Bioinformatics Training Using Cloud Based Computing

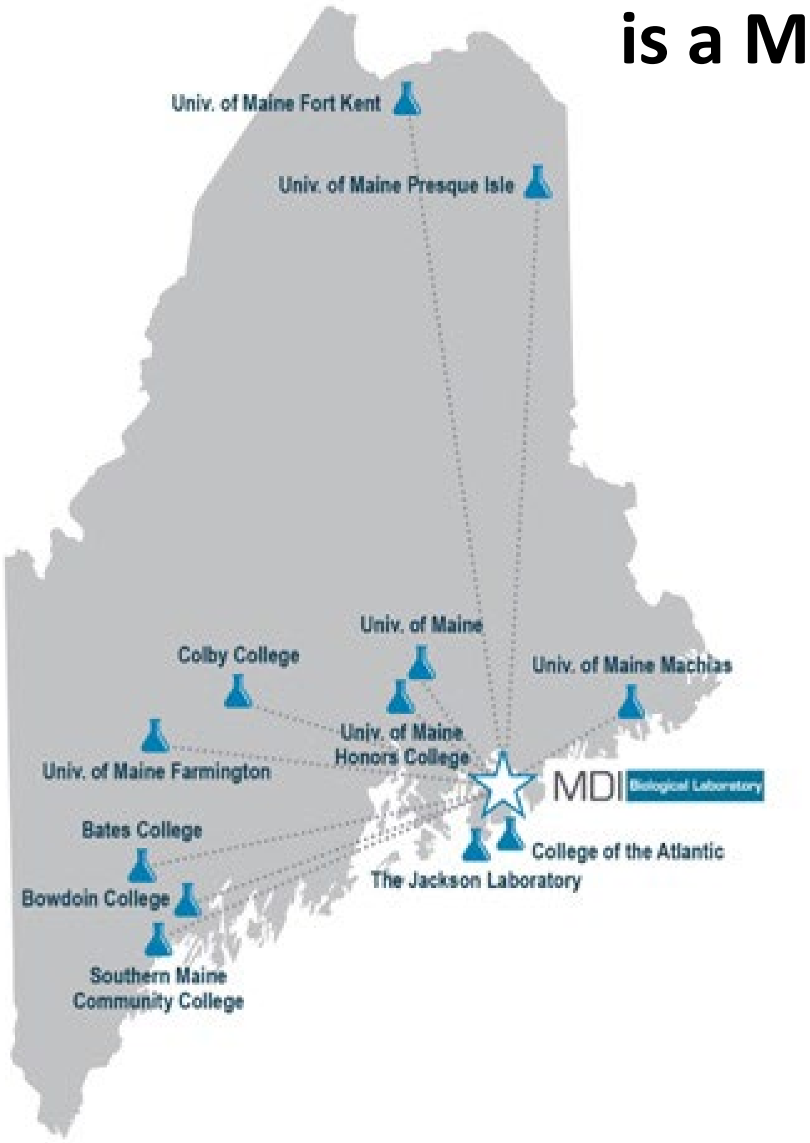
Benjamin King

Co-Director, Maine INBRE Bioinformatics Core

Assistant Professor of Bioinformatics, University of Maine

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Bioinformatics Research Training and Workforce Development is a Major Focus of Maine INBRE



Statewide Collaboration

Support Research Investigators

Undergraduate Short Courses

Bioinformatics Scholars Program

Regional and National Collaboration

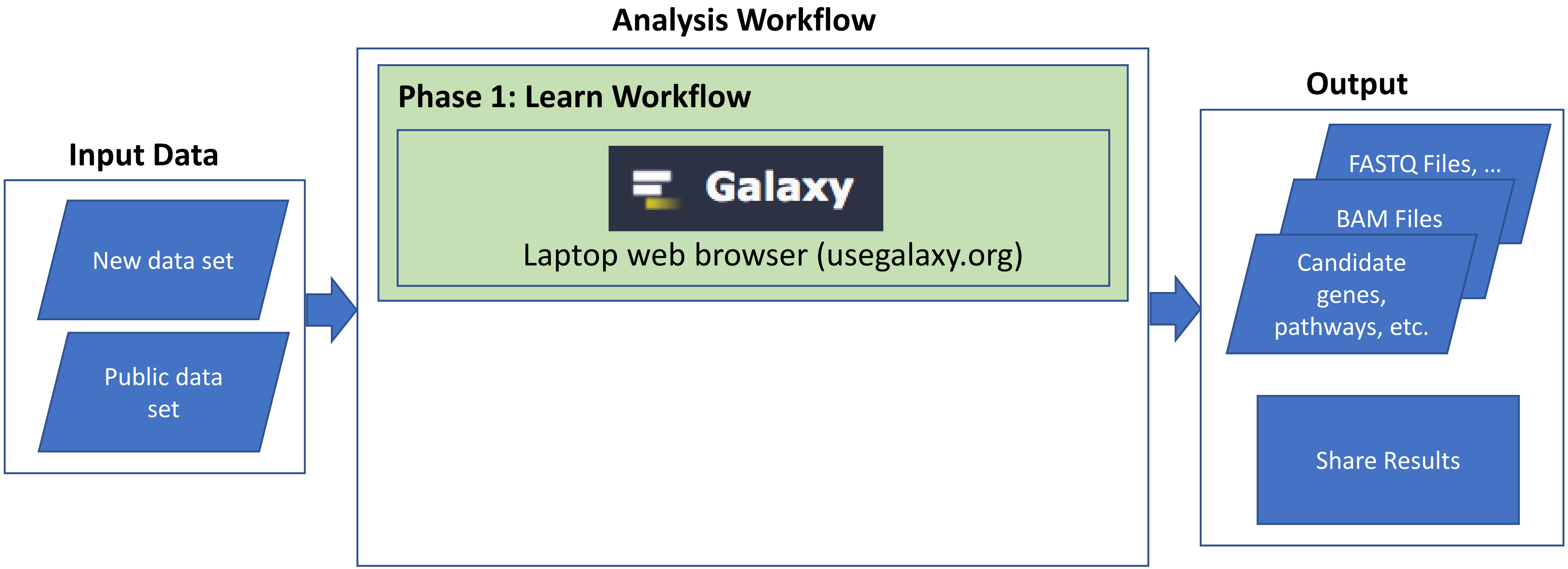
- Bioinformatics Workshops and Courses
  - MINOLTA: Maine INBRE Non-model Organism Transcriptome Analysis Workshop
  - Applied Bioinformatics Course
  - Bioinformatics Train-The-Trainer Course

North East Cyberinfrastructure Consortium

North East Bioinformatics Collaborative

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Bioinformatics Research Training Without Cloud-Based Resources



Input Data

New data set

Public data set

Analysis Workflow

Phase 1: Learn Workflow

Galaxy

Laptop web browser (usegalaxy.org)

Output

FASTQ Files, …

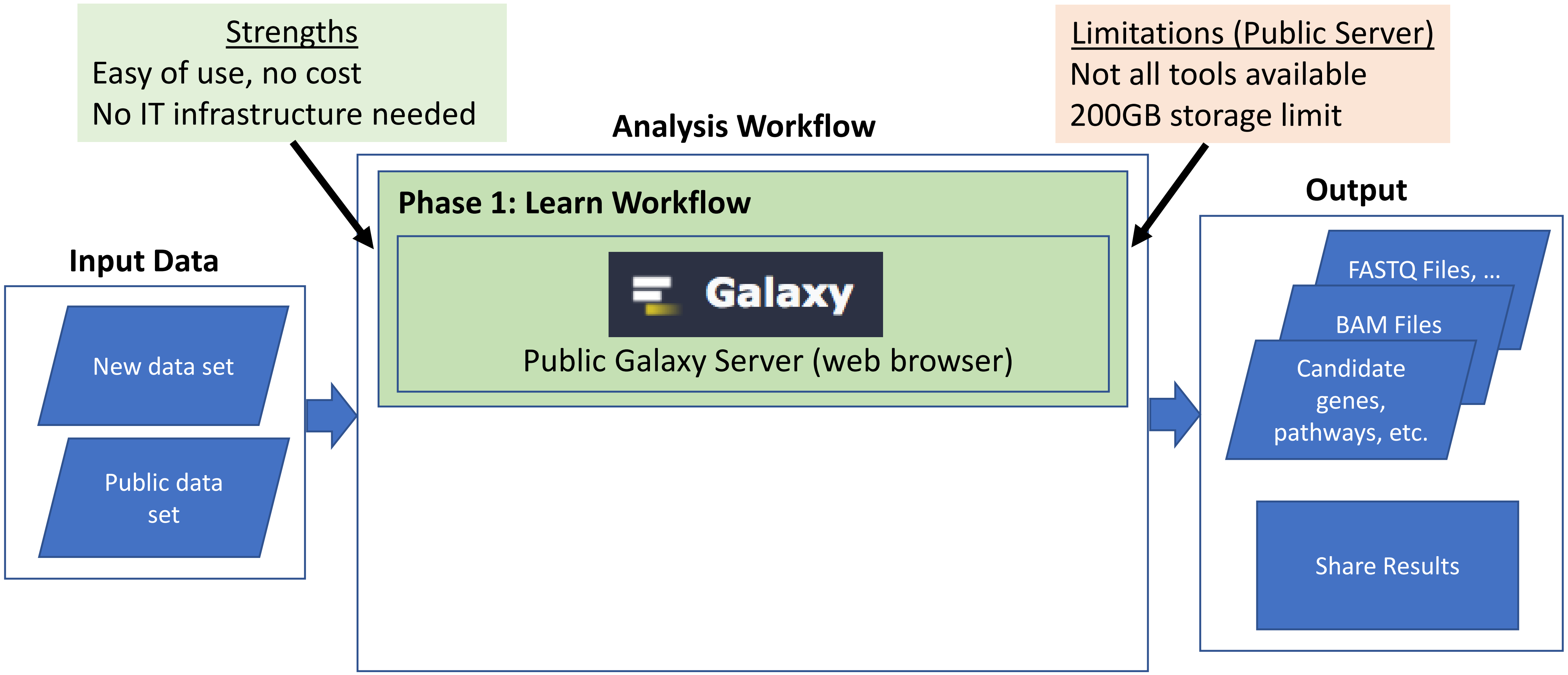
BAM Files

Candidate genes, pathways, etc.

Share Results

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Bioinformatics Research Training Without Cloud-Based Resources



Strengths

Easy of use, no cost

No IT infrastructure needed

Limitations (Public Server)

Not all tools available

200GB storage limit

Input Data

New data set

Public data set

Analysis Workflow

Phase 1: Learn Workflow

Galaxy

Public Galaxy Server (web browser)

Output

FASTQ Files, …

BAM Files

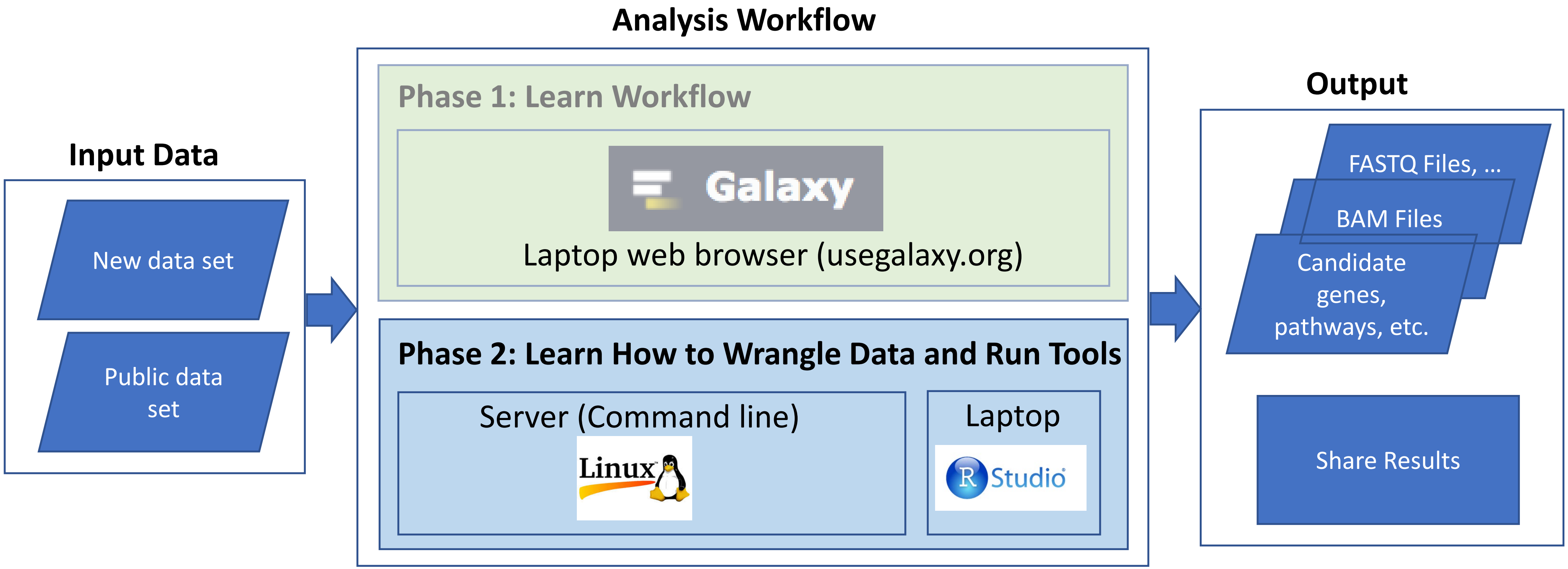
Candidate genes, pathways, etc.

Share Results

MAINE INBRE

IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Bioinformatics Research Training Without Cloud-Based Resources



Input Data

- New data set
- Public data set

Analysis Workflow

Phase 1: Learn Workflow

Galaxy

Laptop web browser (usegalaxy.org)

Phase 2: Learn How to Wrangle Data and Run Tools

Server (Command line)

Linux

Laptop

RStudio

Output

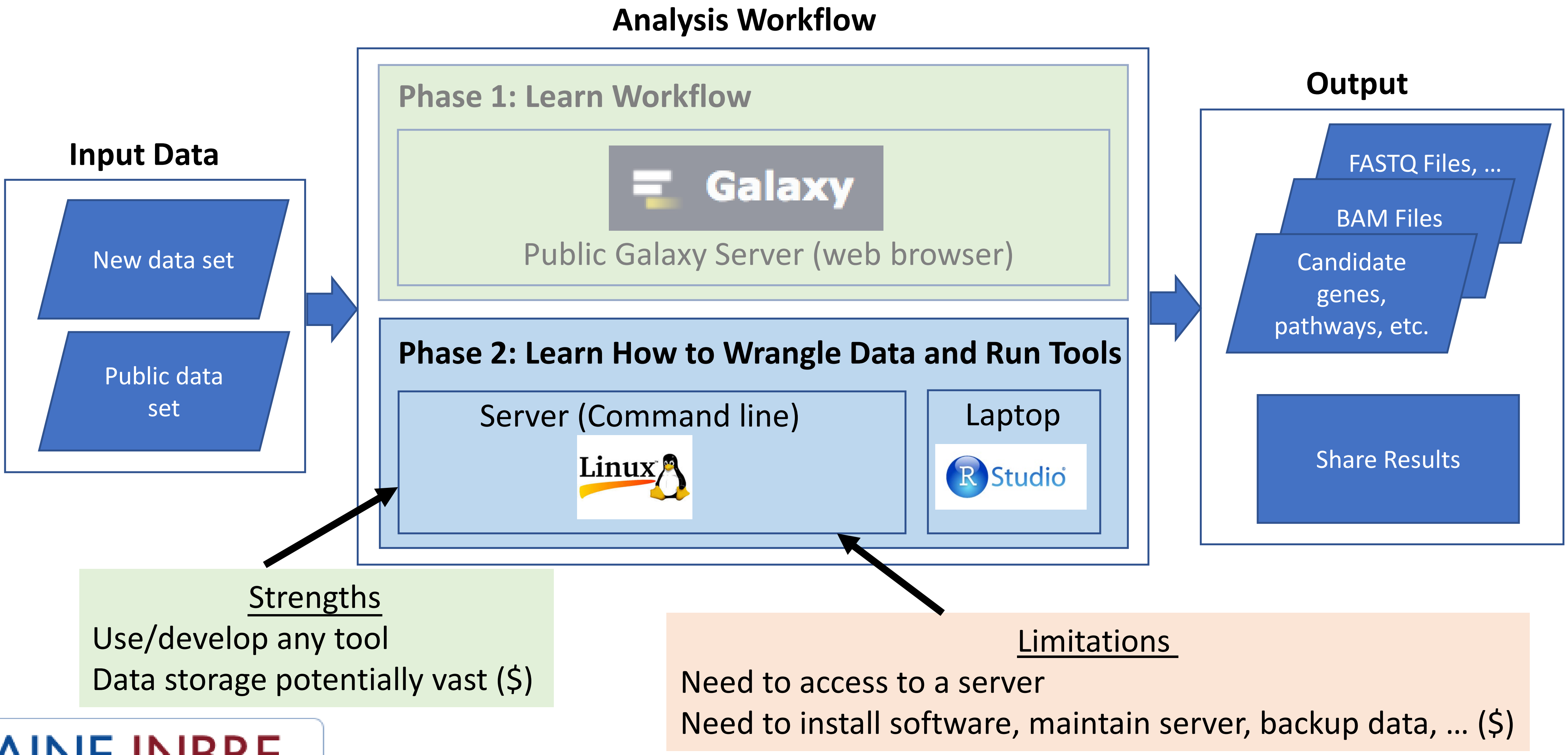
- FASTQ Files, …
- BAM Files
- Candidate genes, pathways, etc.

Share Results

MAINE INBRE

IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Bioinformatics Research Training Without Cloud-Based Resources



Input Data

New data set

Public data set

Analysis Workflow

Phase 1: Learn Workflow

Galaxy

Public Galaxy Server (web browser)

Phase 2: Learn How to Wrangle Data and Run Tools

Server (Command line)

Linux

Laptop

RStudio

Output

FASTQ Files, …

BAM Files

Candidate genes, pathways, etc.

Share Results

Strengths

Use/develop any tool

Data storage potentially vast ($)

Limitations

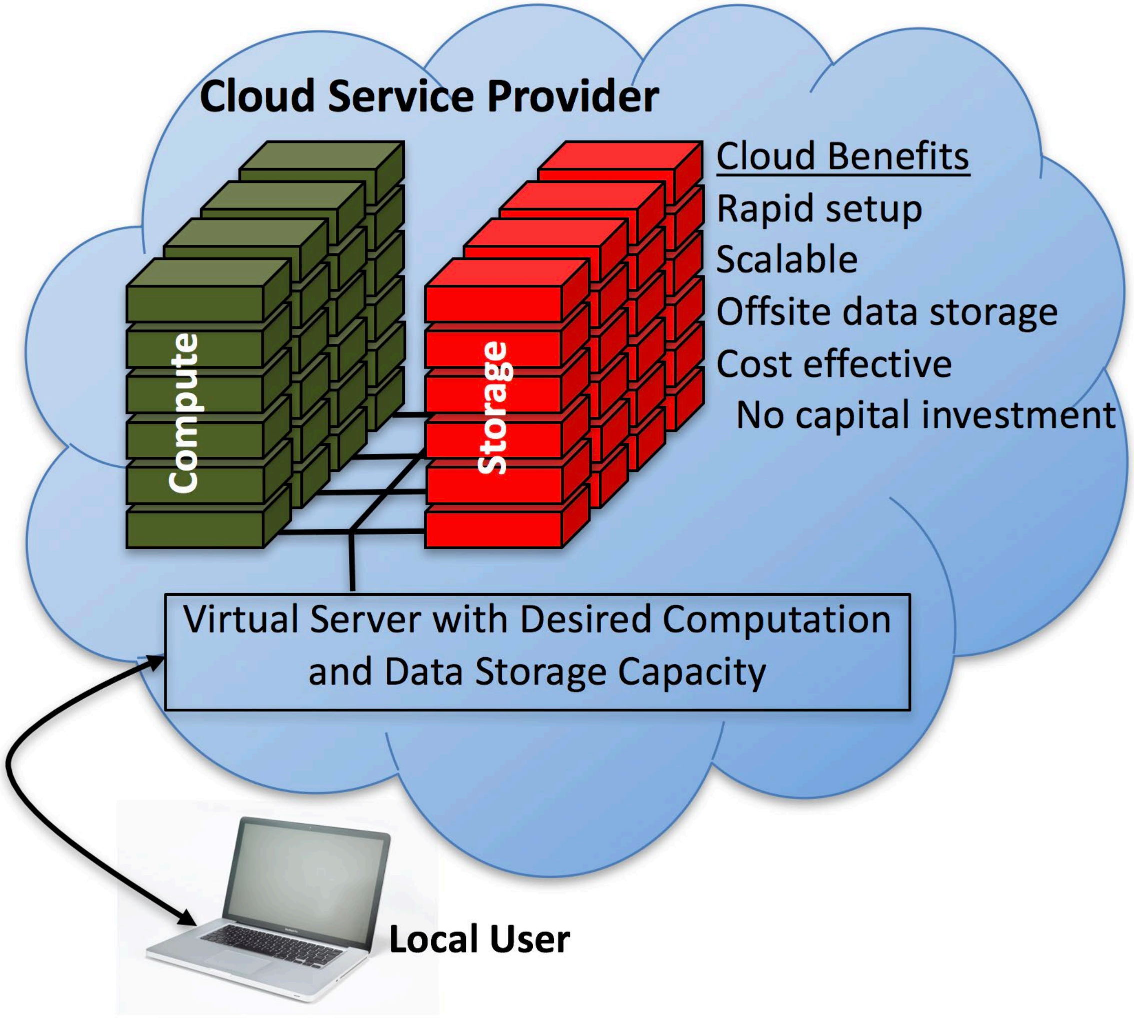
Need to access to a server

Need to install software, maintain server, backup data, … ($)

MAINE INBRE

IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Reusable Cloud-Based Bioinformatics Analysis Environments Can Enhance Bioinformatics Research and Training

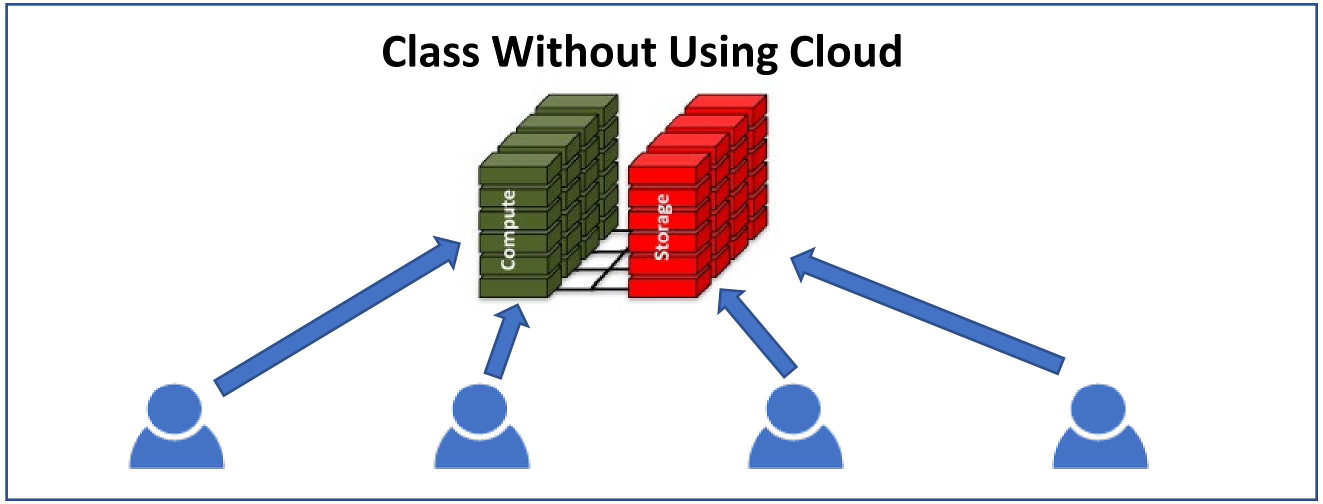


Cloud Service Provider

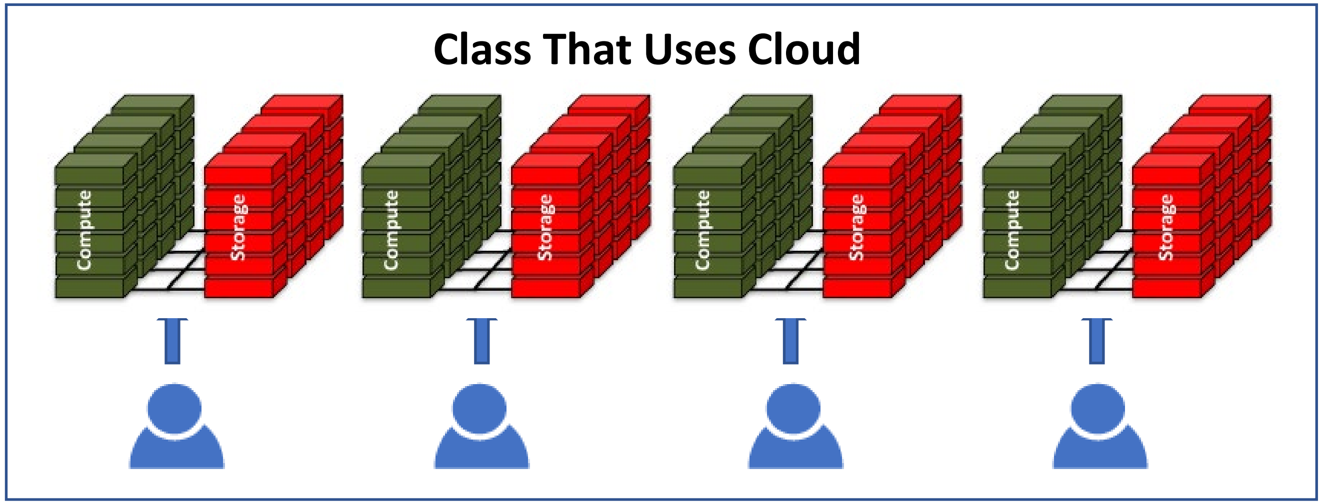
Cloud Benefits
- Rapid setup
- Scalable
- Offsite data storage
- Cost effective
  - No capital investment

Virtual Server with Desired Computation and Data Storage Capacity

Local User



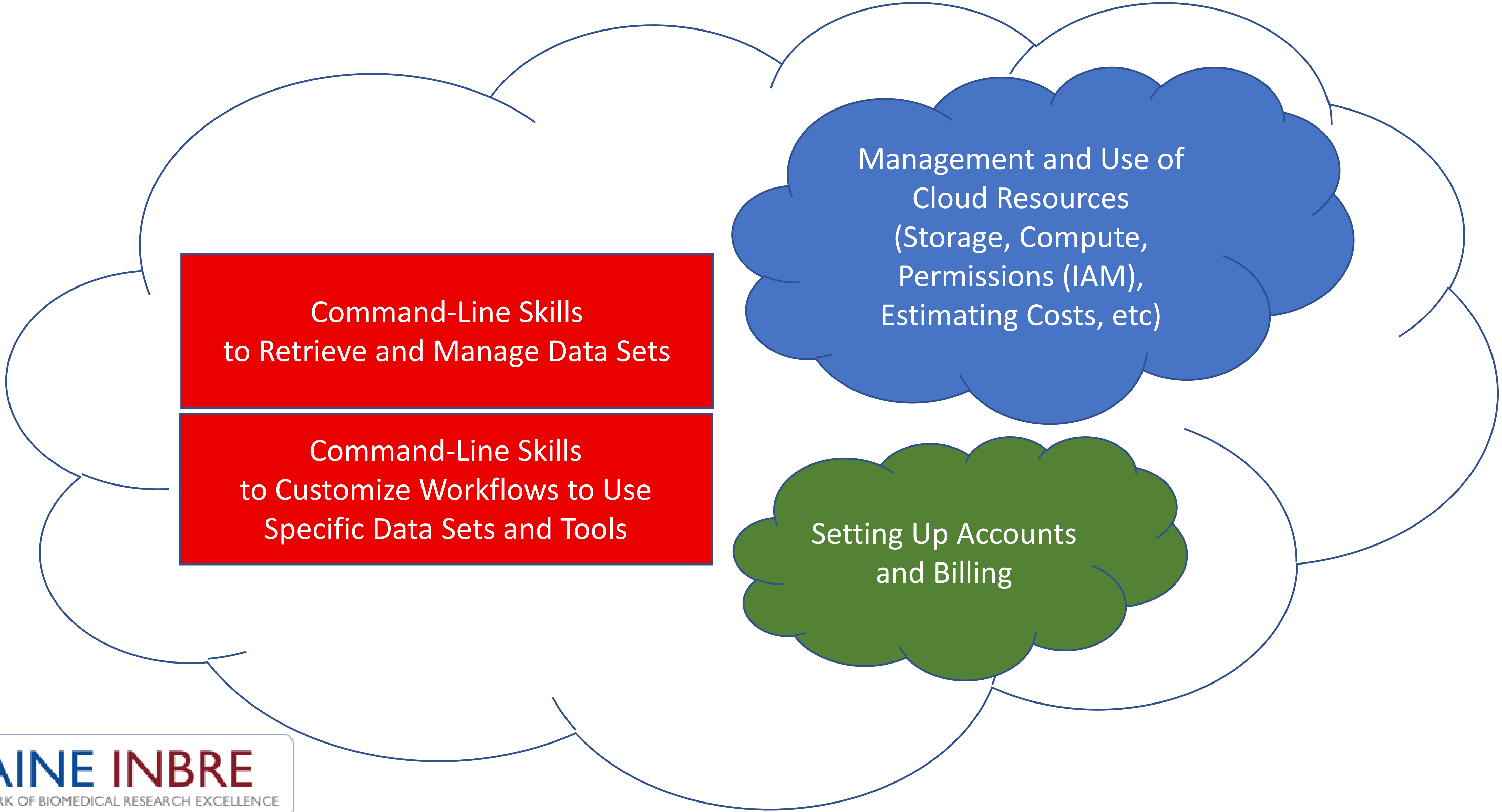
Class Without Using Cloud



Class That Uses Cloud

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Skills Needed to Effectively Use Cloud Computing



Command-Line Skills
to Retrieve and Manage Data Sets

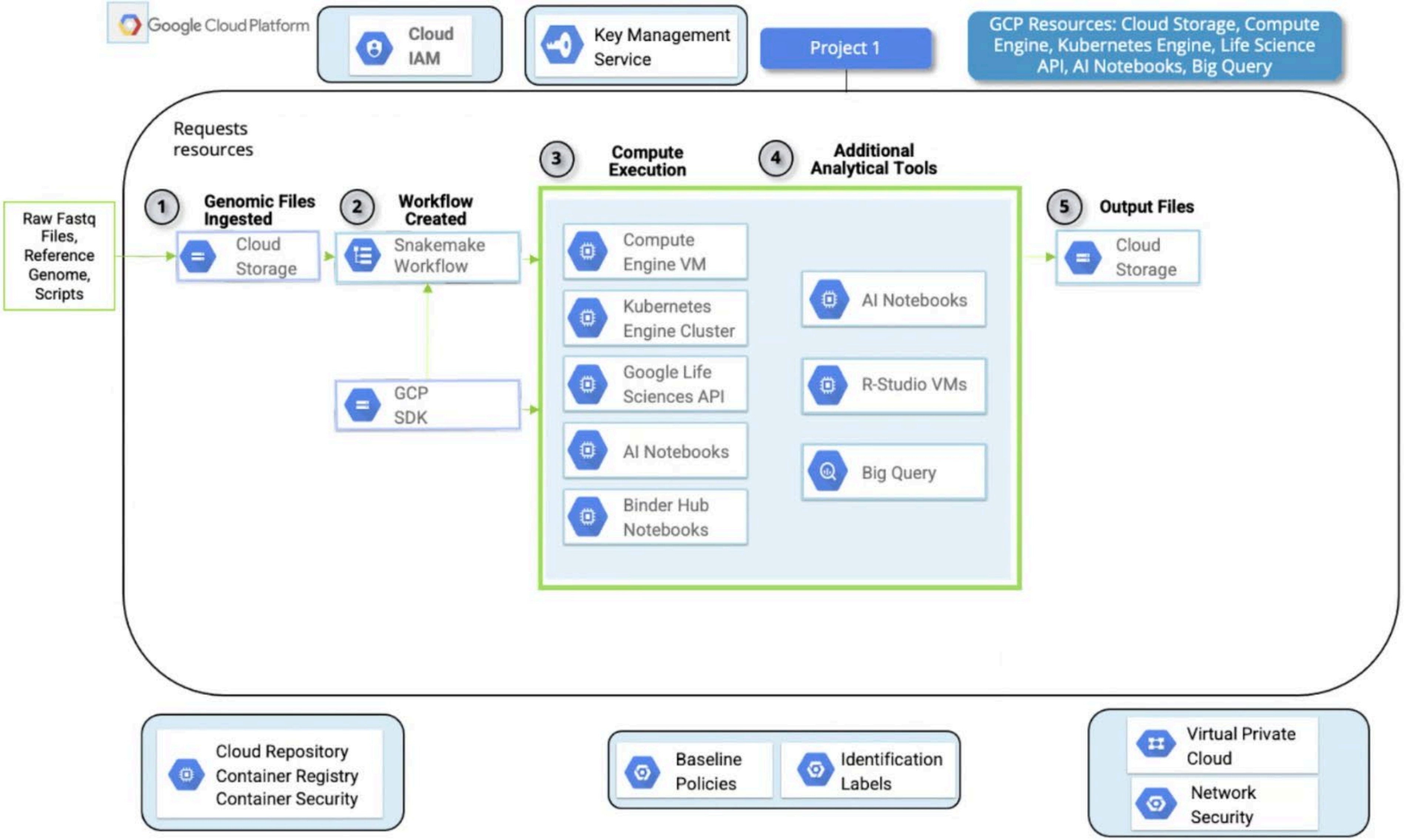
Command-Line Skills
to Customize Workflows to Use
Specific Data Sets and Tools

Management and Use of
Cloud Resources
(Storage, Compute,
Permissions (IAM),
Estimating Costs, etc)

Setting Up Accounts
and Billing

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

RNA Seq Architecture on GCP



Google Cloud Platform

Cloud IAM

Key Management Service

Project 1

GCP Resources: Cloud Storage, Compute Engine, Kubernetes Engine, Life Science API, AI Notebooks, Big Query

Requests resources

Raw Fastq Files, Reference Genome, Scripts

① Genomic Files Ingested

Cloud Storage

② Workflow Created

Snakemake Workflow

GCP SDK

③ Compute Execution

Compute Engine VM

Kubernetes Engine Cluster

Google Life Sciences API

AI Notebooks

Binder Hub Notebooks

④ Additional Analytical Tools

AI Notebooks

R-Studio VMs

Big Query

⑤ Output Files

Cloud Storage

Cloud Repository
Container Registry
Container Security

Baseline Policies

Identification Labels

Virtual Private Cloud

Network Security

Pipeline Stages

① **File Transfer to Cloud Storage**

② **Snakemake Workflow Created**

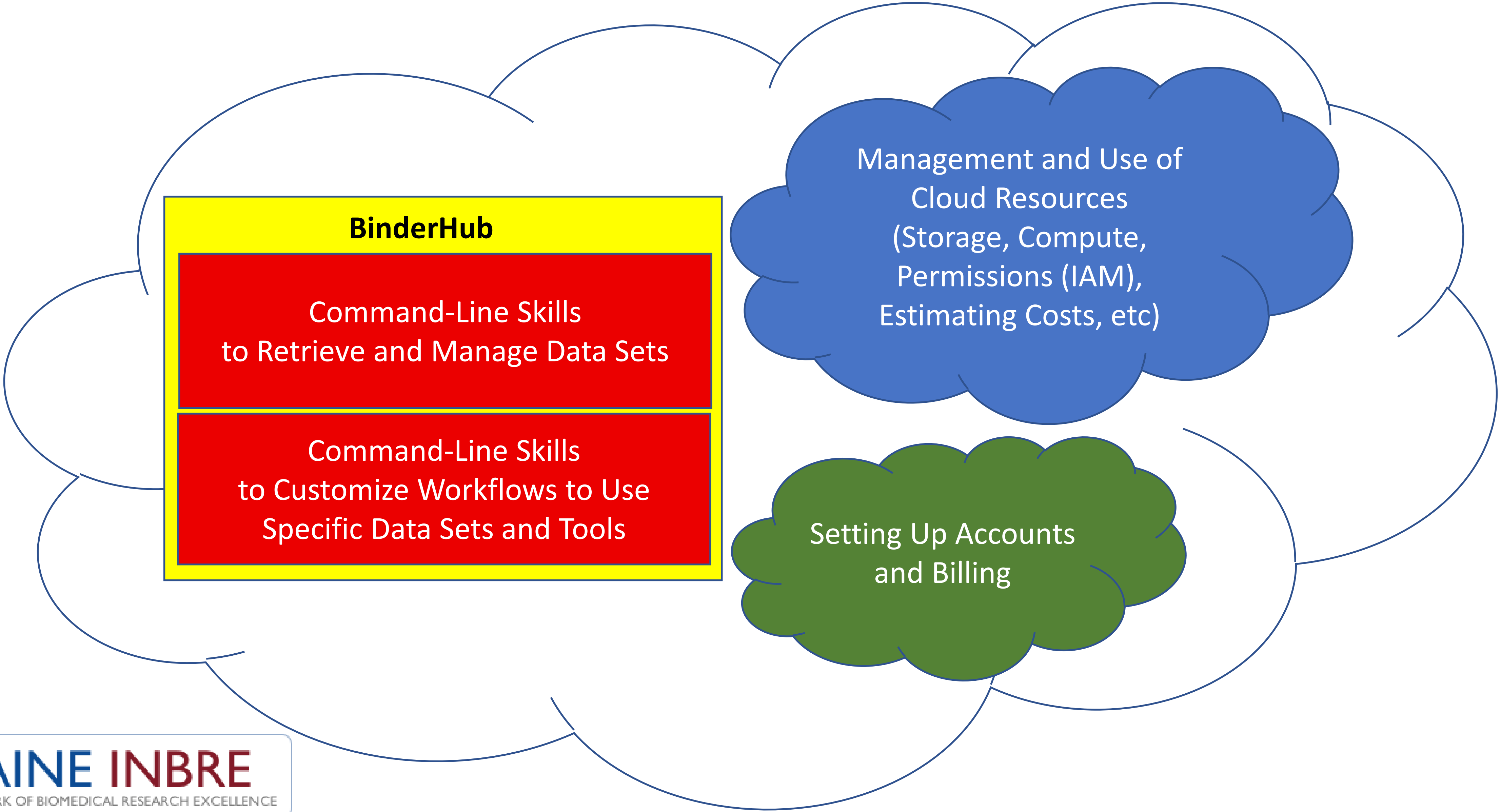
③ **Execution Using Various Resources**

④ **Post Processing Analytical Tools**

⑤ **Finalized Analysis Stored in Cloud Storage**

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

BinderHub Facilitates Training



BinderHub

Command-Line Skills
to Retrieve and Manage Data Sets

Command-Line Skills
to Customize Workflows to Use
Specific Data Sets and Tools

Management and Use of
Cloud Resources
(Storage, Compute,
Permissions (IAM),
Estimating Costs, etc)

Setting Up Accounts
and Billing

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

RNA Sequencing Analysis Training Modules in Development

Basics of Cloud Computing (NIH STRIDES)

How to Upload and Access Data in the Cloud

How to Implement a RNA-Seq Analysis Workflow in the Cloud

Command-Line Training

Learn how to run each step using a Notebook

Workflow Automation Training Using Snakemake

Learn how to run entire workflow using Snakemake with one command.

Training Resources

All materials will be shared on a public Github repository

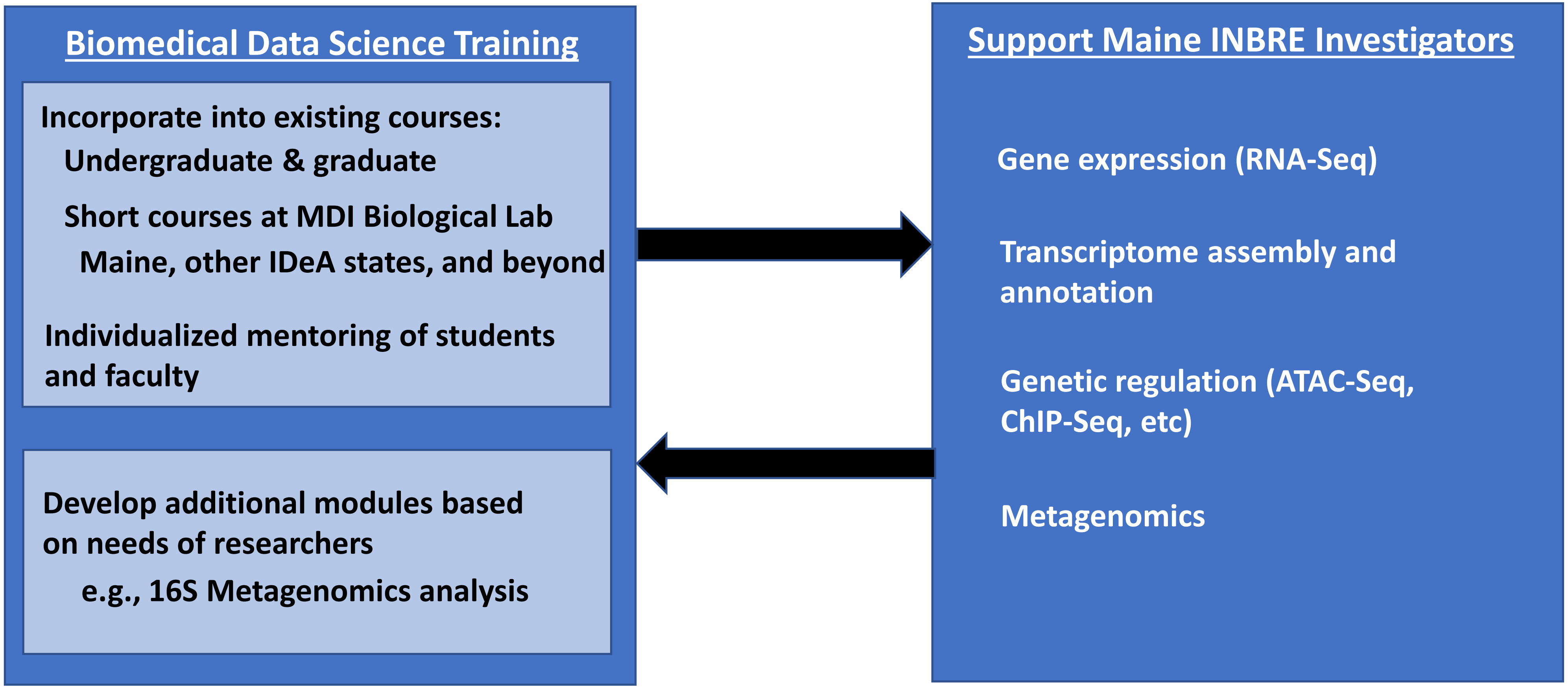
- Notebooks with step-by-step instructions, code, and demonstration data

Easy to run notebooks using Binder

- Provides free access to a small virtual machine

Training videos (under development)

How Training Modules Will Advance Biomedical Data Science Research Training



Biomedical Data Science Training

Incorporate into existing courses:
  Undergraduate & graduate

  Short courses at MDI Biological Lab
    Maine, other IDeA states, and beyond

Individualized mentoring of students and faculty

Develop additional modules based on needs of researchers
  e.g., 16S Metagenomics analysis

Support Maine INBRE Investigators

Gene expression (RNA-Seq)

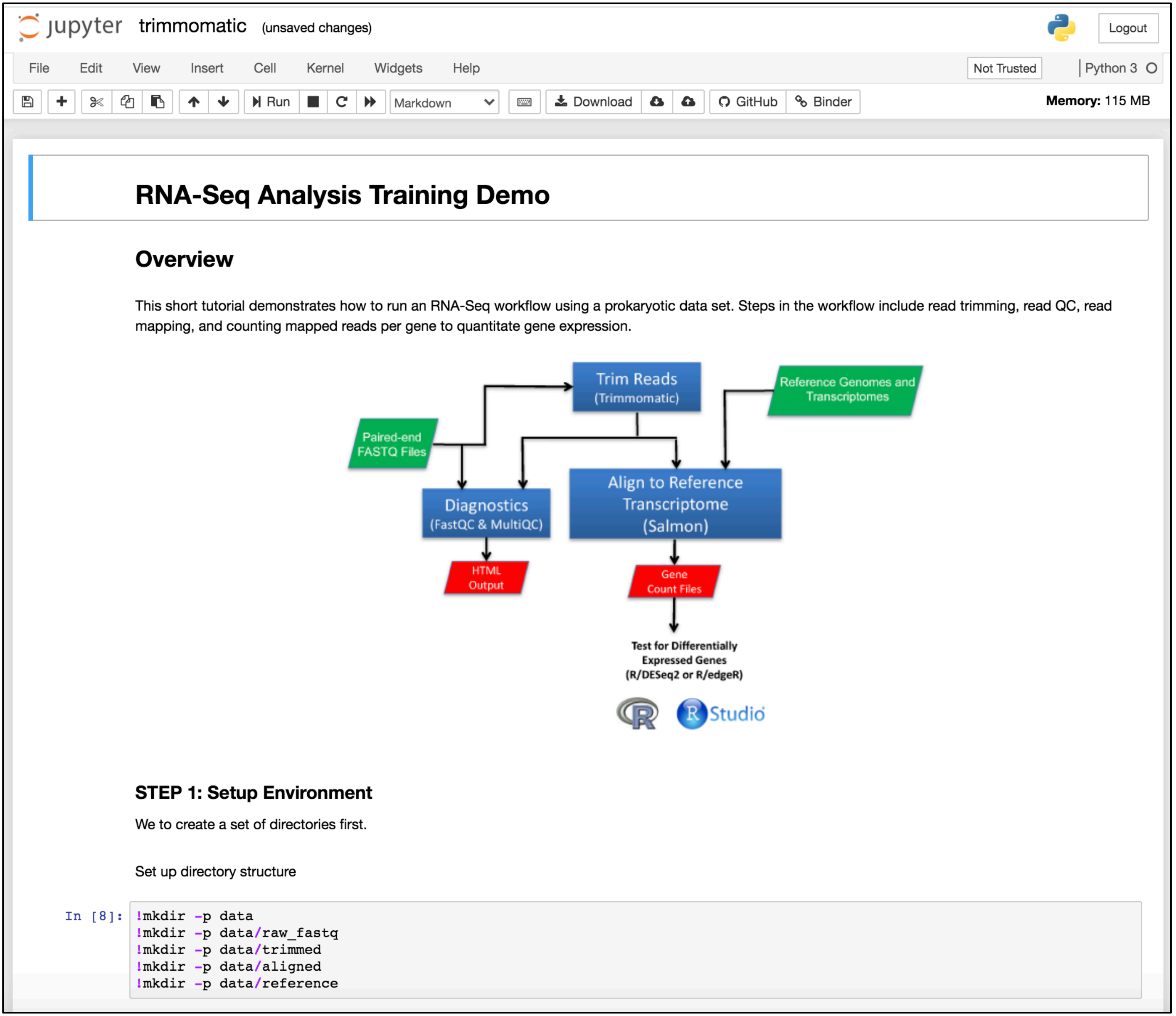
Transcriptome assembly and annotation

Genetic regulation (ATAC-Seq, ChIP-Seq, etc)

Metagenomics

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

Live Demo



RNA-Seq Analysis Training Demo

Overview

This short tutorial demonstrates how to run an RNA-Seq workflow using a prokaryotic data set. Steps in the workflow include read trimming, read QC, read mapping, and counting mapped reads per gene to quantitate gene expression.

STEP 1: Setup Environment

We to create a set of directories first.

Set up directory structure

```
In [8]: !mkdir -p data
        !mkdir -p data/raw_fastq
        !mkdir -p data/trimmed
        !mkdir -p data/aligned
        !mkdir -p data/reference
```

Acknowledgements

NIH National Institutes of Health

Google

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE

NIGMS

Ming Lei
Lakshmi Matukumalli
Krishan Arora

NIH STRIDES

Nick Weber
Matt Gieseke
Tom Shaw
Jared Taylor
Antej Nuhanovic
Joel Mills

Google

Dave Belardo
Phatty Arbuckle
Sam Russ
Asheesh Sharma
Israel Groves
Grant Sorbo
Anne Billak

Deloitte

Kyle O’Conner
Thad Carlson
Ben Kopchick
Dina Mikdadi
Juergen Klenk

Joel Graber
Jim Coffman

C. Titus Brown (UC Davis)

MDIBL
Nathan Maki
Chris Wilson

UMaine
Steven Allers
Sally Molloy
Students in Molloy Lab

MAINE INBRE
IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE