



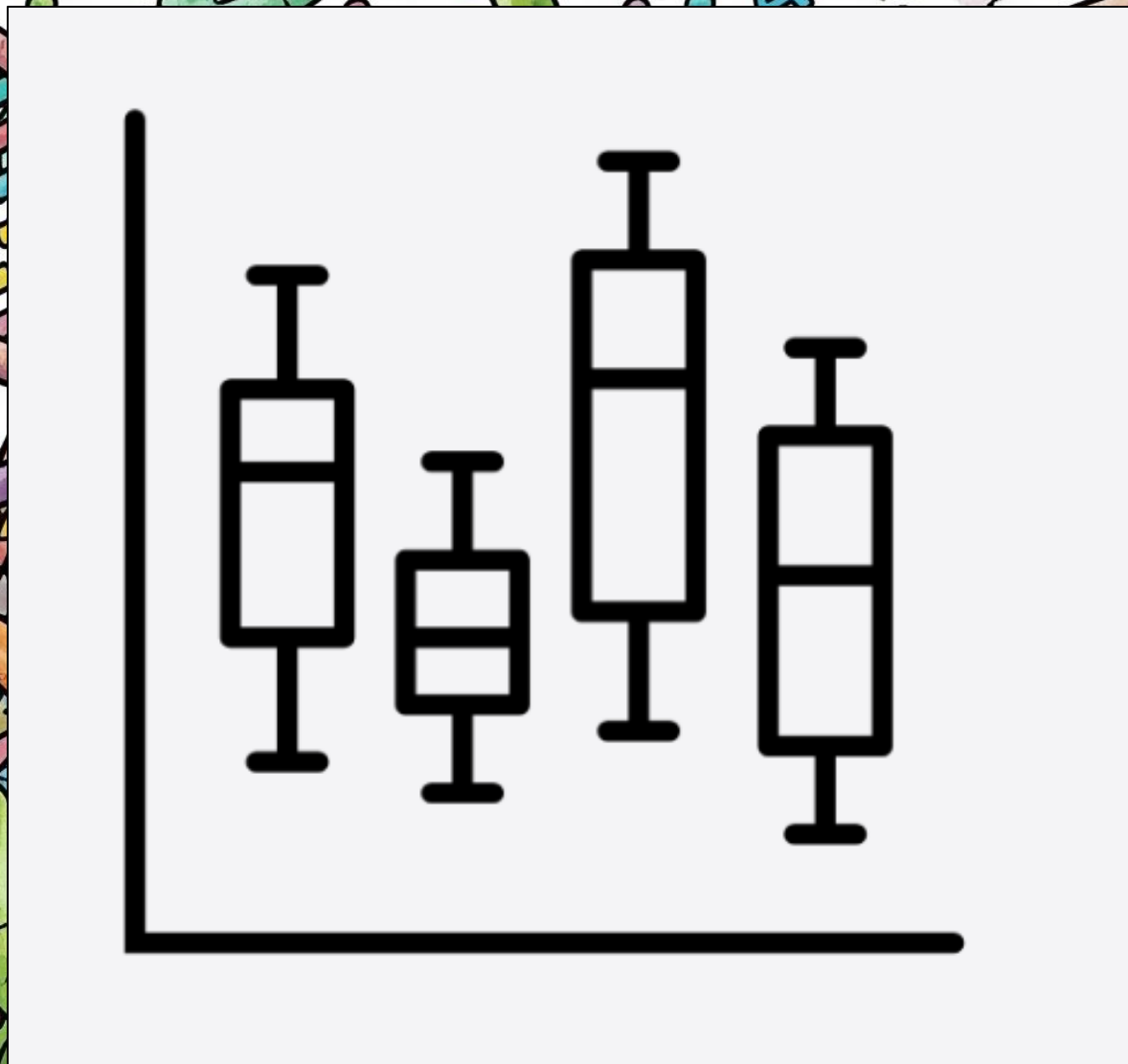
# Essential Tools for Reproducible Research

Pipelines, Workflow Management Systems &  
Containers

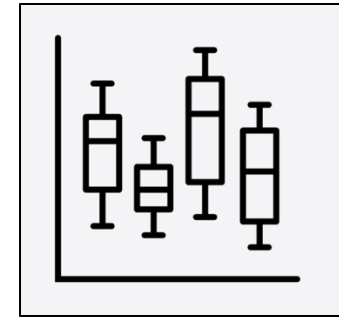




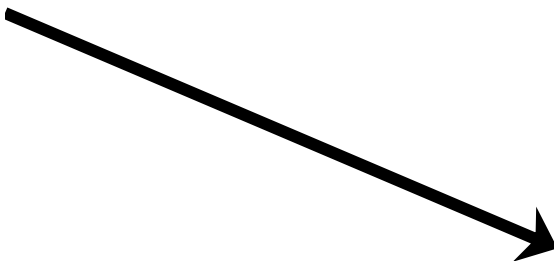
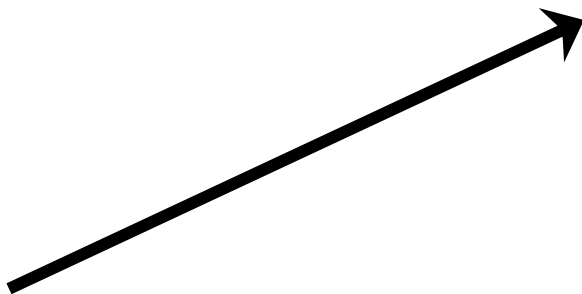


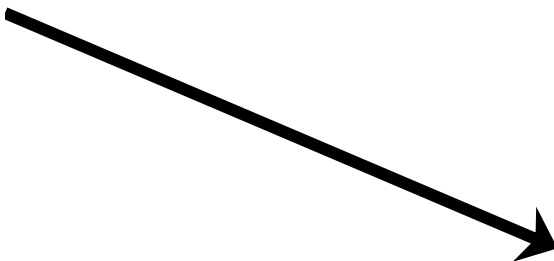
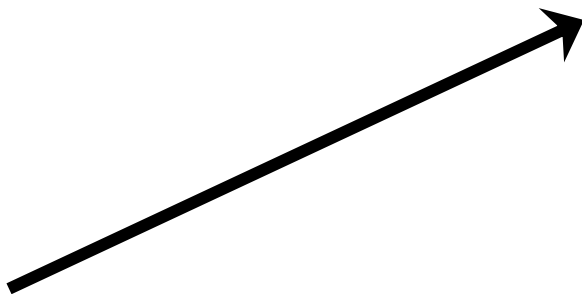












# The Challenge in Bioinformatics Research

- Bioinformatics analyses involve many **distinct, often interdependent**, computational steps.
- Software versions, operating systems, and installed libraries can be **inconsistent** leading to different results.
- "It worked on *my* machine!"



# So... what can we do?

- **Pipelines**
- **Containers**
- **Workflow Management Systems**

# Pipelines - What & Why

**A specific set programs or scripts to that is run in a predefined, automated sequence**

- Reduces manual errors, saves time, and provides detailed, step-by-step record of exactly what happened to your data.

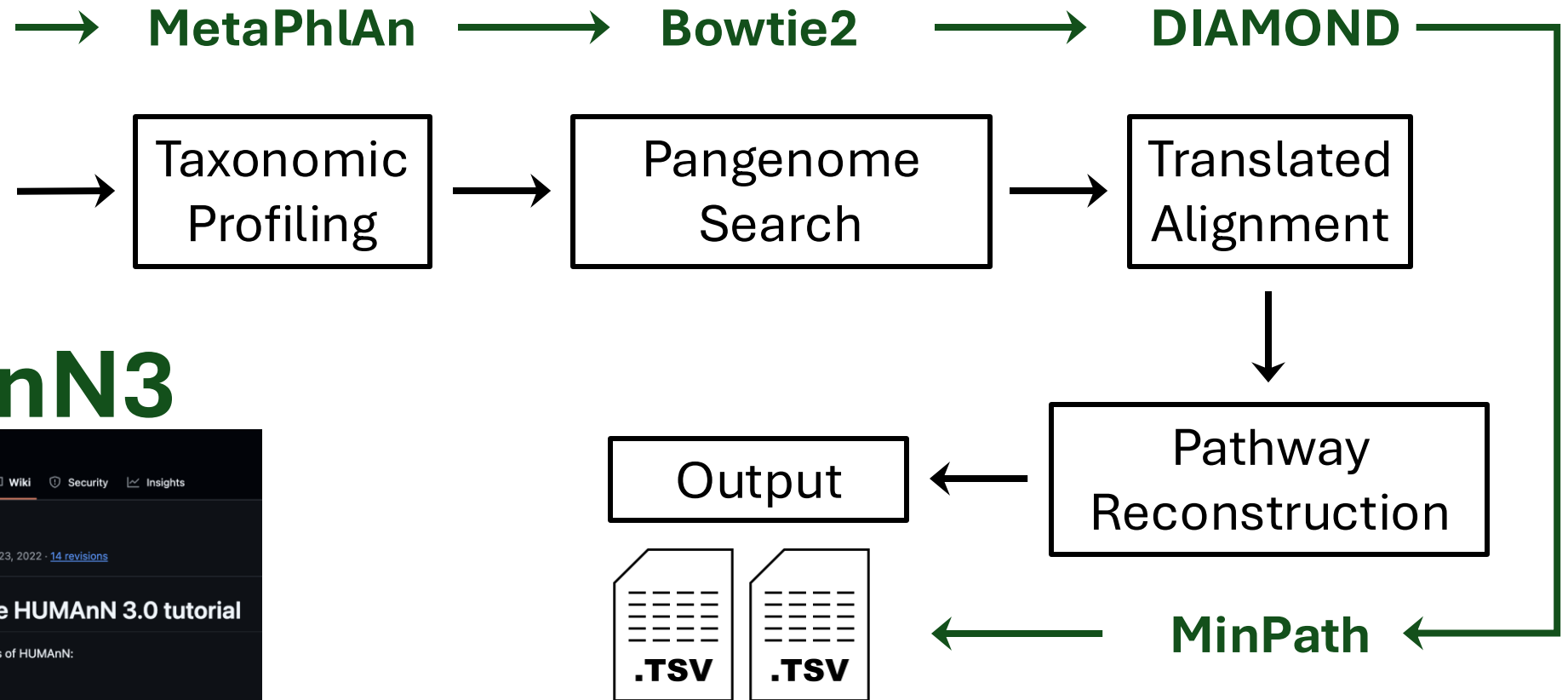
## A 'pattern' for your analysis



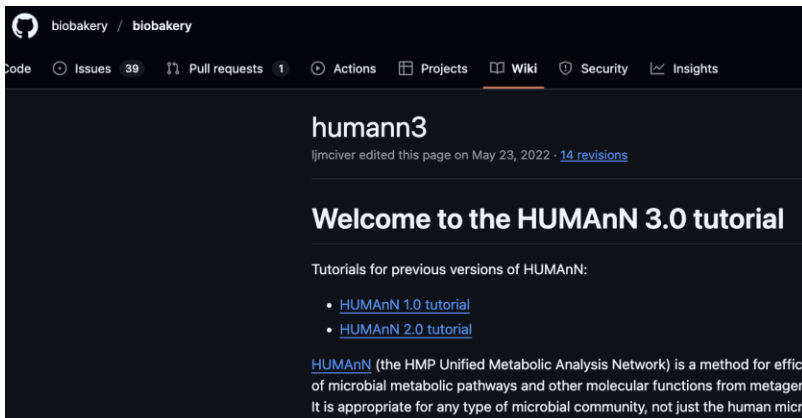


# Example: Metagenomics Analysis

```
@SRR26872690.1 M04209:105:000000000-KBJ
13:5423 length=148
CGTCCACCATACCGCCGCGGAAGCATCATCCCCGCC
ACGCTCAGGAAATGCAGCAGCAAGATAATCAGAGTATC
GCGGCAGACTTGCCACCAAGTCCAACCAATCAAGCAAC
+SRR26872690.1 M04209:105:000000000-KBJ
13:5423 length=148
```



## HUMAnN3



# Containers – Keeping things constant



- Self-contained, executable packages that bundle everything you to run whatever you want to run
- Ensures your analysis runs **identically, anywhere, anytime.**

The '*knitting kit*' of your analysis





# Our tools for making & managing containers

- **Docker:** Most widely used platform for developing,  shipping, and running applications in isolated containers.
- **Apptainer:** Specifically designed for high-performance computing (HPC) and shared cluster environments. 

<https://www.docker.com/>

<https://apptainer.org/>



## Elements:

- **Base Operating System:** Ubuntu (Version 24.04 LTS)
- **R Interpreter:** (R version 4.3.2)
- **R Packages:** (ggplot2 version 3.4.4, dplyr version 1.1.4)
- **R Script:** (my\_analysis.R)
- **Filesystem:** (the layered structure combining all elements above)



How do we keep our pipelines  
and containers organized??


# Workflows Management Systems

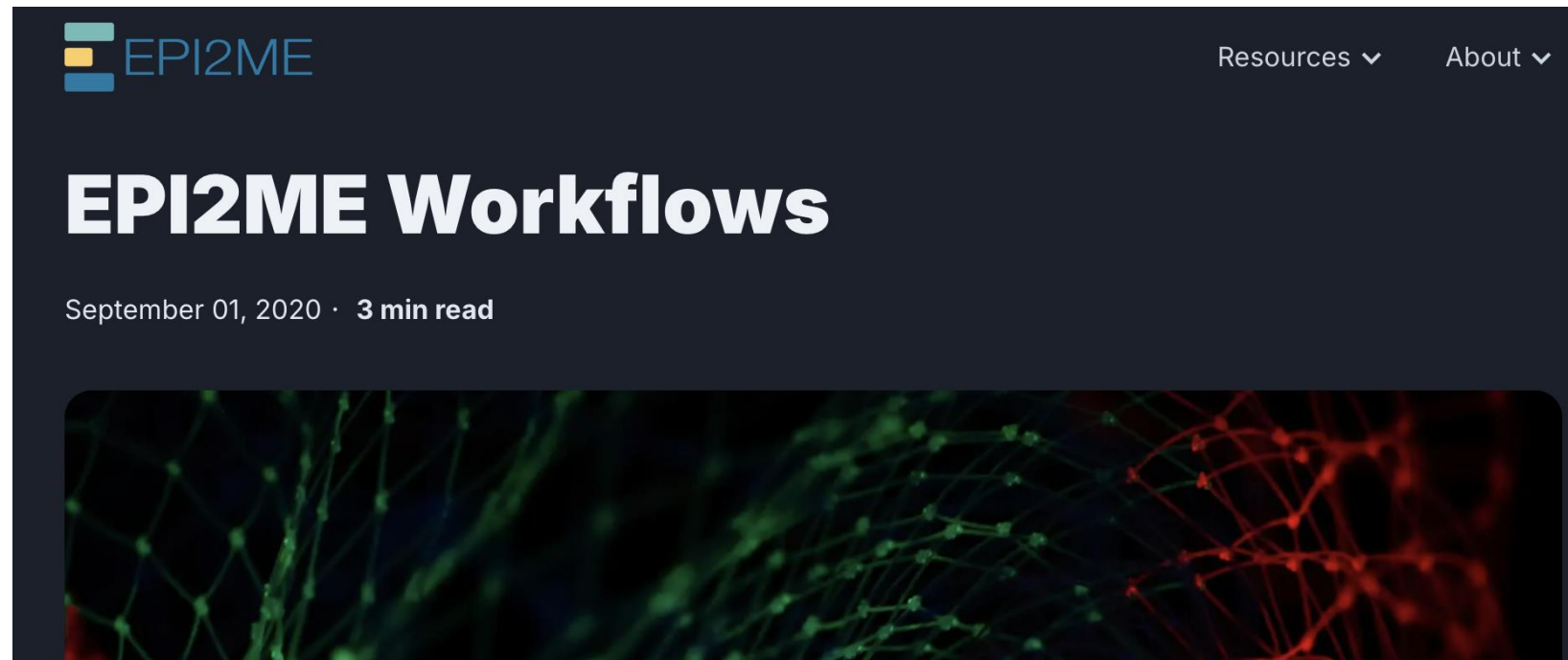
- Software framework or engine that helps you build, manage, and execute pipelines (often within docker containers).
- Handles dependencies, manages failures, and efficiently distributes computations across various resources.
- Improves robustness, scalability, and overall project organization.

*'Master Knitter'*



# Workflows Management Systems

 **nextflow** - Highly scalable and portable pipelines, especially good for distributed computing on clusters or cloud platforms.







# Workflows Management System

Orchestrates the overall analysis strategy

**Containers**  
Consistent and isolated  
environment



+

**Pipelines**  
Automated, sequential steps



**Reproducible  
Results**

