НГТУ
НЭТИ

Processing of High-speed Imaging Data with Galaxy and cwltool

Ilya Baranov, Maxim Gorodnichev, Sergey Starinskiy, Nikolay Miskiv, and Elena Starinskaya

Актуальность

В настоящее время ученые-экспериментаторы сталкиваются с проблемами обработки данных, полученных в ходе проведения экспериментов:

- Ручная обработка занимает много времени и не позволяет охватить весь объем данных
- Написание программы требует от ученых знания программирования и времени на разработку и отладку

Во многих задачах актуальной является проблема автоматизации обработки данных, поскольку её решение позволяет повысить эффективность исследований.

Использование WMS

В настоящее время для решения задачи автоматизации обработки данных используются различные WMS. На данный момент существует более 360 WMS: https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems (Existing Workflow systems)

Примеры систем:

- Galaxy
- CWL
- KNIME
- Nextflow
- Snakemake
- …

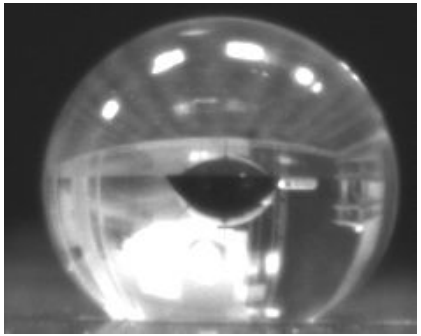
Прикладная задача

Одной из задач обработки, которую необходимо автоматизировать, является прикладная задача по получению характеристик падающих на поверхность капель. Её решением занимается ИТ СО РАН.

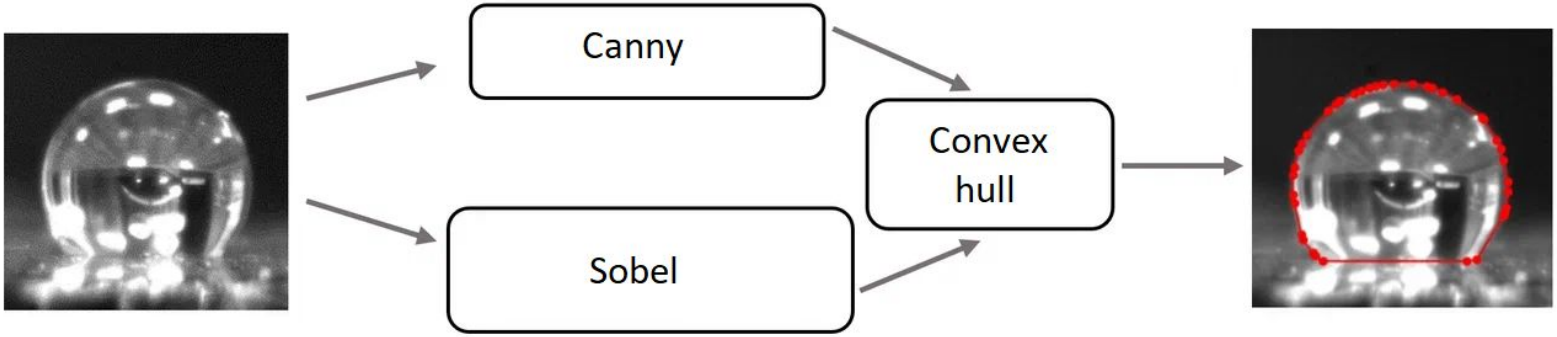
Особенности данных:

- Тысячи кадров для обработки
- Различные типы экспериментов
- Разное качество изображений




Предыдущие исследования



1. Назаров Н. А., Баранов И. Н., Миськив Н. Б., Старинская Е. М. "Методы определения геометрических параметров капель жидкости на основе анализа цифровых изображений" // Автометрия. Т.60, № 2. С. 30-40. DOI: 10.15372/AUT20240204.
2. Development of Methods for Processing of Experimental Data on High-speed Impact of Droplets on a Surface. Параллельные вычислительные технологии – XVIII всероссийская конференция с международным участием, ПаВТ'2024, г. Челябинск, 2–4 апреля 2024 г. Короткие статьи и описания плакатов. Челябинск: Издательский центр ЮУрГУ, 2024. — 196 с. — с. 7–16.

Цель и задачи

Цель работы – решить задачу обработки изображения капли с помощью системы Galaxy и cwltool, выполнить сравнительный анализ систем.

Задачи:

- Анализ системы Galaxy и cwltool
- Решение задачи обработки изображения с помощью Galaxy и cwltool
- Сравнение полученных решений
- Выявление преимуществ и недостатков каждой из систем

Galaxy

Galaxy project — это проект с открытым исходным кодом, который используется учеными по всему миру для обработки больших объемов данных. Изначально, платформа использовалась для хранения и обработки биомедицинских данных, но с развитием проекта количество предметных областей, в которых применяется Galaxy, значительно увеличилось.

Описание платформы Galaxy

В Galaxy работа происходит с рядом базовых сущностей:

- Инструменты (tools)
- Алгоритмы обработки (workflows)
- Истории (histoires)
- Визуализации (visualizations)

Эти сущности не привязаны к конкретной предметной области, позволяя создавать решения для произвольной задачи по обработки данных.

Решение прикладной задачи

В обучающем примере используются готовые инструменты, уже внедренные в Galaxy, но для решения других задач по обработки изображений их может быть недостаточно. Рассмотрим задачу поиска высоты капли по изображению. Прежде чем строить алгоритм обработки в Galaxy нужно определиться с шагами.

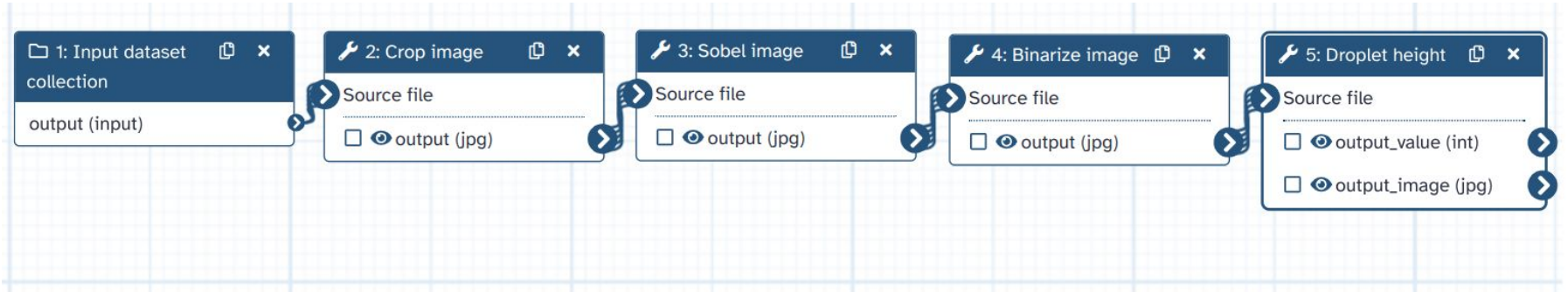
Опишем простейший вариант алгоритма для поиска высоты капли:

1. Обрезать изображение
2. Применить оператор Собеля
3. Применить бинаризацию
4. Вычислить высоту капли по бинарному изображению

Реализация решения задачи с помощью Galaxy

Для реализации решения задачи с помощью Galaxy были проделаны следующие шаги:

1. Реализованы инструменты для каждого шага алгоритма
2. Построен workflow для обработки эксперимента
3. workflow протестирован на тестовых данных
4. Реализован инструмент визуализации полученных данных

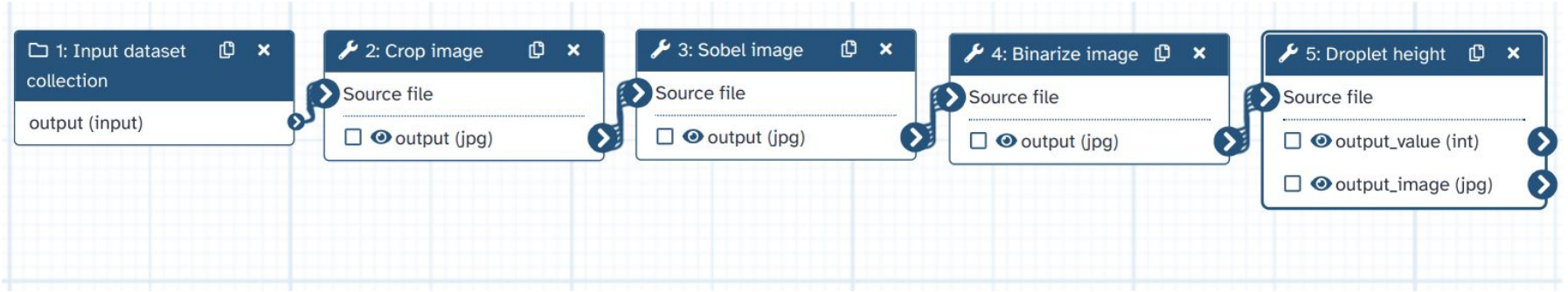


Отладка полученного алгоритма

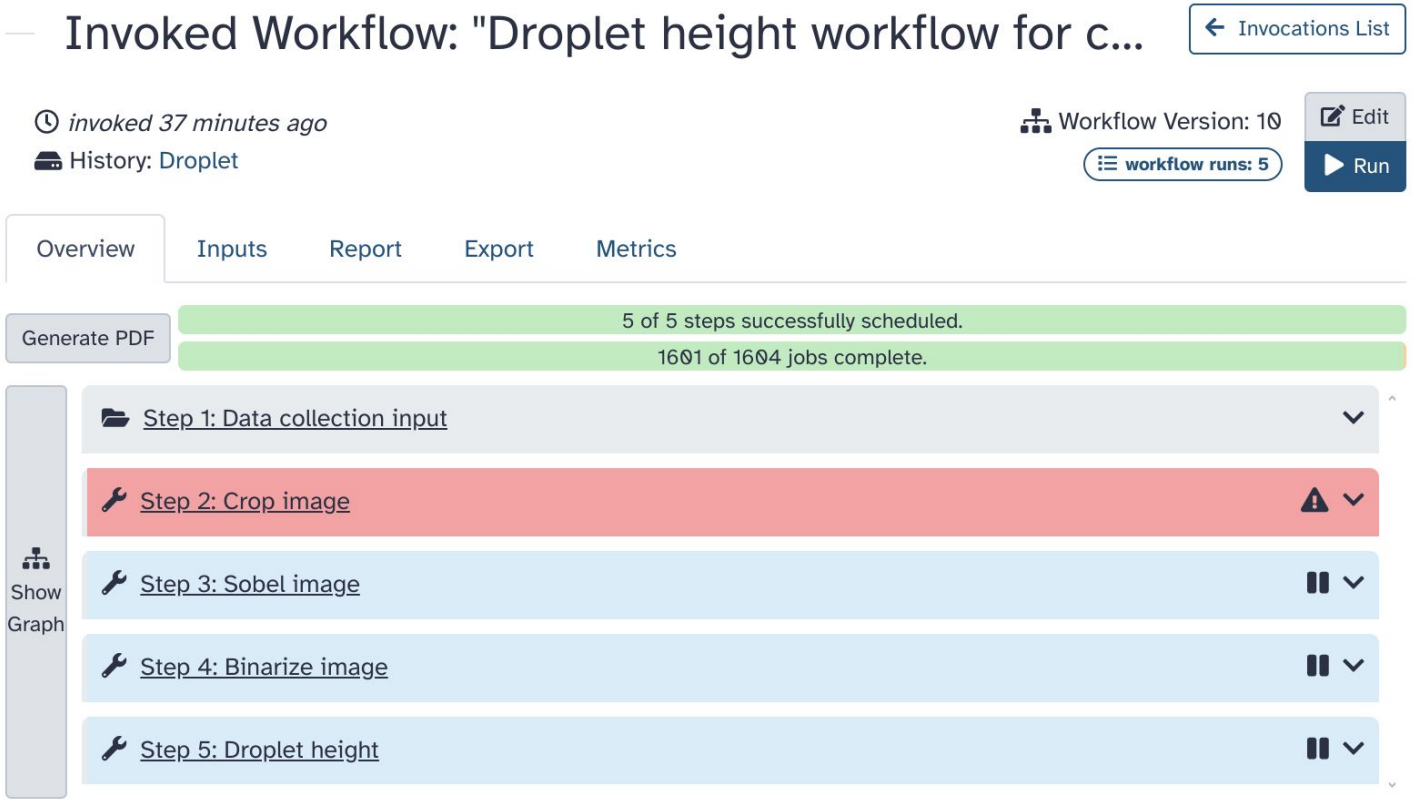
Протестируем алгоритм на тестовом изображении. Результаты работы каждого шага алгоритма обработки показаны на рисунках. Последний шаг помимо изображения возвращает значение высоты капли в пикселях.

Работа с коллекциями

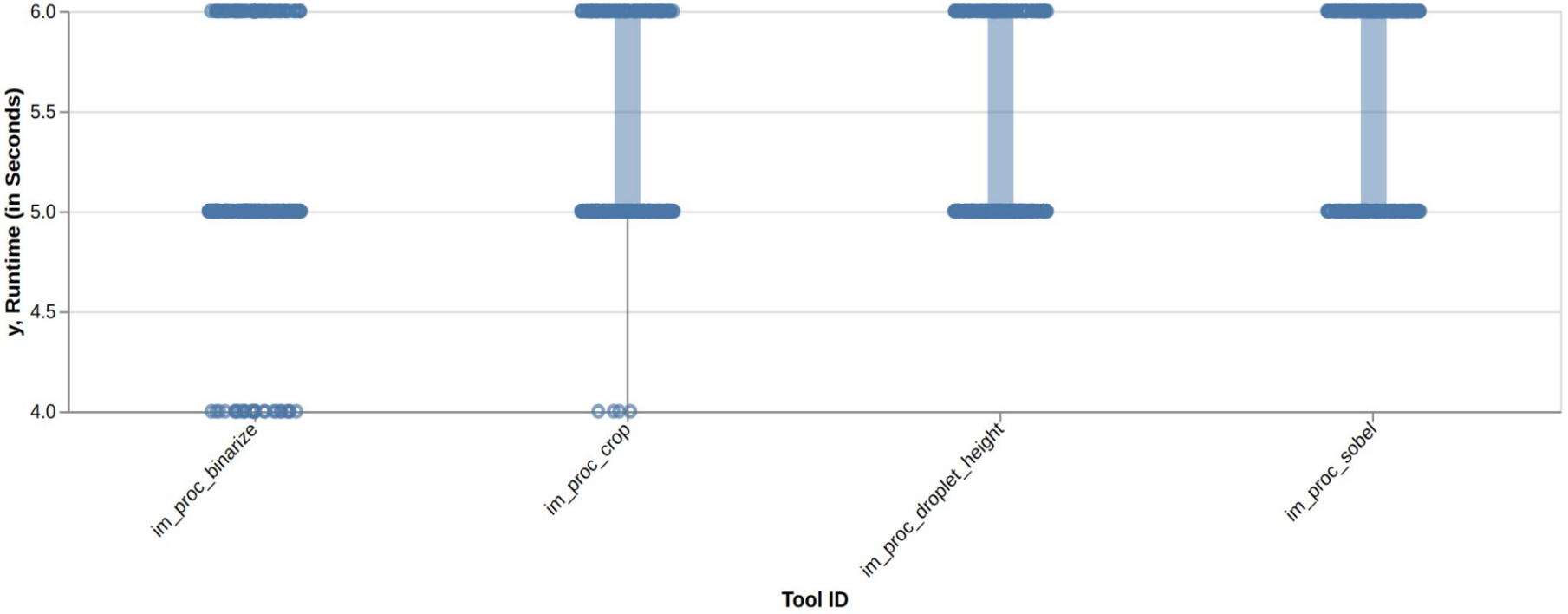
После того, как алгоритм отлажен на одном изображении, перейдем к обработки датасета целиком, весь эксперимент содержит 401 кадр. Для того, чтобы инструменты работали с коллекциями модифицируем созданный ранее workflow, добавив в начало инструмент для подачи коллекции на вход. Модифицированный алгоритм представлен на рисунке.



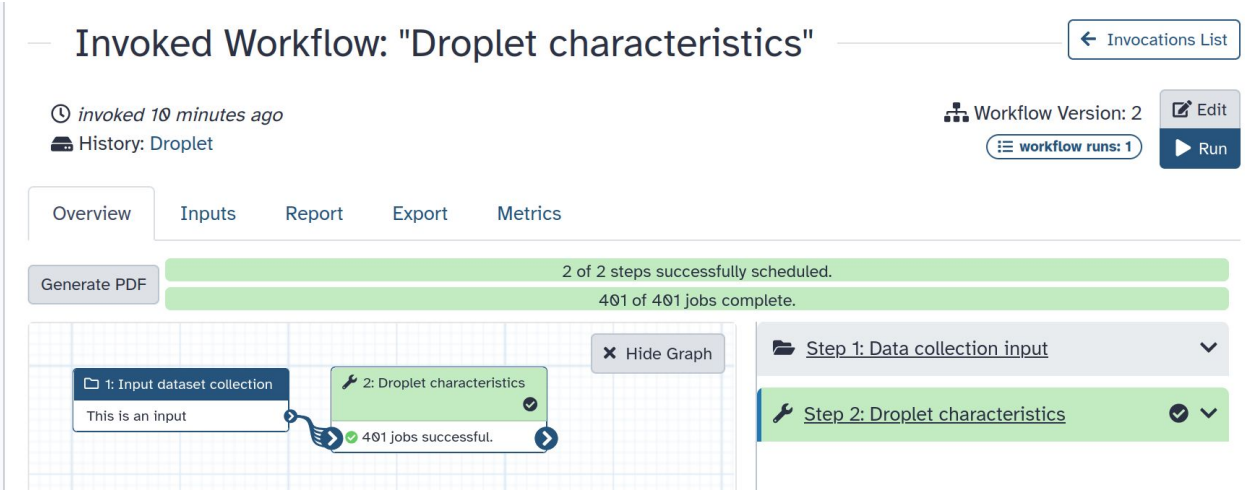
Тестирование полученного алгоритма на эксперименте

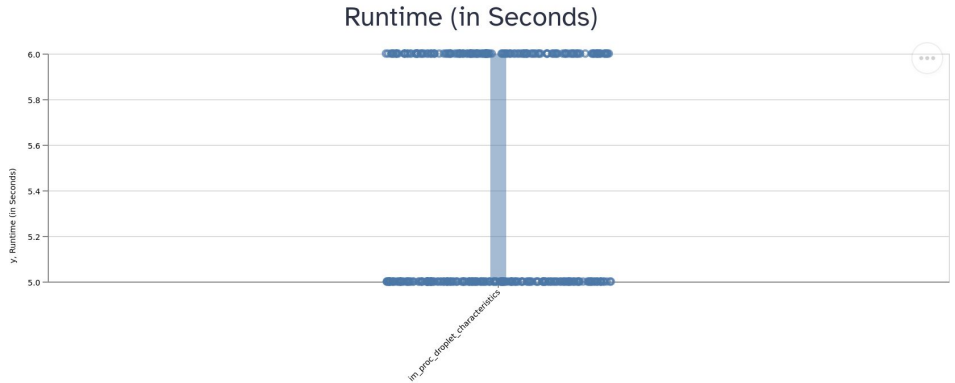


Время работы алгоритма



Оптимизация алгоритма

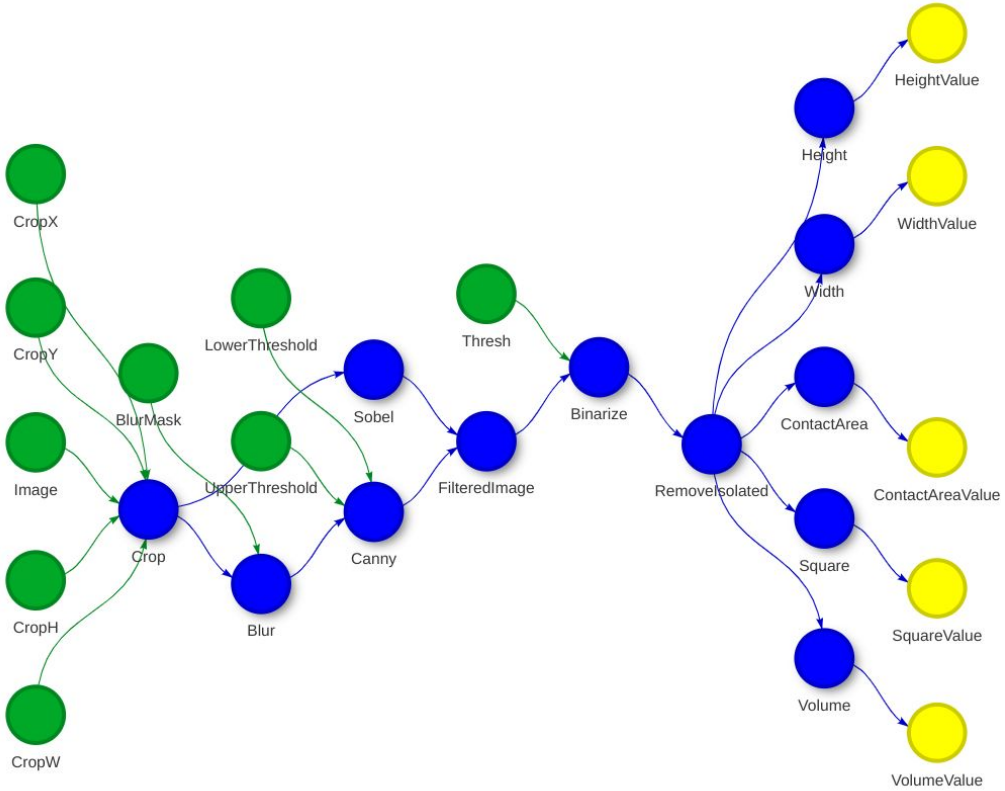




Common Workflow Language (CWL)

CWL - это открытый стандарт описание workflow, последовательности связанных между собой команд для командной строки.

Существует ряд различных инструментов, которые могут запускать процессы, описанные с помощью CWL, например, cwltool.

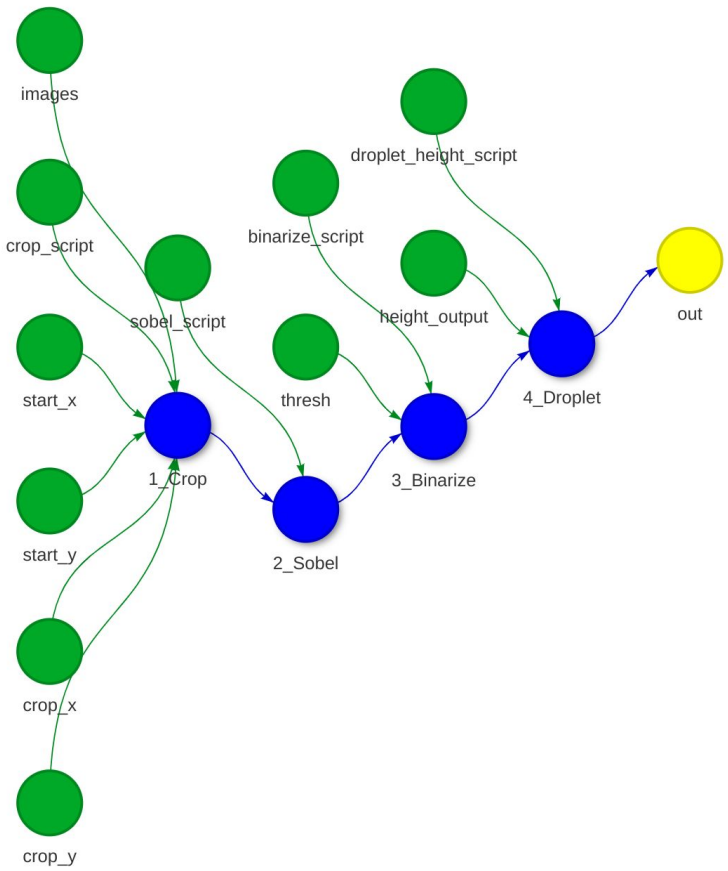


Реализация решения задачи с помощью CWL

Для реализации решения задачи с помощью CWL были проделаны следующие шаги:

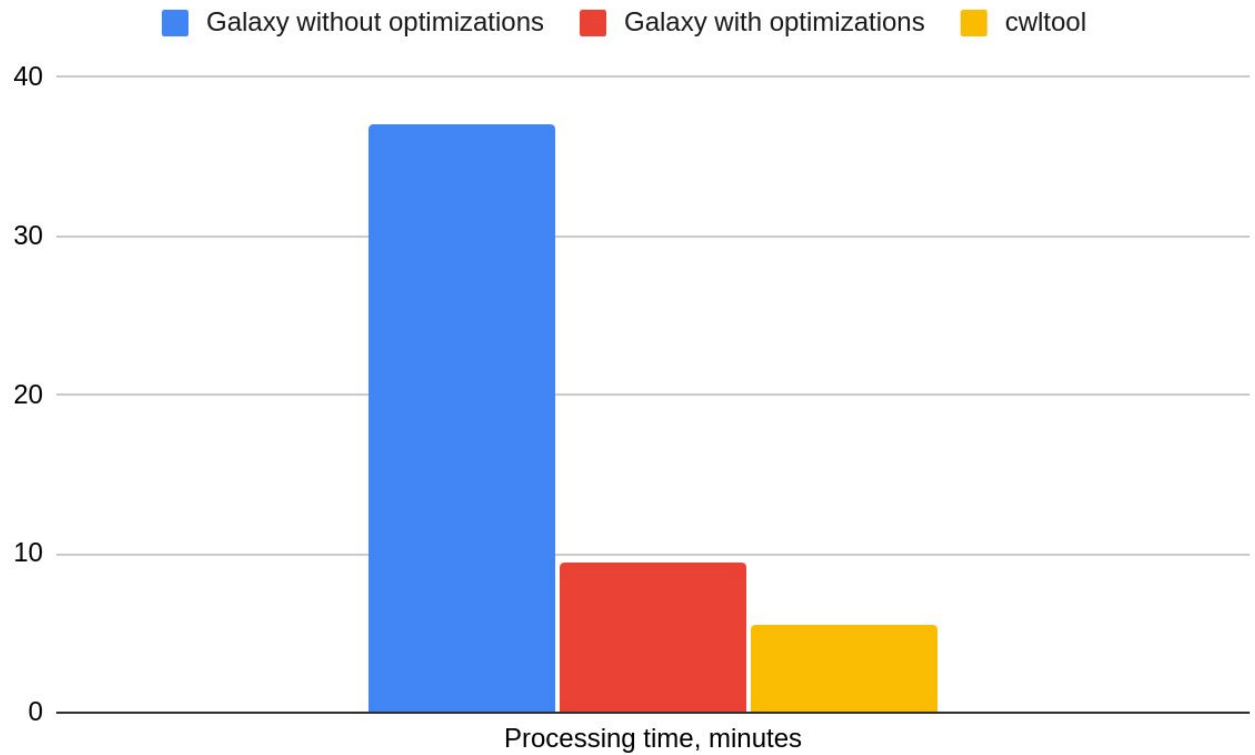
1. Для каждого шага алгоритма реализован инструмент в формате cwl
2. Входные данные описаны в формате XML
3. Построен workflow, состоящий из последовательности cwl инструментов
4. Полученный workflow запускается с помощью cwltool, эталонной реализации CWL
5. Итоговый workflow протестирован на тестовых данных

Итоговый CWL workflow




```yaml
steps:
  1_Crop:
    run: cwl_tools/crop.cwl
    in:
      images: images
      script: crop_script
      start_x: start_x
      start_y: start_y
      crop_x: crop_x
      crop_y: crop_y
    out: [output_files]
  2_Sobel:
    run: cwl_tools/sobel.cwl
    in:
      images:
        source: 1_Crop/output_files
      script: sobel_script
    out: [output_files]
  3_Binarize:
    run: cwl_tools/binarize.cwl
    in:
      images:
        source: 2_Sobel/output_files
      script: binarize_script
      thresh: thresh
    out: [output_files]
  4_Droplet:
    run: cwl_tools/droplet_height.cwl
    in:
      images:
        source: 3_Binarize/output_files
      script: droplet_height_script
      height_output: height_output
    out: [output_files]
```

Производительность Galaxy и cwltool



Сравнение Galaxy и cwltool

Criterion	Galaxy	cwltool
UI	+	-
Documentation	good	good
Open-source	+	+
Overheads[1]	high	medium
Parallelism	+	+
Data management	yes	no
Cloud deployment	native support	requires manual setup

Заключение

В данной работе показано:

- использование системы Galaxy и инструмента cwltool для решения проблемы обработки изображений;
- какие могут возникнуть проблемы с производительностью в системе Galaxy, и как можно оптимизировать полученные инструменты;
- возможности WMS в контексте обработки изображений:
  - гибкая отладка алгоритма;
  - автоматизация управления данными;
  - автоматическое распараллеливание обработки.
- сравнение системы Galaxy и подхода с использованием cwltool с точки зрения обработки изображений.

Спасибо за внимание!