

Slides at [fjukstad.github.io](https://fjukstad.github.io)  
Code at [bdps.cs.uit.no](https://bdps.cs.uit.no)



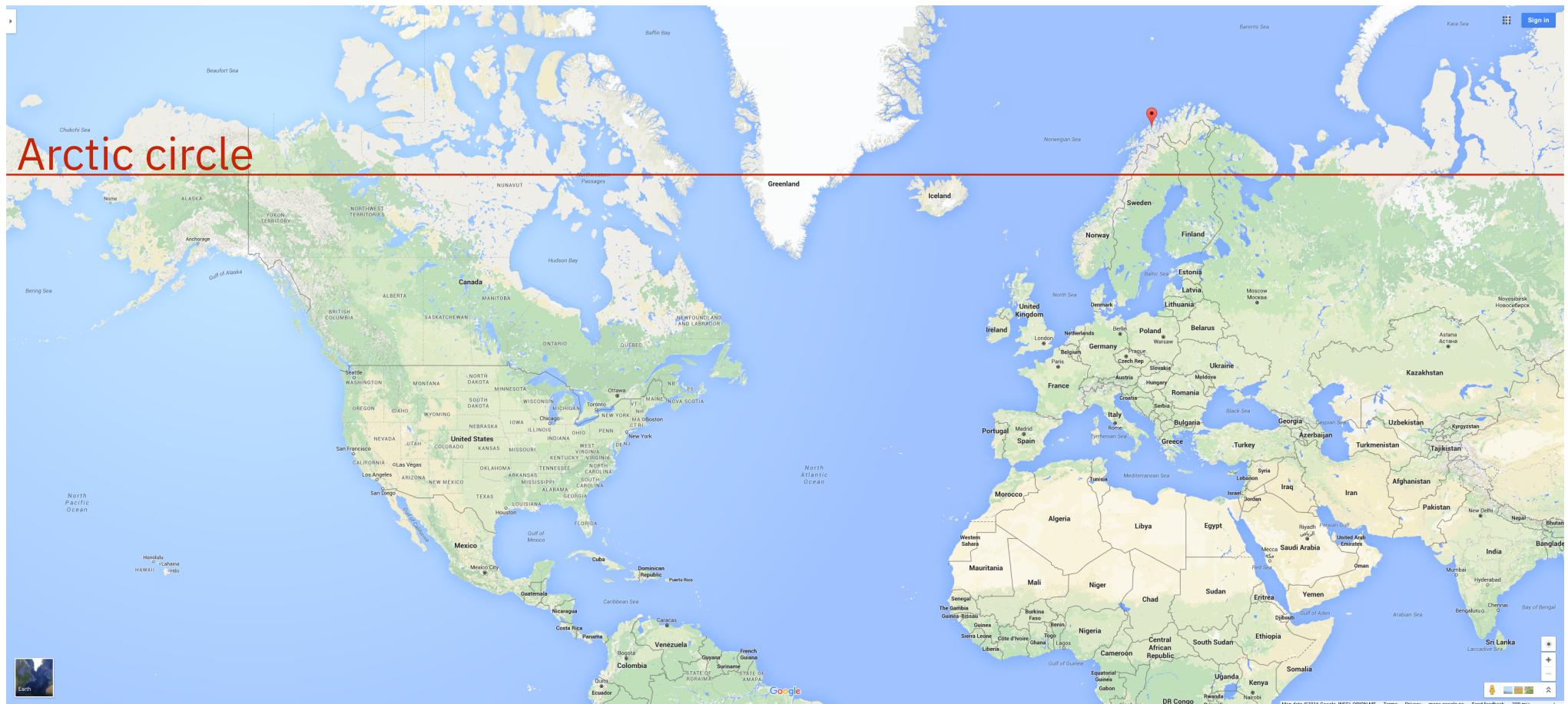
# Reproducible Data Analysis Pipelines for Precision Medicine

Bjørn Fjukstad, PhD

Department of Computer Science, UiT The Arctic University of Norway  
DIPS AS



## Tromsø



Tromsø, 9.2.2019



# Motivation

# Cornerstone of Science

- The principal goal of scientific publications is to teach new concepts, show the resulting implications of those concepts in an illustration, and provide enough detail to make the work reproducible.

# Great Potential

- High-throughput technologies and simpler access to datasets have revolutionized biology
- Datasets can reveal genetic basis of disease in patients, but require a collaborative effort
- Constantly ensure the quality of the data and analyses behind any interpretation
- Inaccurate results from improperly developed analyses can lead to negative consequences for patient care

# Precision Medicine

- Precision medicine uses patient-specific molecular information to diagnose and categorize disease to tailor treatment to improve health outcome.
- High-throughput sequencing is currently the main technology to facilitate personalized diagnosis and treatment
- Downstream computational analysis and interpretation is the major cost. \$1000 genome, \$1,000,000 analysis

# Genomic Analysis

- Analysis pipelines consist of many steps that transform raw data into interpretable results
- It is necessary to carefully explore different tools and parameters to answer a dedicated question
- Improperly developed analysis pipelines may generate inaccurate results, which may have negative consequences for patient care

# Definitions

# Different definitions

- **Replicate:** carry out the same task as the original researcher, and get the same results
  - Same code, same compiler, same hardware and same operating system (OS), gives the same result.
- **Reproduce:** carry out tasks that are similar to the original and get the same result
  - Same data, similar code, similar compiler, similar hardware, similar OS, gives the same result.

# Different definitions

- **Preproducibility:** An experiment is preproducible if it has been described in adequate detail for others to undertake it.

# Different definitions

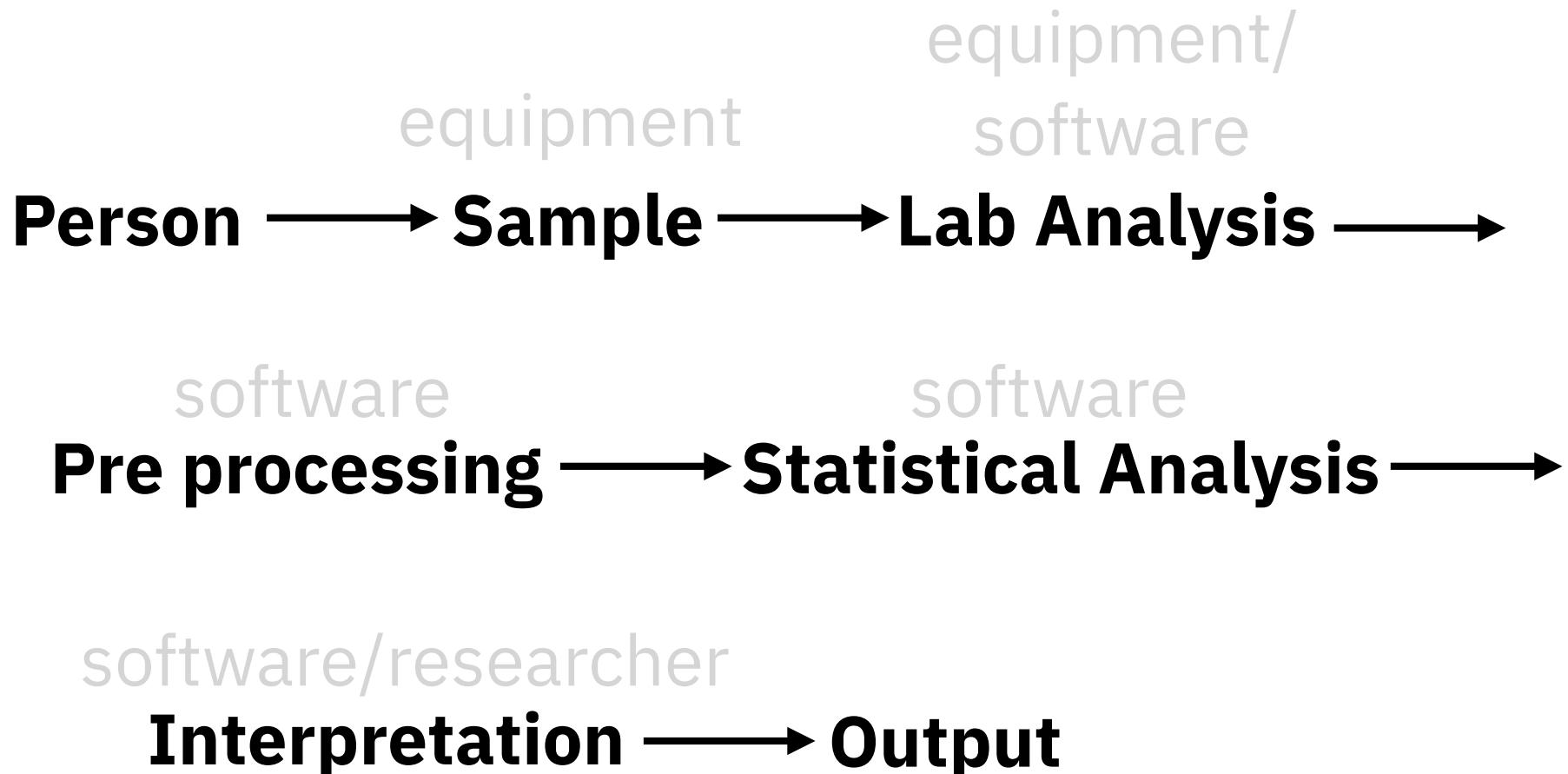
- **Verify:** The task of replicating an experiment to see if it yields the same results
- **Validate:** *the task of evaluating a result to see if the researcher's conclusions are warranted.*

# Consequences

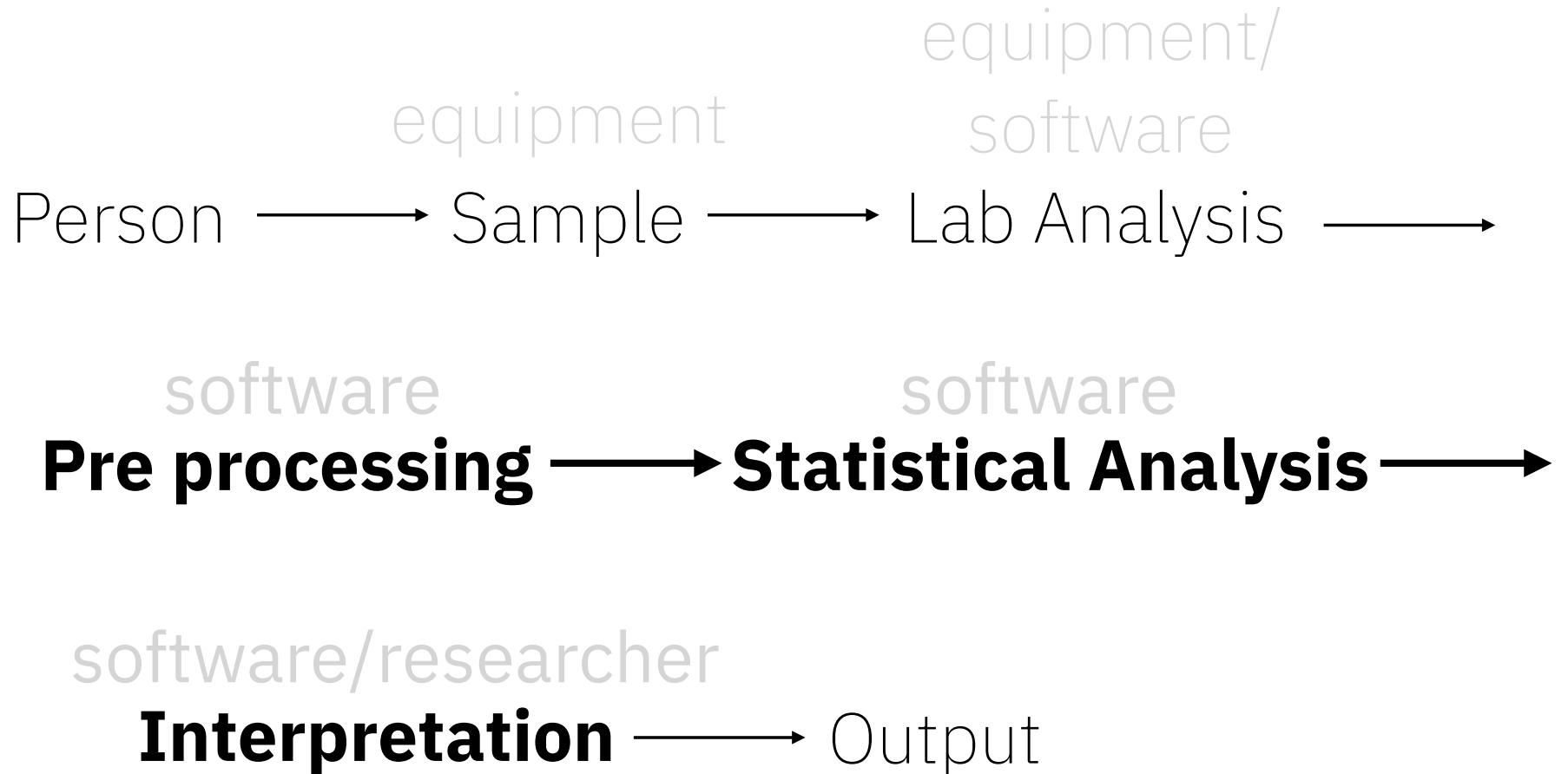
- Science moves forward when researchers verify other's results. Science advances faster when people waste less time pursuing false leads
- The society and companies can potentially waste hundreds of millions on failed drug development programs
- If science is not useful, why should we fund or support it? Why should people donate blood to a study that can't be reproduced?

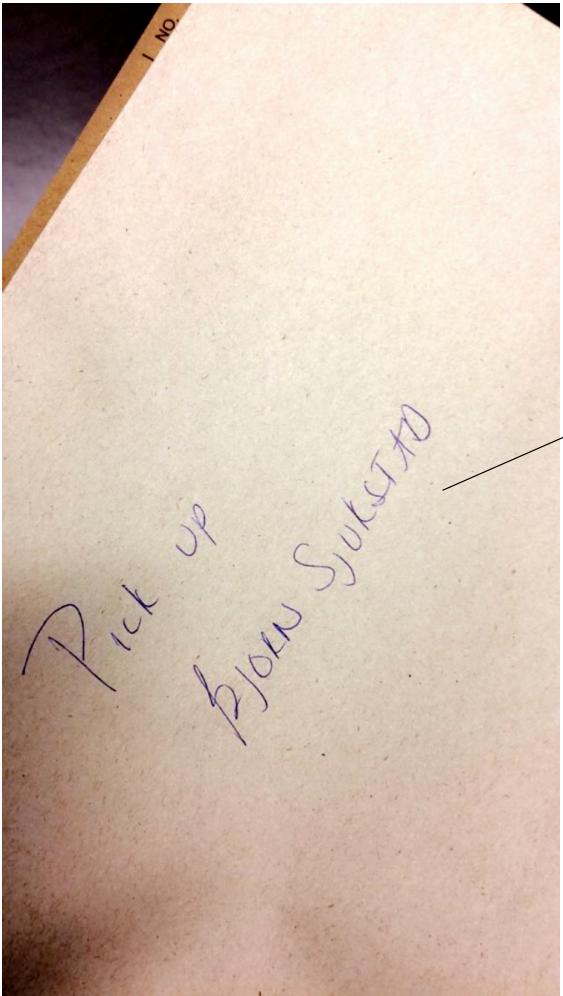
Where things (can) go wrong

# The Trajectory of a Biomedical Research Project



# The Trajectory of a biomedical research project



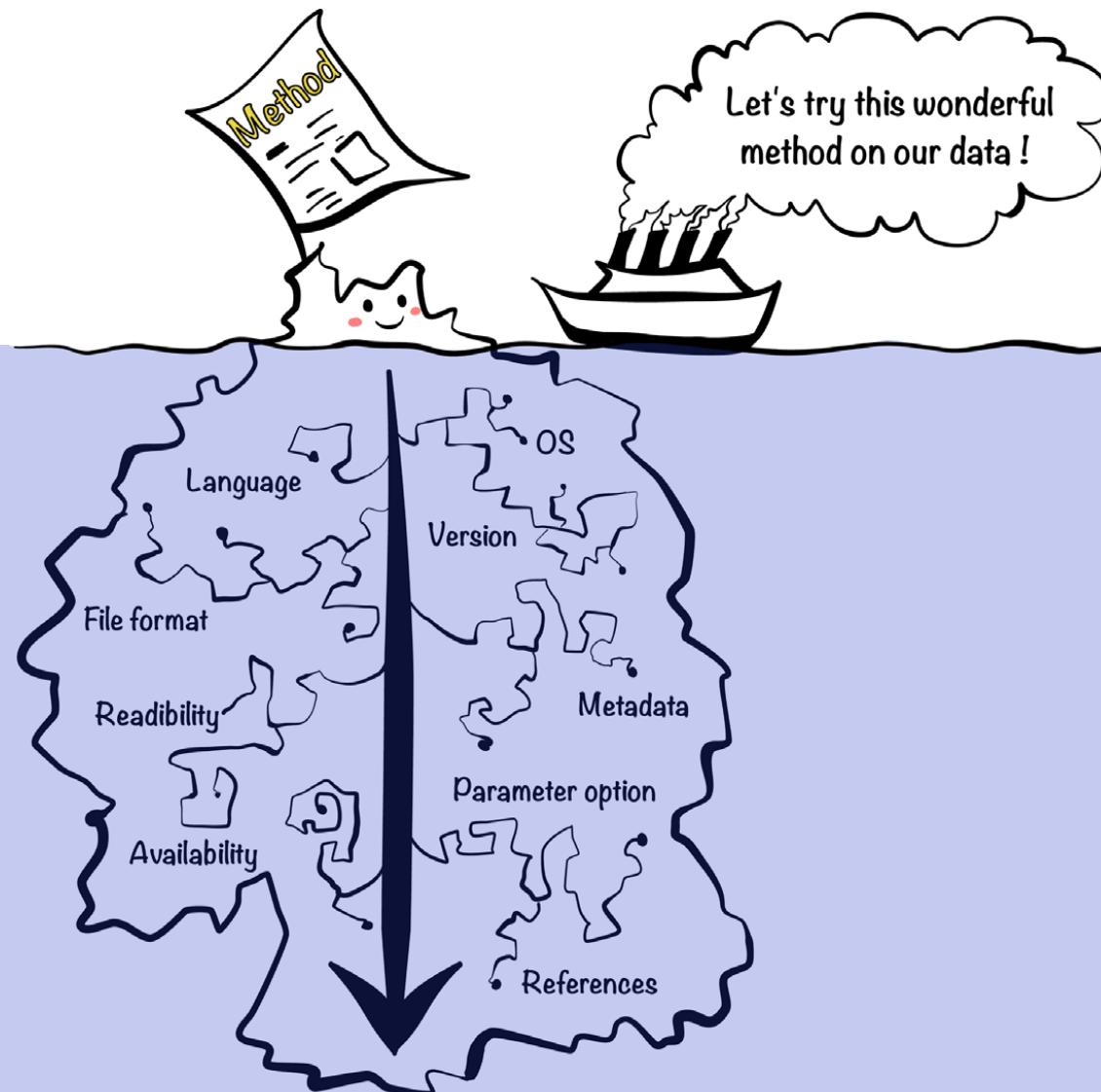


Bjørn ~~Fjukstad~~  
Sjukstad ≈ Sicktown

14 October 2016, pick up of biological sample at McGill University Health Centre

# The Problem





# 2016 Nature Survey

- Nature asked **1,576** researchers to respond to a brief online questionnaire on reproducibility in research
- **70%** responded that they have tried and failed to reproduce another scientist's experiment.
- More than **50%** have failed to reproduce their own experiments
- **52%** said it's a significant reproducibility crisis, **38%** said it's a slight crisis

# Cancer Research

- Biotechnology company Amgen tried to reproduce 53 “landmark” papers in cancer research : could only reproduce 11% of the findings
- Bayer HealthCare surveyed 67 in-house projects, but could only reproduce 21% of the projects

# “Reproducing” Computer Science Papers

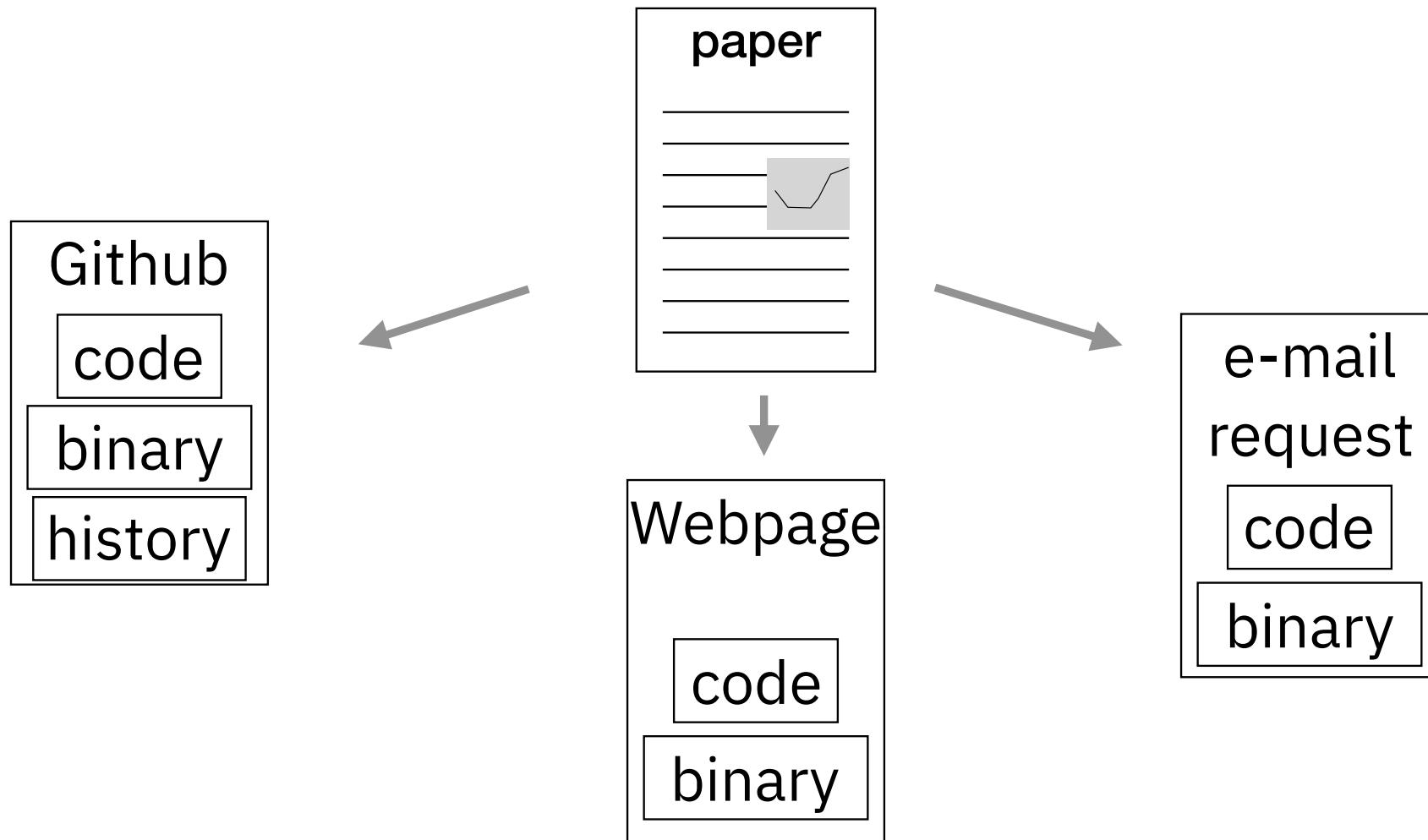
- Study from University of Arizona downloaded **601** papers from eight ACM conferences (including SOSP, SIGMOD, OSDI) and five journals.
- Investigated if the code was available, and if they could build it
  - Excluded **199** papers
  - Did not find, or receive, code for **176** papers.
  - **226** papers backed by code: **130** built within 30 minutes; **64** after more than 30 minutes; **23** not built, but authors claim it will; **9** cannot be built by team or original authors

# **BONUS:** Why researchers can't/won't share code

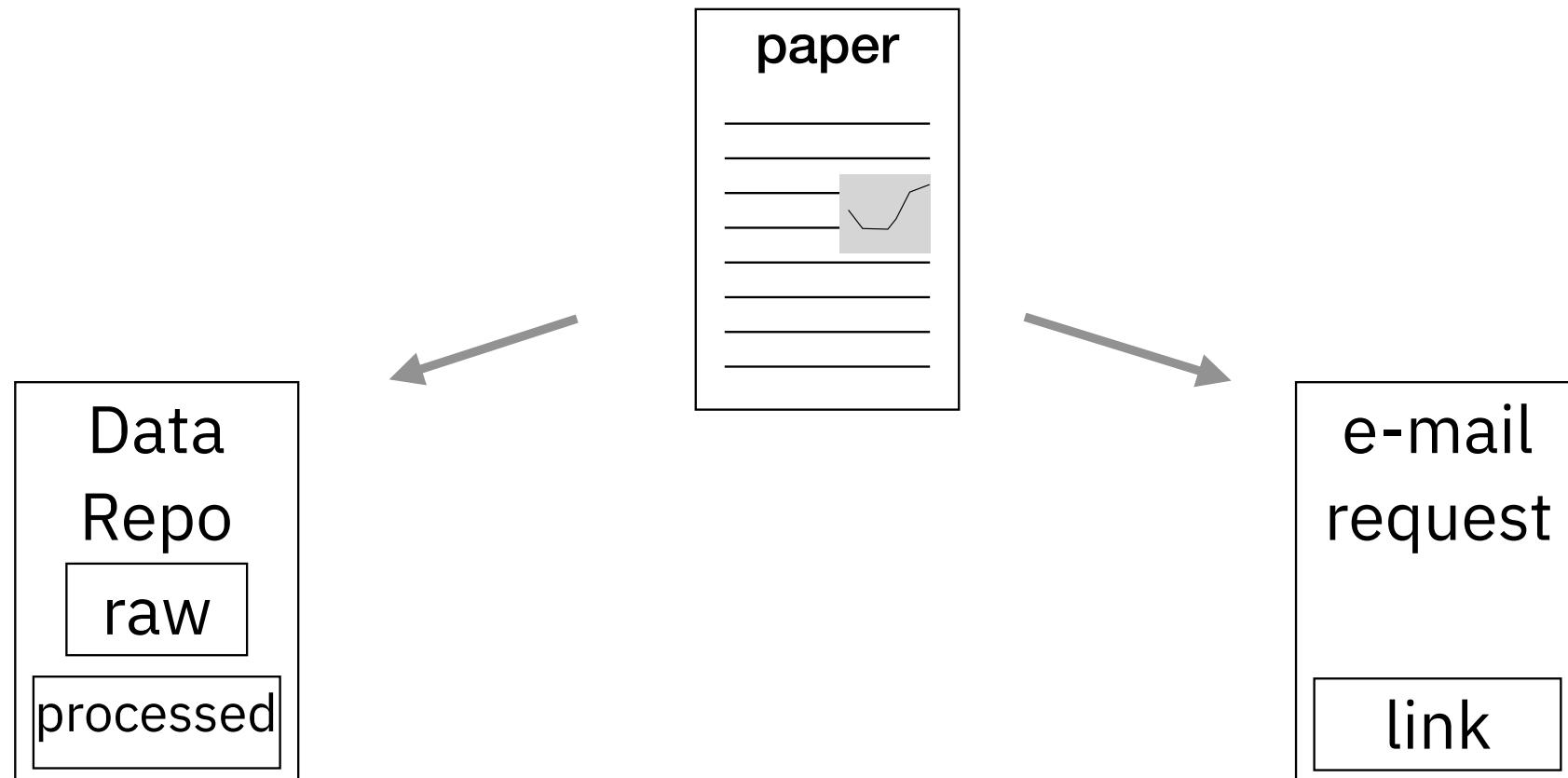
- *Versioning: Don't know if this version is the final one*
- *Available soon*
- *We don't intend to make the code available, ever*
- *The programmer left*
- *The system physically ceased to exist, got stolen or crashed*
- *The code is commercial or proprietary, and won't be open-sourced or given a permissive license*
- *Too busy to help out*

# The technical and practical difficulties of reproducing results

# Simply getting the required program

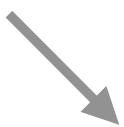


# Get a hold of the data



**paper**

# Example



```
> Rscript analysis.R --method="a" --input=data.rda --output=output.csv
```

**data  
.rda**



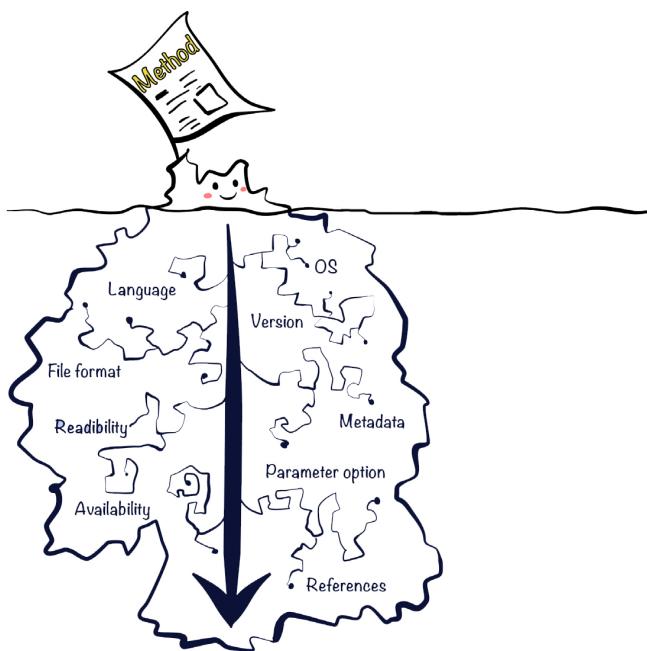
**analysis.R**



**output.csv**

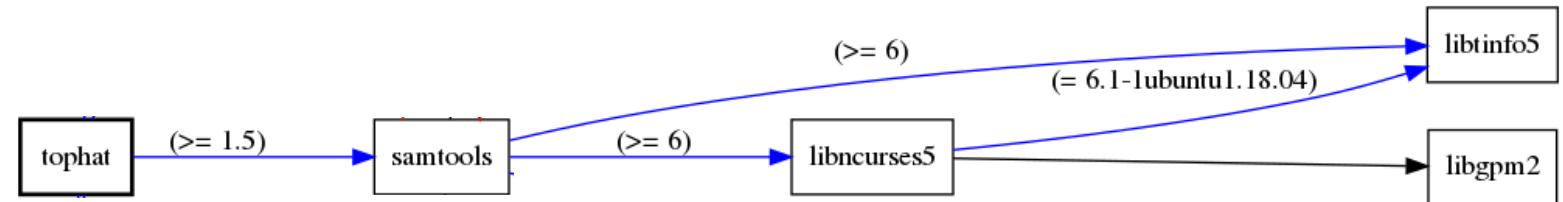
# Difficulties all the way down

```
> Rscript analysis.R --method="a" --input=data.rda --output=output.csv
```



Command, data, and parameters	a, data.rda, output.csv
Software	R/Rscript 3.40, plotly 3.4.1, ...
Operating System	OS X High Sierra 10.13.6
Kernel	Darwin Kernel Version 17.7.0 xnu-4570.71.2~1/RELEASE_X86_64 x86_64
Hardware	Macbook Pro (Retina, 13-inch, Mid 2014), 3 GHz Intel Core i7, 16 GB RAM, 250 GB SSD

# TopHat Dependencies

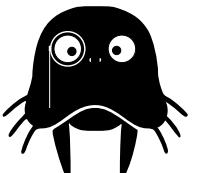


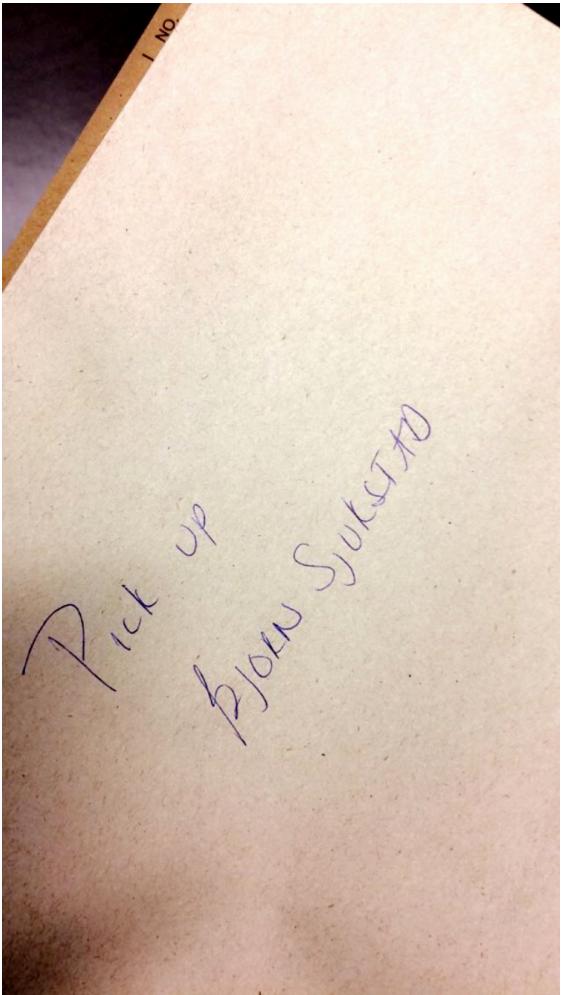
# TopHat Dependencies



# walrus

[github.com/uit-bdps/walrus](https://github.com/uit-bdps/walrus)





# Use Case

- Previous: Identify the molecular signature of the patient's tumor and germline with DNA sequence data from the patient's primary tumor and adjacent normal cells
- **Provide an extensive comparison of the metastasis against the primary, and to identify the molecular drivers of the tumor**

14 October 2016, pick up of biological sample at McGill University Health Centre

# Lessons from the initial analysis pipeline

- The pipeline tools should be kept for later use, either as binaries or code
- Datasets and databases should be stored along with the pipeline description
- It should be easy to install and include new tools to a new or existing pipeline

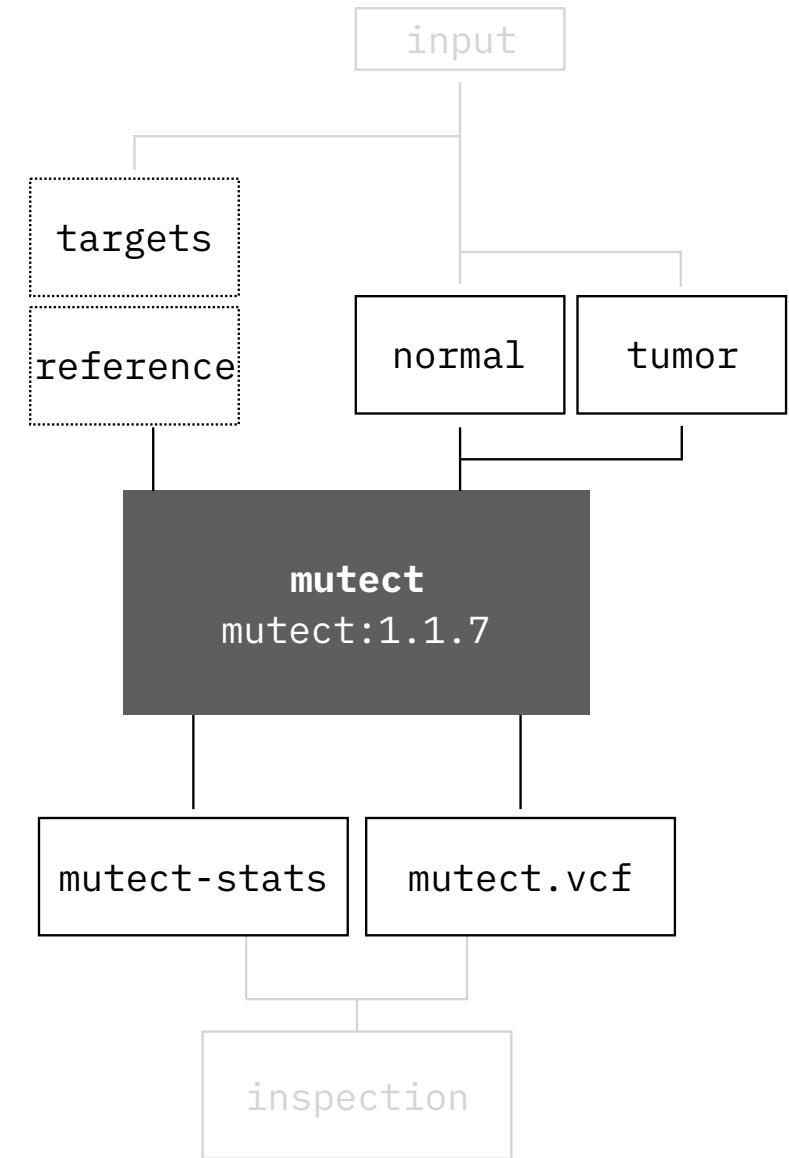
# Related Work

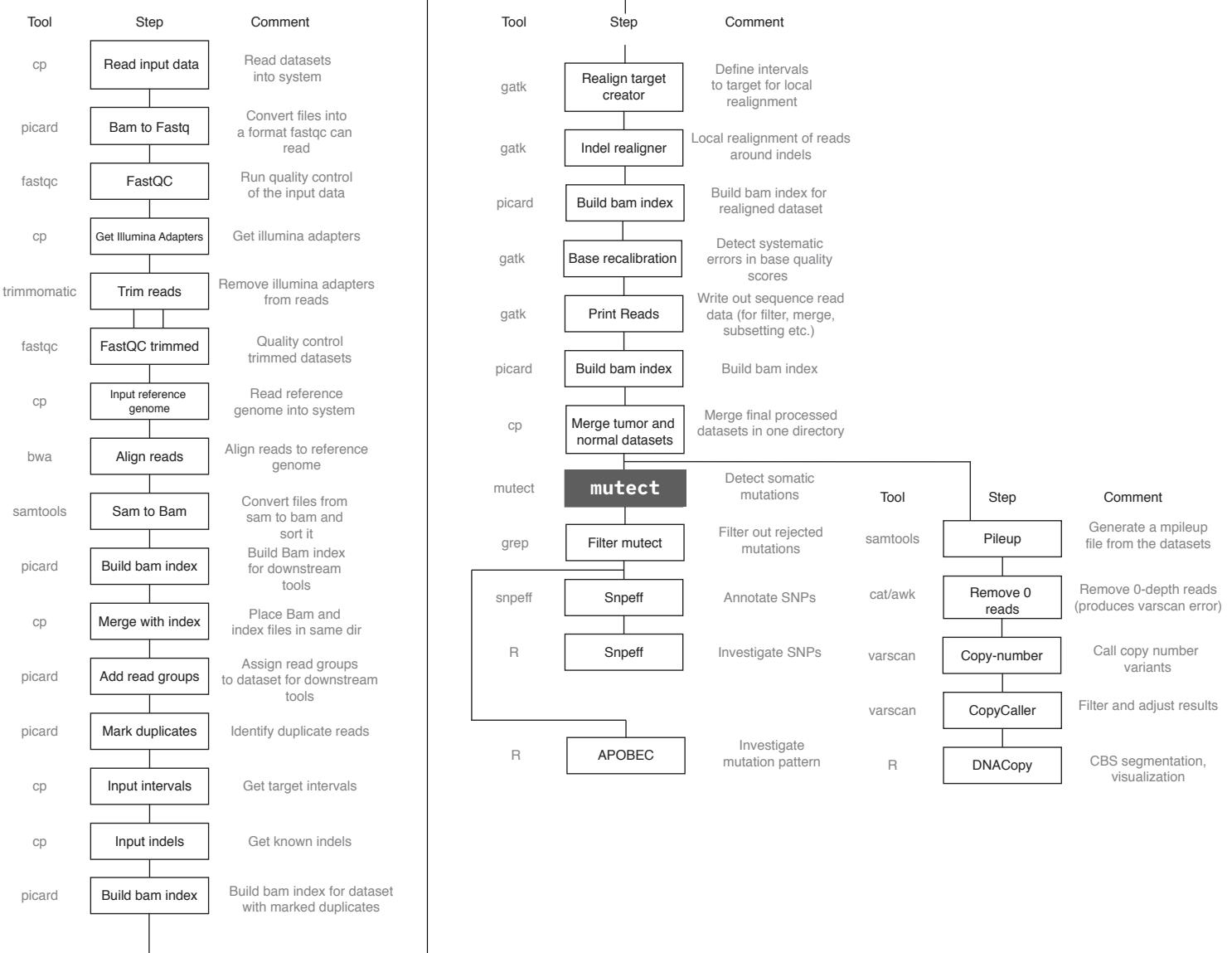
- A wealth of pipeline description formats and workflow managers available
- Common Workflow Language (CWL) is a popular specification formation that is supported by many workflow managers, e.g. Galaxy, Rabix, Toil, and AWE. Implementation specific data management.
- Pachyderm provides version control for data, and reproducible data analysis pipelines. Runs on top of Kubernetes.
- Snakemake is a popular tool that has lately got support software containers

# walrus

- Tool for developing and executing data analysis pipelines
- Stores: tool versions, tool parameters, input data, intermediate data, output data, along with the execution environment
- Users write pipeline descriptions in JSON/YAML files
- Users use the commandline interface which executes the pipeline

```
{
  "Name": "mutect",
  "Image": "fjukstad/mutect:1.1.7",
  "Cmd": [
    "--analysis_type", "MuTect",
    "--reference_sequence", "/walrus/input/reference.fasta",
    "--input_file:normal", "/walrus/input/normal.bam",
    "--input_file:tumor", "/walrus/input/tumor.bam",
    "-L", "/walrus/input/targets.bed",
    "--out", "/walrus/mutect/mutect-stats-txt",
    "--vcf", "/walrus/mutect/mutect.vcf"
  ],
  "Inputs": [
    "input"
  ]
}
```





# Pipeline Execution

- All tools are Docker images, and each pipeline step is run within a Docker container. Simplifies dependency management
- Users can create these images themselves, or use existing images from e.g. BioContainers
- A pipeline is run on a single server, but supports data and task parallelism

# Data Management

- We version control all data: input, intermediate, and output
- We use git for the pipeline description, and git-lfs for the datasets
- The commandline interface takes care of tracking, and optionally restoring previous datasets

# Results

- Has shown its usability in a clinical setting. We analyzed a patient's metastasis to discover Single Nucleotide Polymorphisms(SNPs), genomic variants and somatic mutations
- Implemented a variant calling pipeline for a public dataset to evaluate performance and resource usage.

# Discussion

- Still no standard for representing or sharing workflows. CWL is gaining popularity and more systems now support it
- In walrus we propose a solution where researchers can share pipeline descriptions *and* datasets
- Forcing users to package tools in Docker images can help ensure reproducibility

# Summary

- A system for reproducible analysis of genomic data in a clinical setting
- Analyzed a patient's metastasis to discover SNPs, genomic variants and somatic mutations
- Performance evaluation and example pipelines to get started

# Acknowledgement

- Advisors: **Lars Ailo Bongo, Karina Standahl Olsen, and Eiliv Lund**
- Biological Data Processing Systems Lab: **Einar Holsbø, Morten Grønnesby, Edvard Pedersen, Nikita Shvetsov, Jo Inge Arnes, Tengel Skar**
- Norwegian Women and Cancer (NOWAC)
- Concordia University: **Vanessa Dumeaux and Mike Hallett**
- Department of Computer Science at UiT

Slides at [fjukstad.github.io](https://fjukstad.github.io)  
Code at [bdps.cs.uit.no](https://bdps.cs.uit.no)



# Reproducible Data Analysis Pipelines for Precision Medicine

Bjørn Fjukstad, PhD

Department of Computer Science, UiT The Arctic University of Norway  
DIPS AS

