

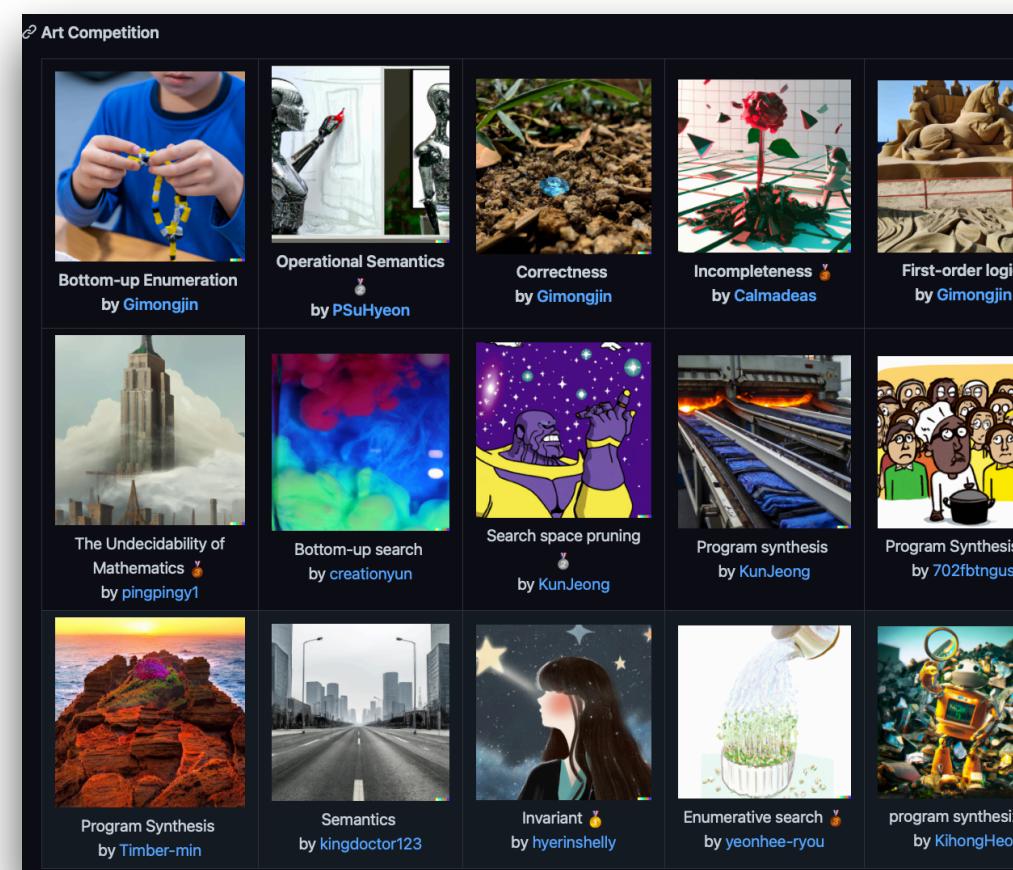
田舎暮らし아쿠아; 여행가고싶다 @hanaosakas · 2013년 3월 19일
@urimal365 님에게 보내는 답글
@urimal365 트랜스 지방의 트랜스는 "변이"라는 의미가 아닙니다.. trans-/cis-로 구
조체의 형태가 다른 것입니다... 굳이 번역할 필요도 다듬을 필요도 없는 과학용어입니
다.. 과학과 관련된 말을 다듬을 때는 과학계에 자문을 구해보심이

국립국어원 @urimal365 · 2013년 3월 19일
@hanaosakas 건의하실 의견은 공식적인 민원(j.mp/zUyQoR)을 통하여 주시거나,
'우리말 다듬기(malteo.net)'를 이용해 주시기 바랍니다. #순화

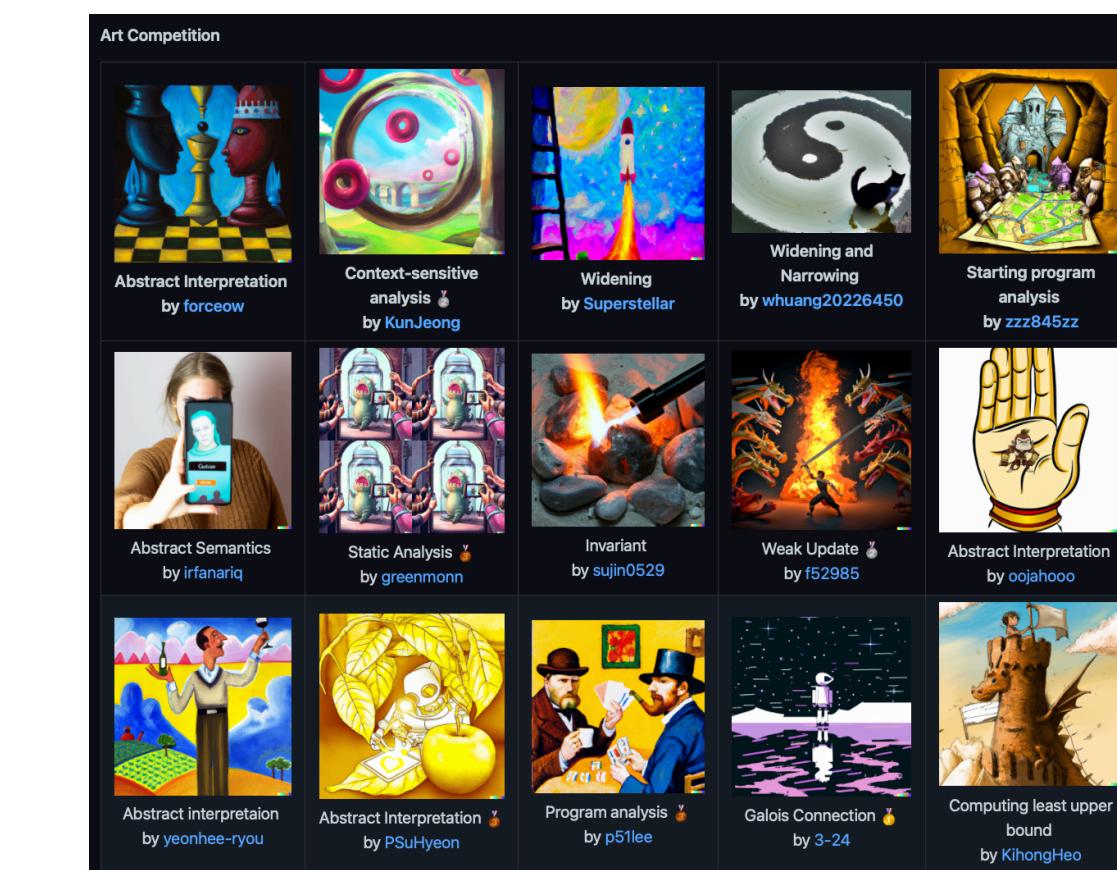
PYLON / 수정탑 @Pylon_kr · 2013년 3월 19일
@urimal365 님에게 보내는 답글
@urimal365 trans가 변이면 cis는 뭐라고 해야하나요? 유기화학 책 좀 보고 오세요.

Activity (2): Art Competition (천하제일 그림대회)

- What does X mean to you?
- Pick a concept X you have learned in this class such as cryptography, program analysis
- Draw a picture that succinctly represents the concept using DALL-E (or similar tools)
- Detailed instructions will be announced later



CS492: Program Reasoning, 2022



CS524: Program Analysis, 2022

Homework

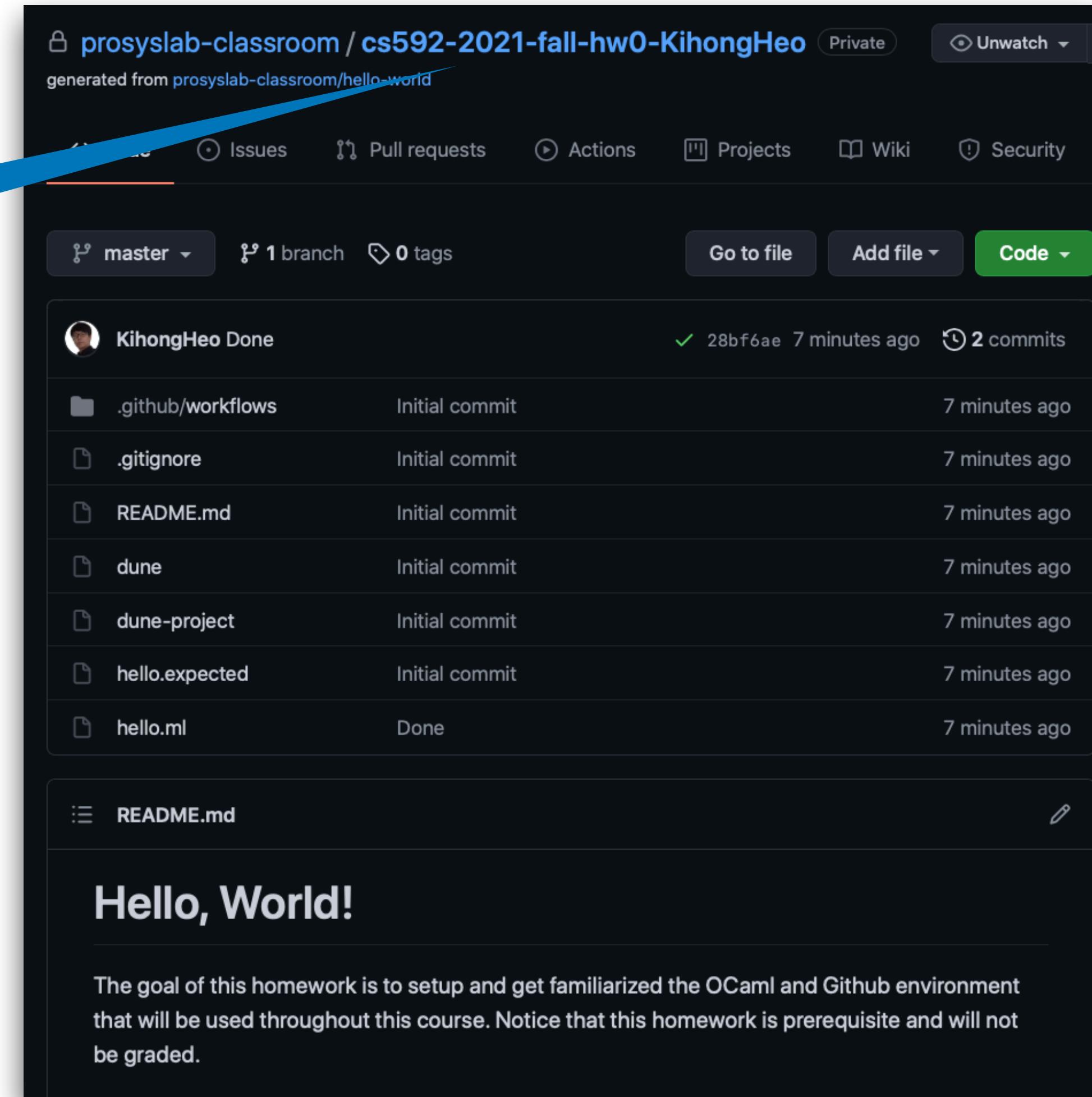
- All submissions will be managed using Github / Github Classroom
 1. For each HW, a unique invitation URL will be posted on the issue board
 2. Once you accept, a private repo for your assignment will be created
 3. You can push as many commits as you want before the deadline
 4. The final commit of your main branch will be graded
- Gradescope will be used for written assignments and exams
- 80% credit for 1-day late, 50% credit for 2-days late, NO credit otherwise

Homework 0.1: Hello, World!

- Goal: setting up and getting familiarized with OCaml and Git environments
 - Implement your “hello-world” program in OCaml
 - Test on your machine
 - Push to your Github repository
 - See the result in Github Action
- The invitation URL will be posted on the course webpage
- Will not be graded

Homework 0.1: Hello, World!

1. Accept the invitation
and have your repository



Homework 0.1: Hello, World!

A screenshot of a GitHub repository page for "prosylab-classroom / cs592-2021-fall-hw0-KihongHeo". The repository is private and was generated from "prosylab-classroom/hello-world". The "Code" tab is selected, showing the master branch with 1 branch and 0 tags. A commit by "KihongHeo Done" was made 7 minutes ago, containing 2 commits. The commit details show files like .github/workflows, .gitignore, README.md, dune, dune-project, hello.expected, and hello.ml. The README.md file contains the text "Hello, World!". Below the README, a note states: "The goal of this homework is to setup and get familiarized the OCaml and Github environment that will be used throughout this course. Notice that this homework is prerequisite and will not be graded."

2. Commit your code

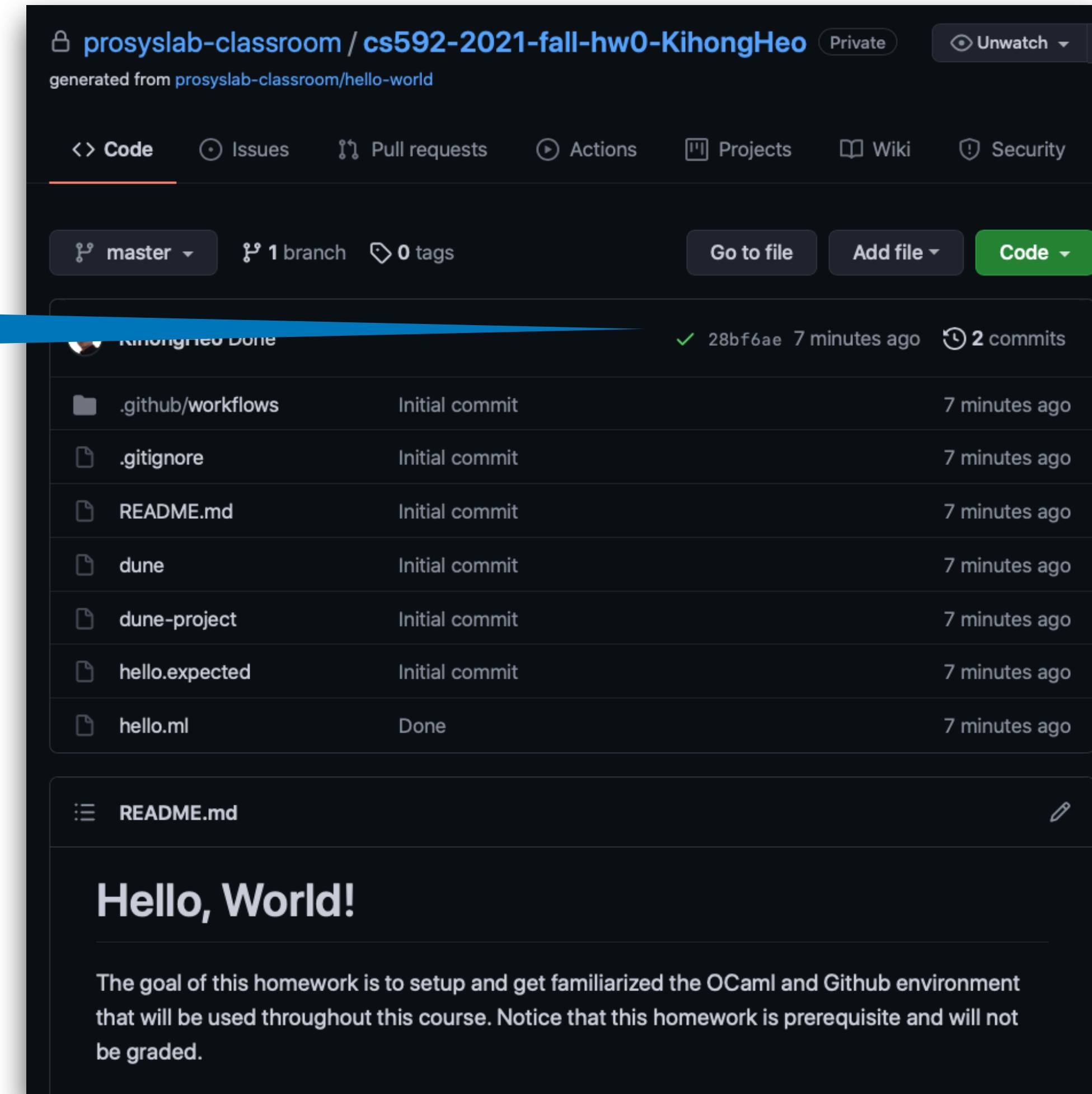
File	Type	Commit Message	Time
.github/workflows		Initial commit	7 minutes ago
.gitignore		Initial commit	7 minutes ago
README.md		Initial commit	7 minutes ago
dune		Initial commit	7 minutes ago
dune-project		Initial commit	7 minutes ago
hello.expected		Initial commit	7 minutes ago
hello.ml		Done	7 minutes ago

Hello, World!

The goal of this homework is to setup and get familiarized the OCaml and Github environment that will be used throughout this course. Notice that this homework is prerequisite and will not be graded.

Homework 0.1: Hello, World!

3. See your result



Homework 0.2: OCaml Programming

- Goal: getting familiarized with basic OCaml programming
 - Solve a few programming problems in OCaml
 - Test on your machine
 - Push to your Github repository
 - See the result in Github Action
- The invitation URL will be posted on the course webpage
- Will not be graded but highly recommended if you are not familiar with OCaml

Requirement: Software Engineering Practices

- All programming assignments submissions must follow basic SE practices

- Remove all compile errors and warning

- Remember: warning is error

- Clean code via formatting

- Acceptable code coverage

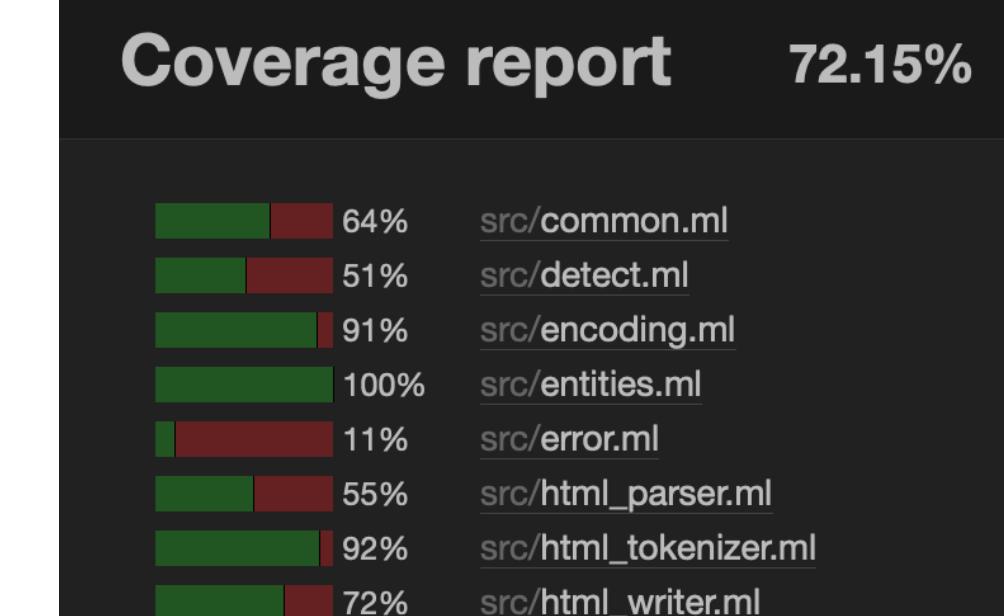
- Otherwise, NO points will be given

- Detailed instructions will be provided

```
File "src/semantics.ml", line 39, characters 6-11:  
39 |   let count = 1 in  
      ^^^^  
Warning 26 [unused-var]: unused variable count.
```

```
let rec  
  fact n =  
    match n with  
    | 0 | 1 -> 1  
    | _ -> n *  
           fact (n - 1)
```

```
let rec fact n =  
  match n with  
  | 0 | 1 -> 1  
  | _ -> n * fact (n - 1)
```



```
14 module Ns :  
15 sig  
16   val to_string : ns -> string  
17 end =  
18 struct  
19   let to_string = function  
20   | HTML -> html_ns  
21   | MathML -> mathml_ns  
22   | SVG -> svg_ns  
23   | Other s -> s  
24 end
```

Homework 1: Essay

- Read the stories of Google, Facebook, and Apple (available from the course webpage)
- Write a critique essay
- Syntactic requirements
 - Typeset your document in Latex using a provided template (maximum 2 pages)
 - Korean (if you are a native Korean speaker), English (otherwise)
- Semantic requirements: Top-down (두괄식)
 - For each paragraph, write the topic sentence first followed by the details.
- See the README for the details
- The invitation URL will be posted on the course webpage

Misc (1): Github, Gradescope

- Submit your Github account via the google form (see the Github issue board)
- Join Gradescope (see the Github issue board)
 - For written assignments and the final exam
- Rules for programming assignments
 - Preserve the structures (directories, files, types, etc)
 - Don't install further Github App

Misc (2): Email Communication

- Primary communication tool for business
 - NOT for instant messages like Kakao Talk
 - But try to respond as soon as possible (within 24hrs)
- Include all relevant persons by using “CC” (참조) and “Reply All” (전체답장)
- International communication: you can send emails anytime
 - Professionals do not use notifications for emails
 - Do you send it to the US at 3 am?
- Other details: Google “email etiquette”