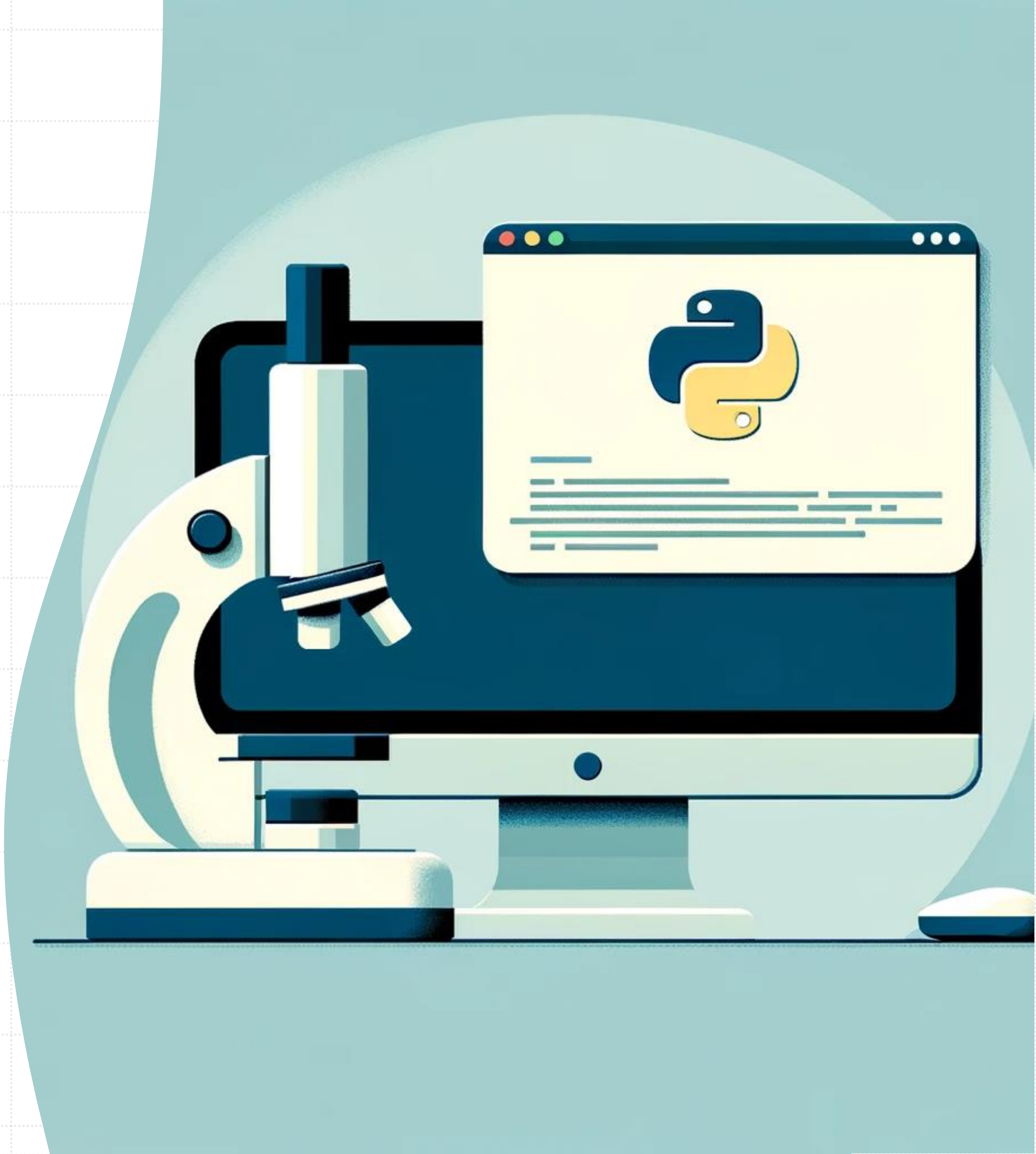
Segment Anything in Python for Microscopy Data

Ranit Karmakar

Image Analysis Collaboratory

Harvard Medical School

October 25, 2024



Who are we?

We collaborate with biologists at **HMS** to extract useful information from their data quantitively.

Image Analysis Collaboratory

https://iac.hms.harvard.edu

Simon Nørrelykke

Director

Antoine Ruzette

Associate

Federico Gasparoli

Research Associate

Maria Theiss

Specialist Postdoc

Ranit Karmakar

Specialist Postdoc

Housekeeping

We code together

Feel free to stop me any time if I am going too fast or too slow

Ask questions

There's no such thing as a dumb question*. Please ask any question you have

Objective

“is the process of partitioning a digital image into multiple image segments”

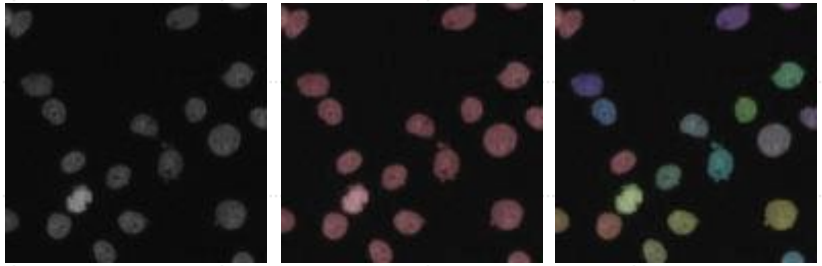
Generalizable (Segment cats, dogs, cars as well as cells, nuclei, tubules etc.)

Segment Anything in Python

for Microscopy Data

Before we start

Semantic vs Instance segmentation



Input image

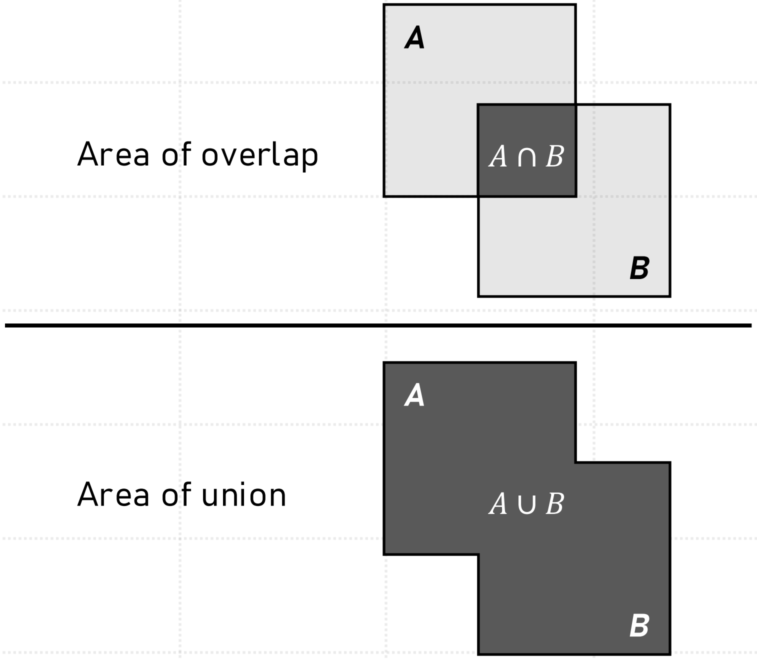
Semantic segmentation

Instance segmentation

Cells vs not-cell → semantic segmentation

Cell 1, Cell 2, Cell 3, … → Instance segmentation

Intersection over Union (IoU)



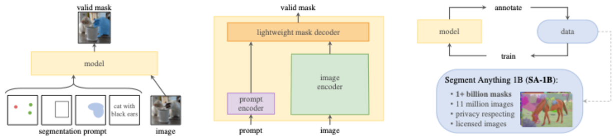
Chapters

Chapter 1: Introduction to SAM

Vision Transformer (ViT) based image segmentation model trained on 1.1 million images and over 1 billion masks to perform zero-shot segmentation

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a promptable segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Abstract

We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 ***billion*** *masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive – often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and 11M images at https://segment-anything.com to foster research into foundation models for computer vision.*

1. Introduction

Large language models pre-trained on web-scale datasets are revolutionizing NLP with strong zero-shot and few-shot generalization [10]. These "foundation models" [8] can generalize to tasks and data distributions beyond those seen during training. This capability is often implemented with *prompt engineering* in which hand-crafted text is used to prompt the language model to generate a valid textual response for the task at hand. When scaled and trained with abundant text corpora from the web, these models' zero and few-shot performance compares surprisingly well to (even matching in some cases) fine-tuned models [10, 21]. Empirical trends show this behavior improving with model scale, dataset size, and total training compute [56, 10, 21, 51].

Foundation models have also been explored in computer vision, albeit to a lesser extent. Perhaps the most prominent illustration aligns paired text and images from the web. For example, CLIP [82] and ALIGN [55] use contrastive learning to train text and image encoders that align the two modalities. Once trained, engineered text prompts enable zero-shot generalization to novel visual concepts and data distributions. Such encoders also compose effectively with other modules to enable downstream tasks, such as image generation (*e.g.*, DALL·E [83]). While much progress has been made on vision and language encoders, computer vision includes a wide range of problems beyond this scope, and for many of these, abundant training data does not exist.

In this work, our goal is to build *a foundation model for image segmentation*. That is, we seek to develop a promptable model and pre-train it on a broad dataset using a task that enables powerful generalization. With this model, we aim to solve a range of downstream segmentation problems on new data distributions using prompt engineering.

The success of this plan hinges on three components: **task**, **model**, and **data**. To develop them, we address the following questions about image segmentation:

1. What **task** will enable zero-shot generalization?
2. What is the corresponding **model** architecture?
3. What **data** can power this task and model?

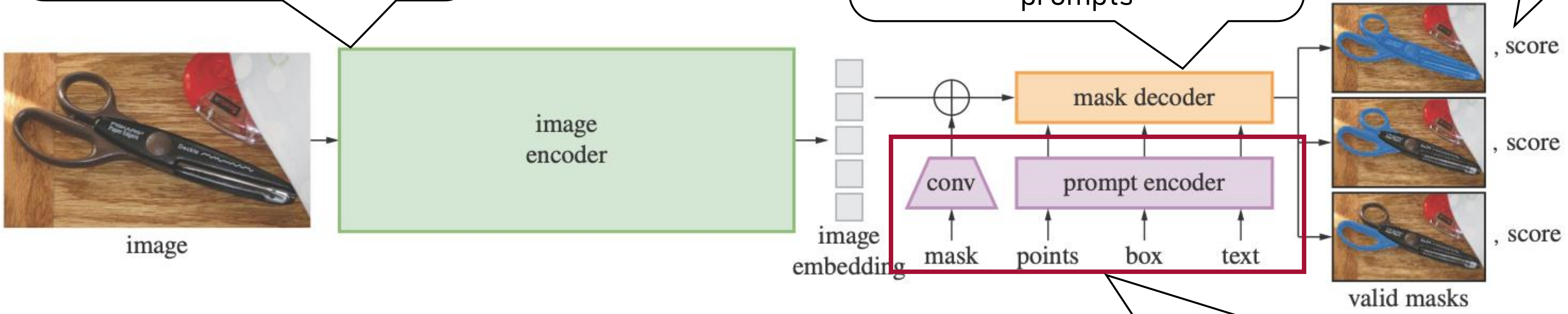
1

Chapter 1: Introduction to SAM

Masked Auto Encoder (MAE)
Pre-trained ViT that extracts image features creates image embedding

Mask Decoder
Generates the mask based on image embedding and prompts

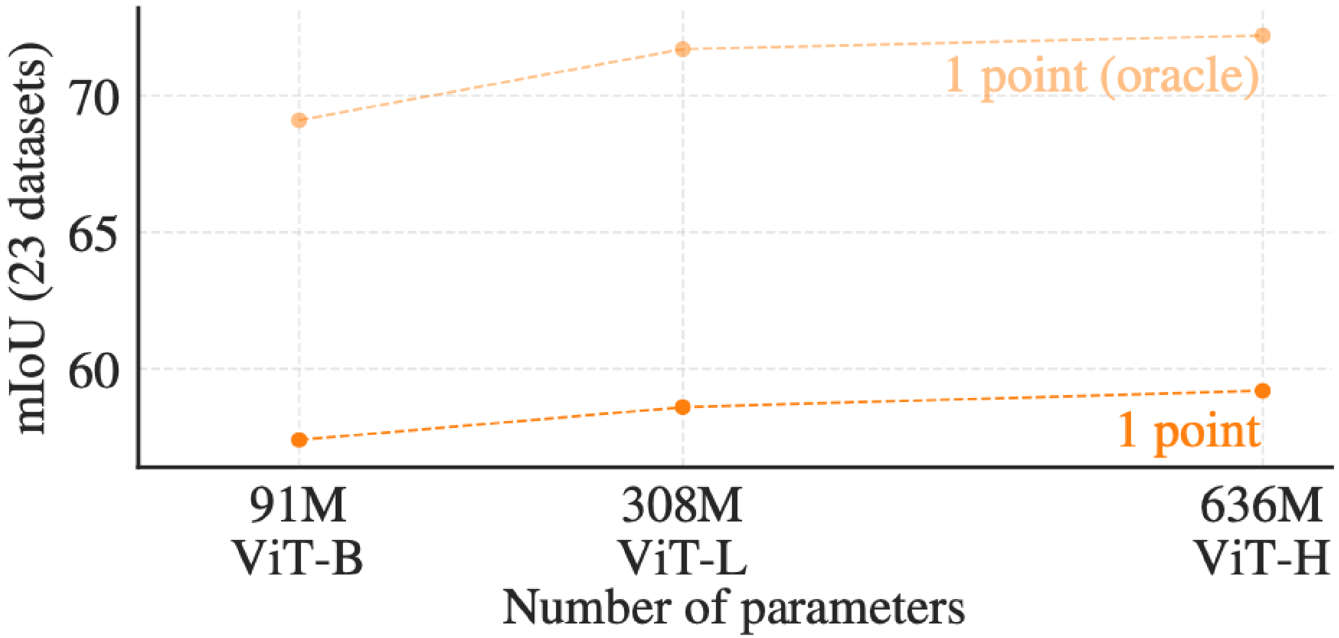
Produces multiple masks and segmentation score



Prompts
To indicate the model the region or object of interest

Chapter 1: Introduction to SAM

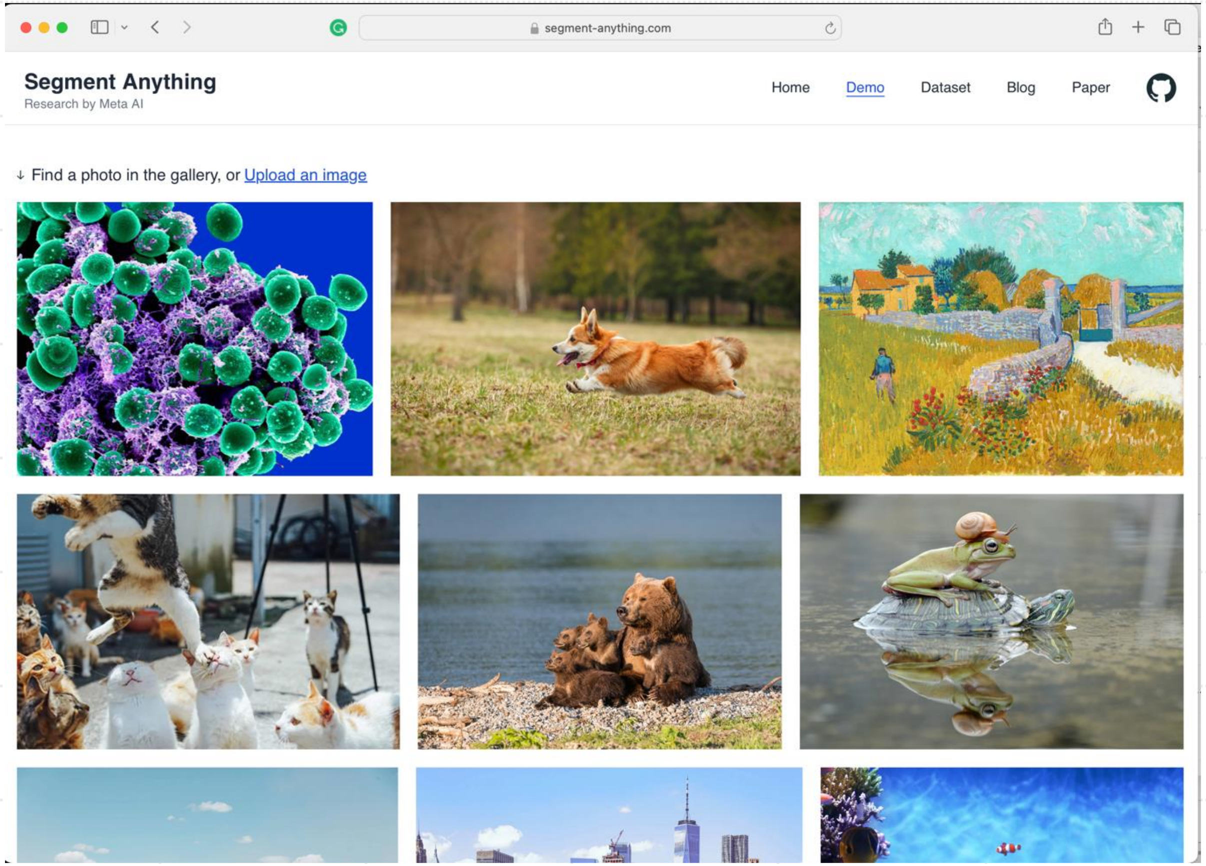
There are three different model checkpoints that can be used



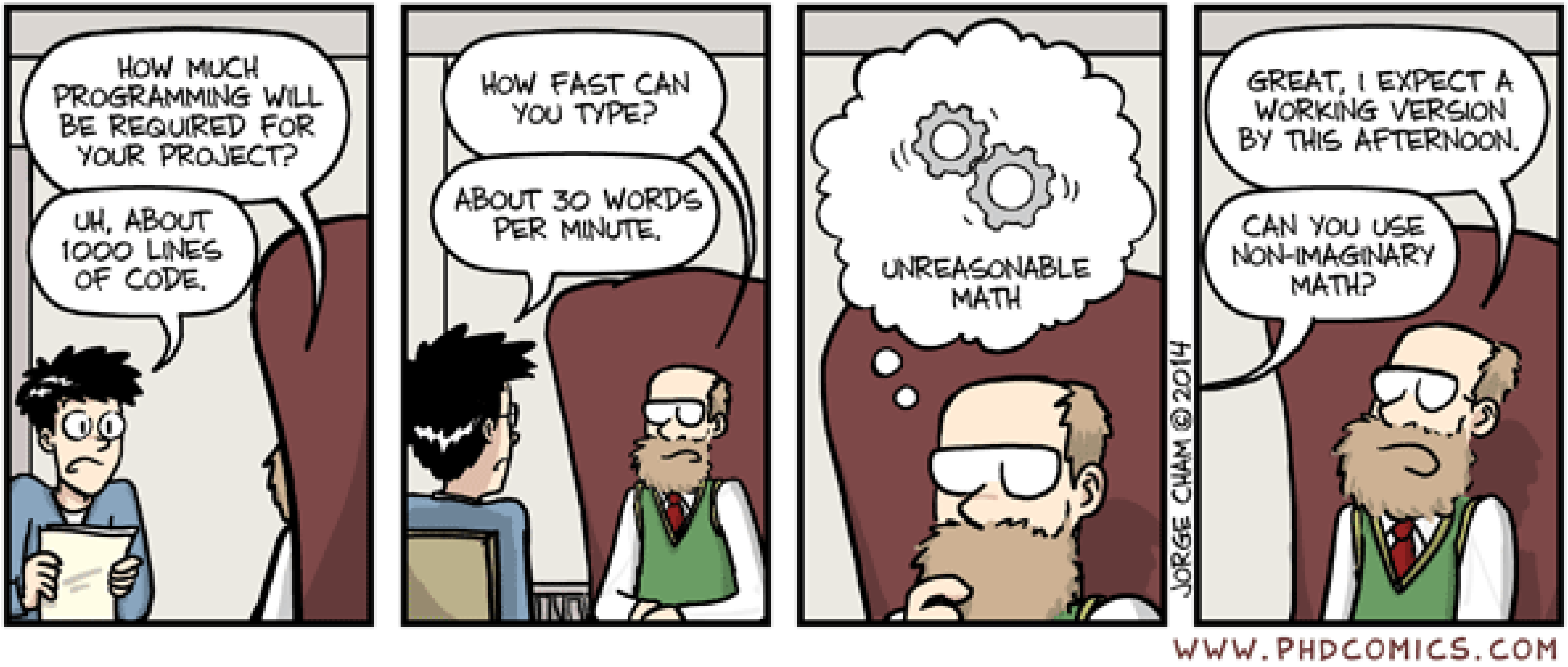
Chapter 1: Introduction to SAM

SAM on the web

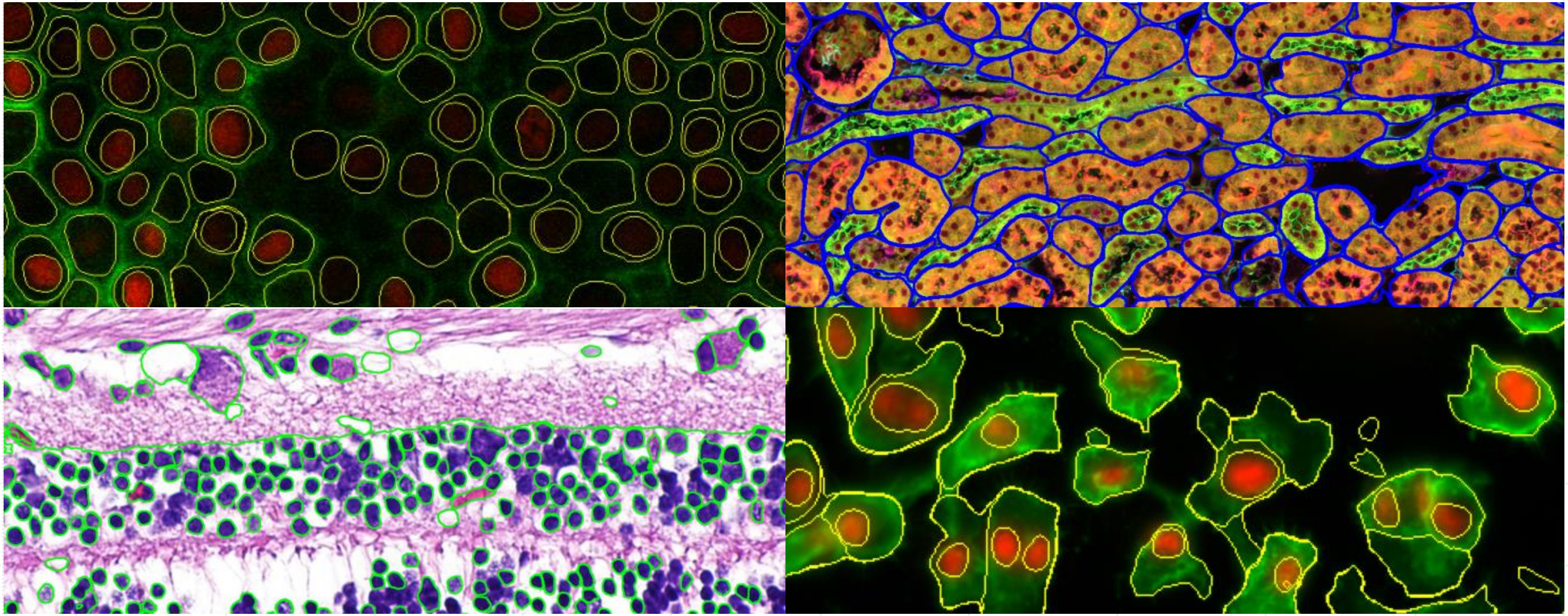
http://segment-anything.com/



Chapter 2: Setting up SAM on Google Colab

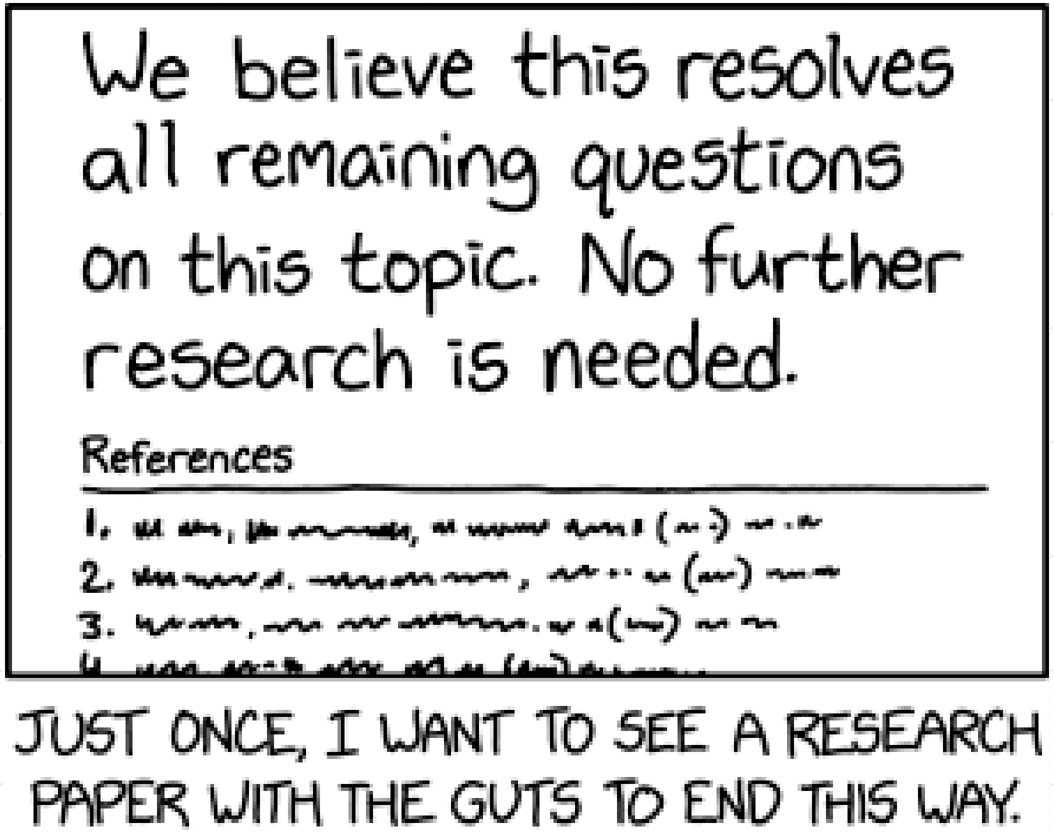


Chapter 5: Few other examples



Chapter 6: Conclusion

- SAM is a very powerful tool but it has its limitations
- For microscopy images as well, SAM performs well right out of the box
- Pre and post processing will almost always improve the results



Thank you!

in/rkarmaka/
rkarmaka/
ranit@hms.Harvard.edu