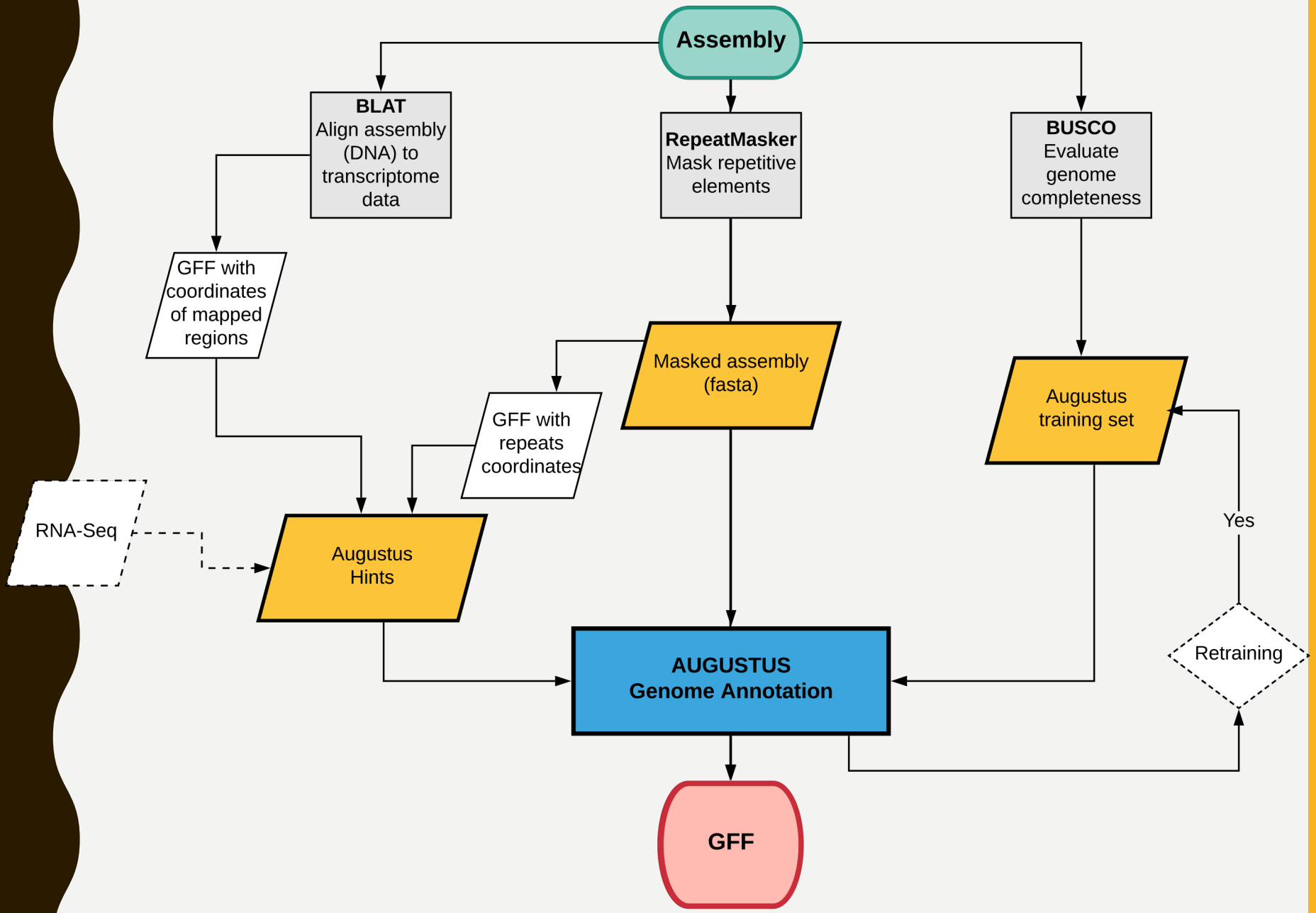
GENOME ANNOTATION WORKSHOP

MIRIAN T. N. TSUCHIYA
DATA SCIENCE POSTDOCTORAL FELLOW
DATA SCIENCE LAB - OCIO

WHERE ARE WE?


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
Masked assembly (fasta)
Augustus training set
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Yes
AUGUSTUS
Genome Annotation
Retraining
GFF

WHAT DID WE DO

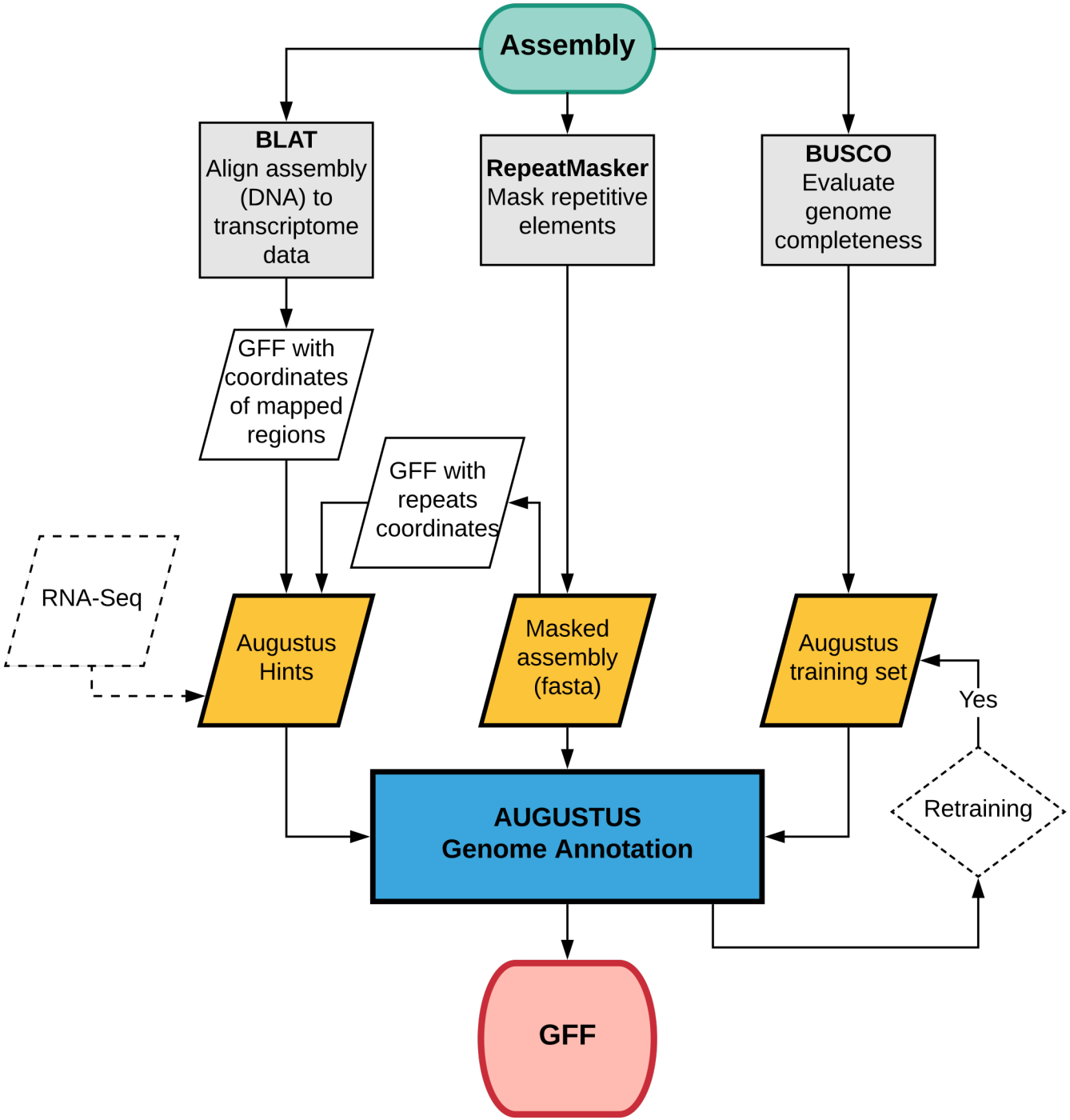
1. BUSCO
    - BUSCO summary ✓
    - Retraining parameters ✓
2. BLAT
    - Hints ✓
3. RepeatMasker
    - Masked assembly 

WHAT ARE WE DOING TODAY?

1. Setting up an embarrassingly parallel AUGUSTUS run
2. Combine the results
3. Visualize it all using JBrowse

AUGUSTUS

FINALLY!


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Masked assembly (fasta)
Augustus training set
Yes
Retraining
AUGUSTUS
Genome Annotation
GFF

AUGUSTUS

- ab initio (internal) + evidence-driven(external)

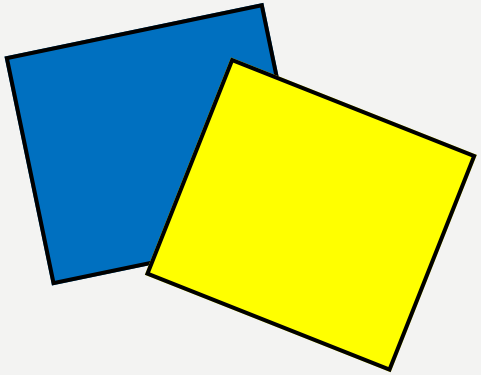
AUGUSTUS is based on a generalized hidden Markov model (GHMM), which defines probability distributions for the various sections of genomic sequences. Introns, exons, intergenic regions, etc. correspond to states in the model and each state is thought to create DNA sequences with certain pre-defined emission probabilities.

REQUIRED FILES

1. Masked assembly ✓
2. Hints file ✓
3. Retraining parameters ✓
4. Extrinsic file ✓

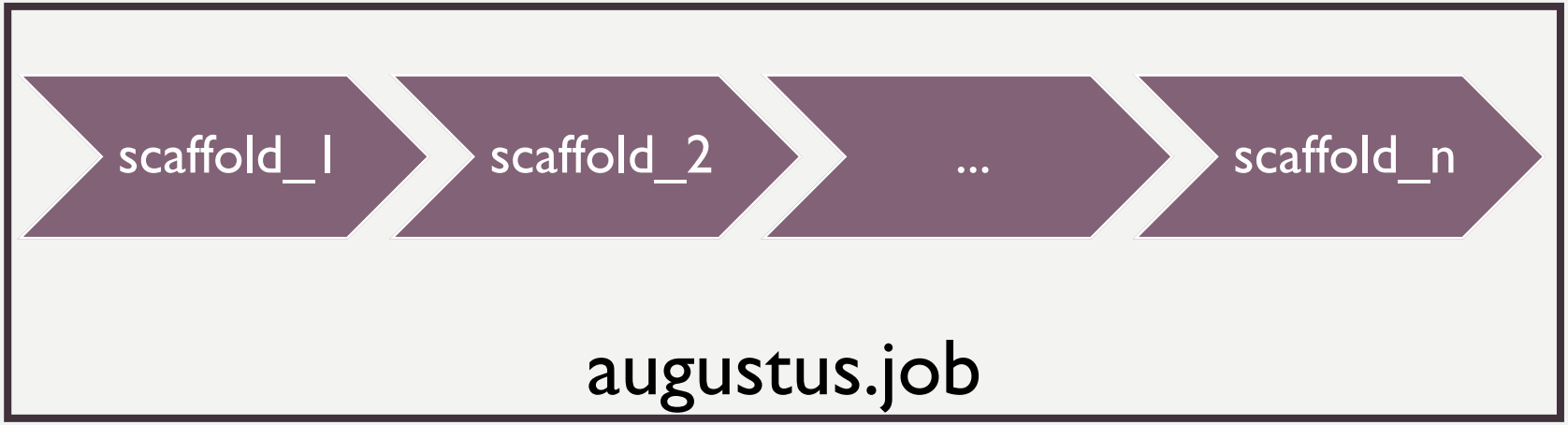
TASKS

1. Copy the masked assembly to your augustus folder
2. Download and extract EVidenceModeler to your augustus folder



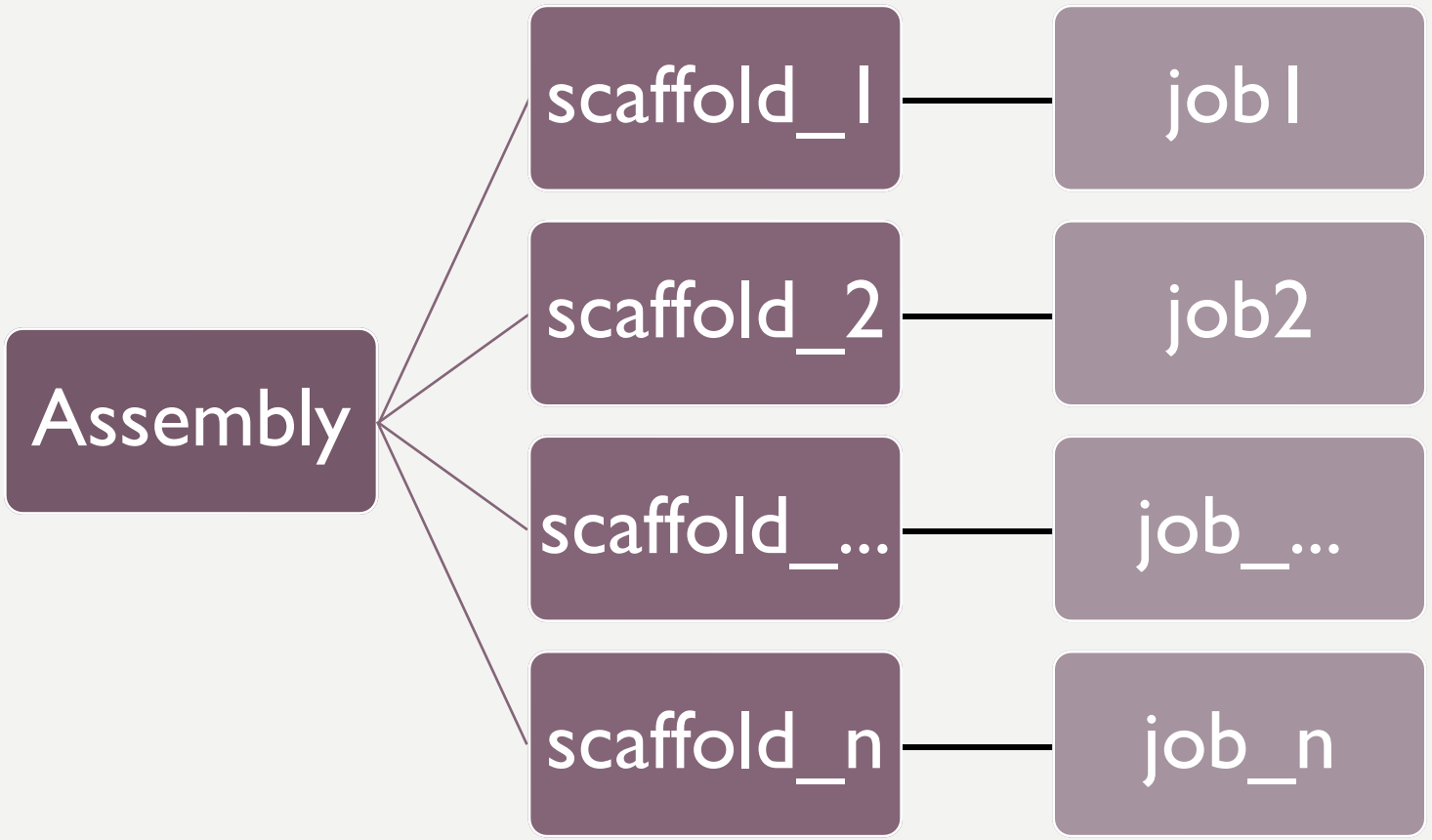
EMBARRASSINGLY PARALLEL

- AUGUSTUS runs serially (aka one scaffold at a time)



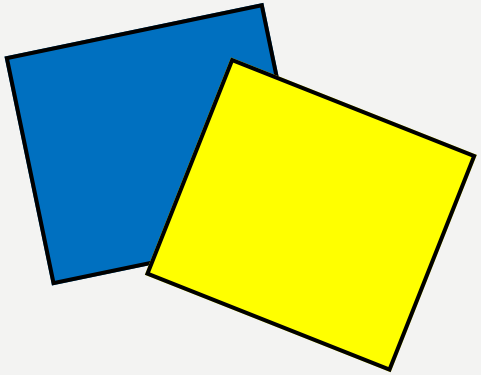
EMBARRASSINGLY PARALLEL

- But we can "force" it to run in parallel



TASKS

- Login to the interactive queue
- Run the EVM script `partition_EVM_inputs.pl` from your scaffolds folder
  - Don't forget that you copied the masked assembly to your augustus folder. Adjust the paths accordingly.



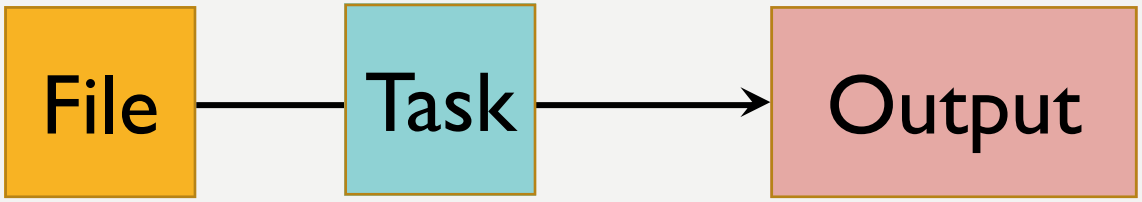
SCAFFOLDS FOLDER

- How many folders do we have?
- Take a look in the folder scaffold_110.
  - What's inside?
  - Use less to look at the files.

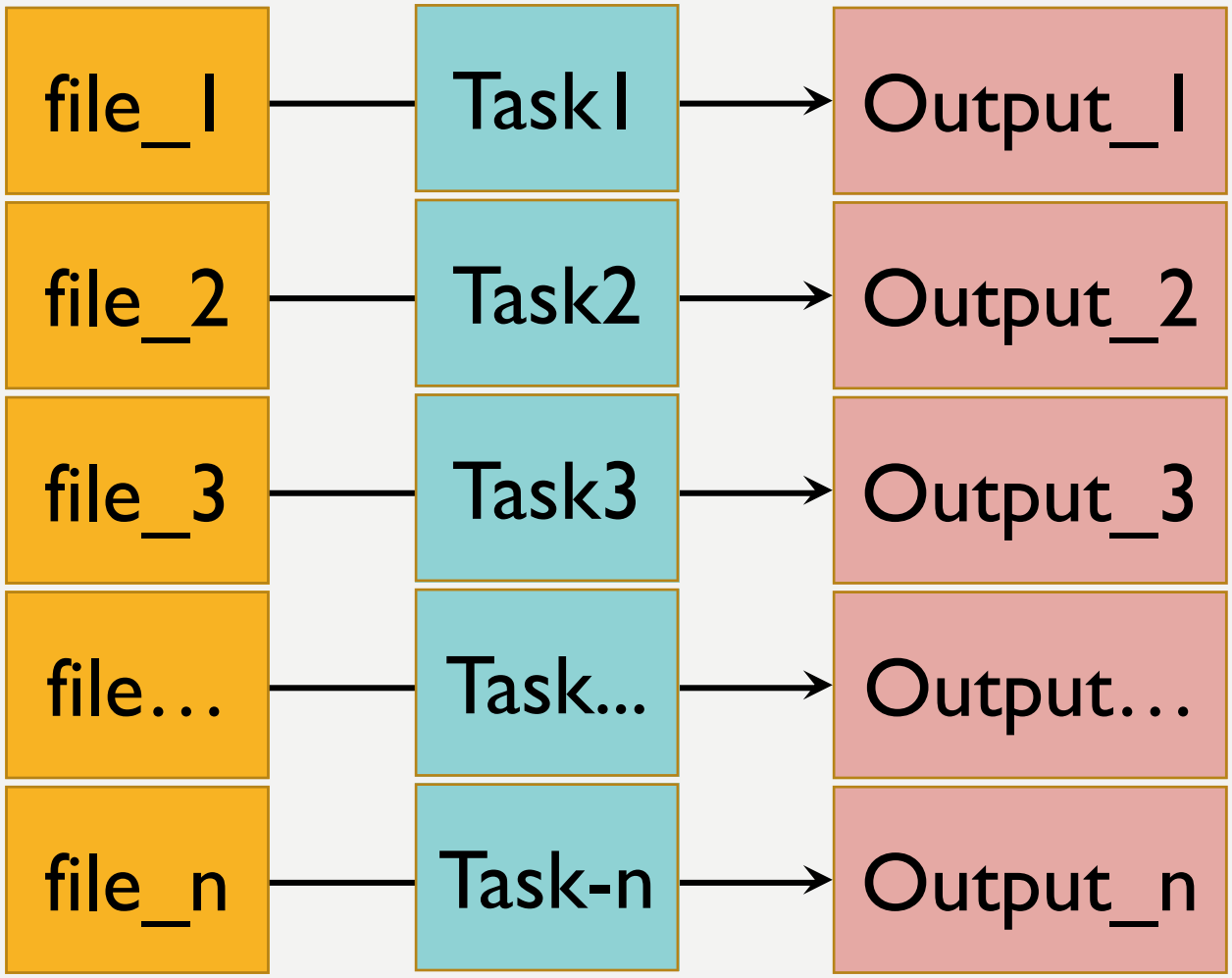
AUGUSTUS JOB FILE

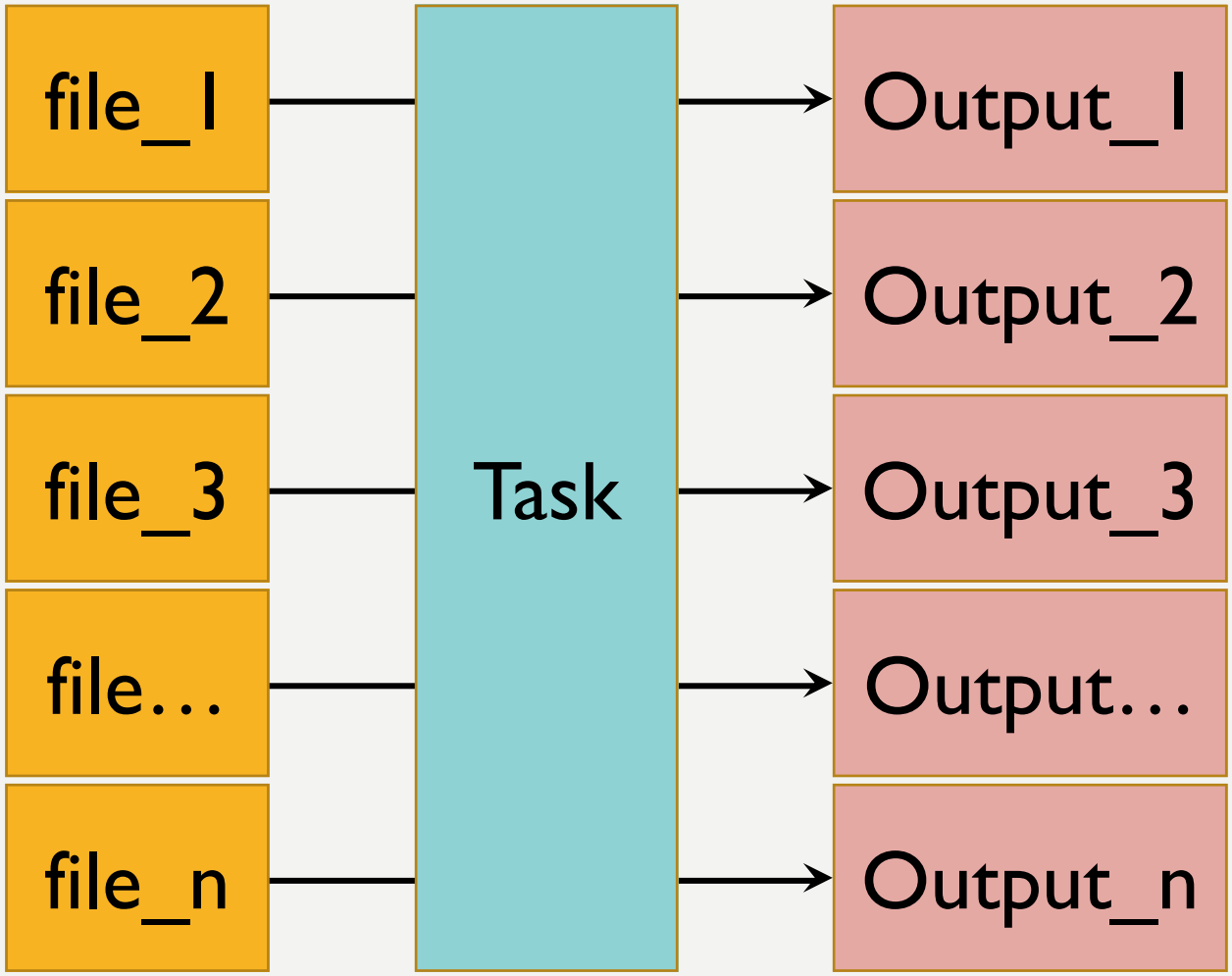
- How are we running augustus on each scaffold?
  - Option 1: Create one job file for each scaffold… manually
  - Option 2: Create one job file and use a loop to submit it.

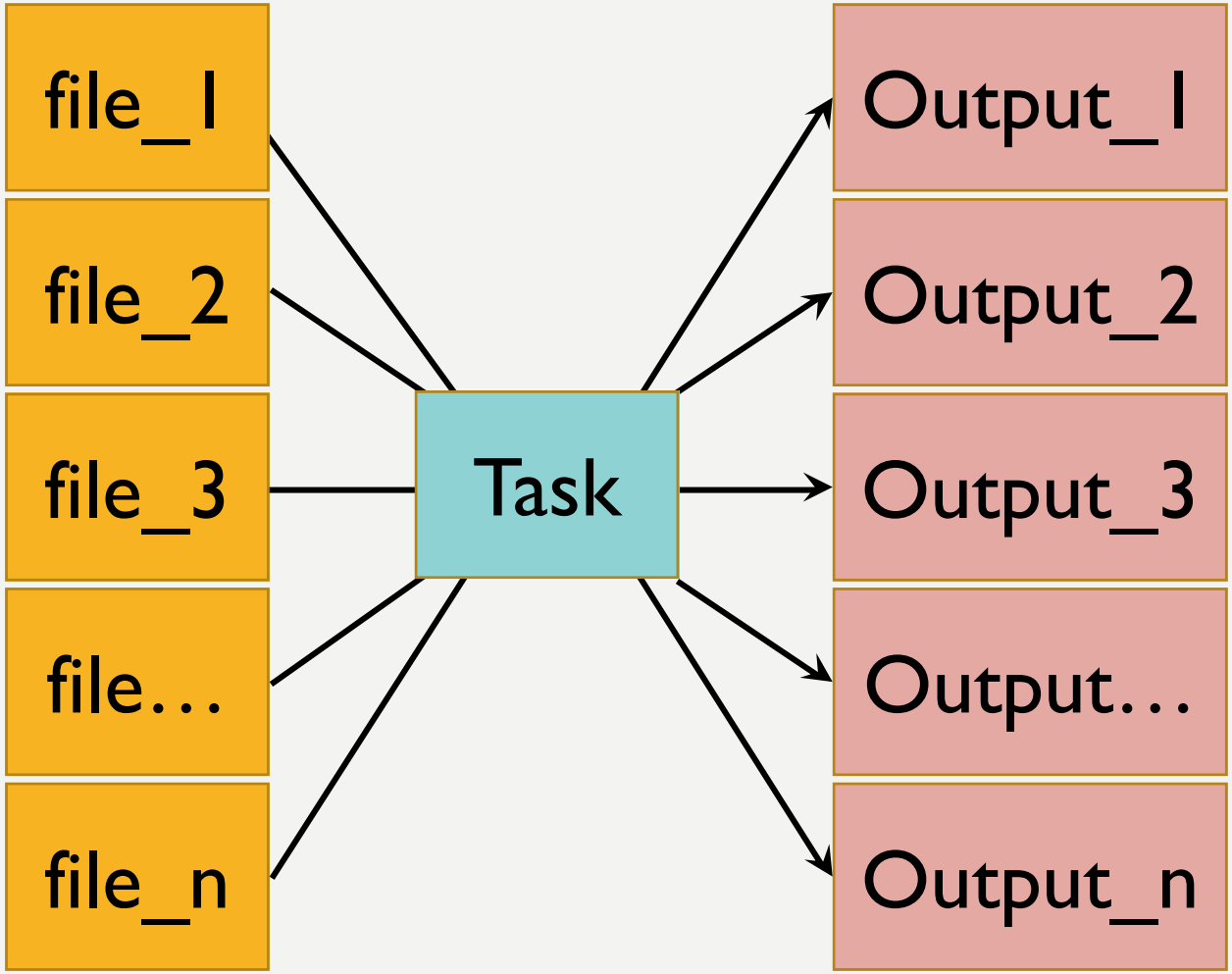
BRIEF INTRO ABOUT LOOPS



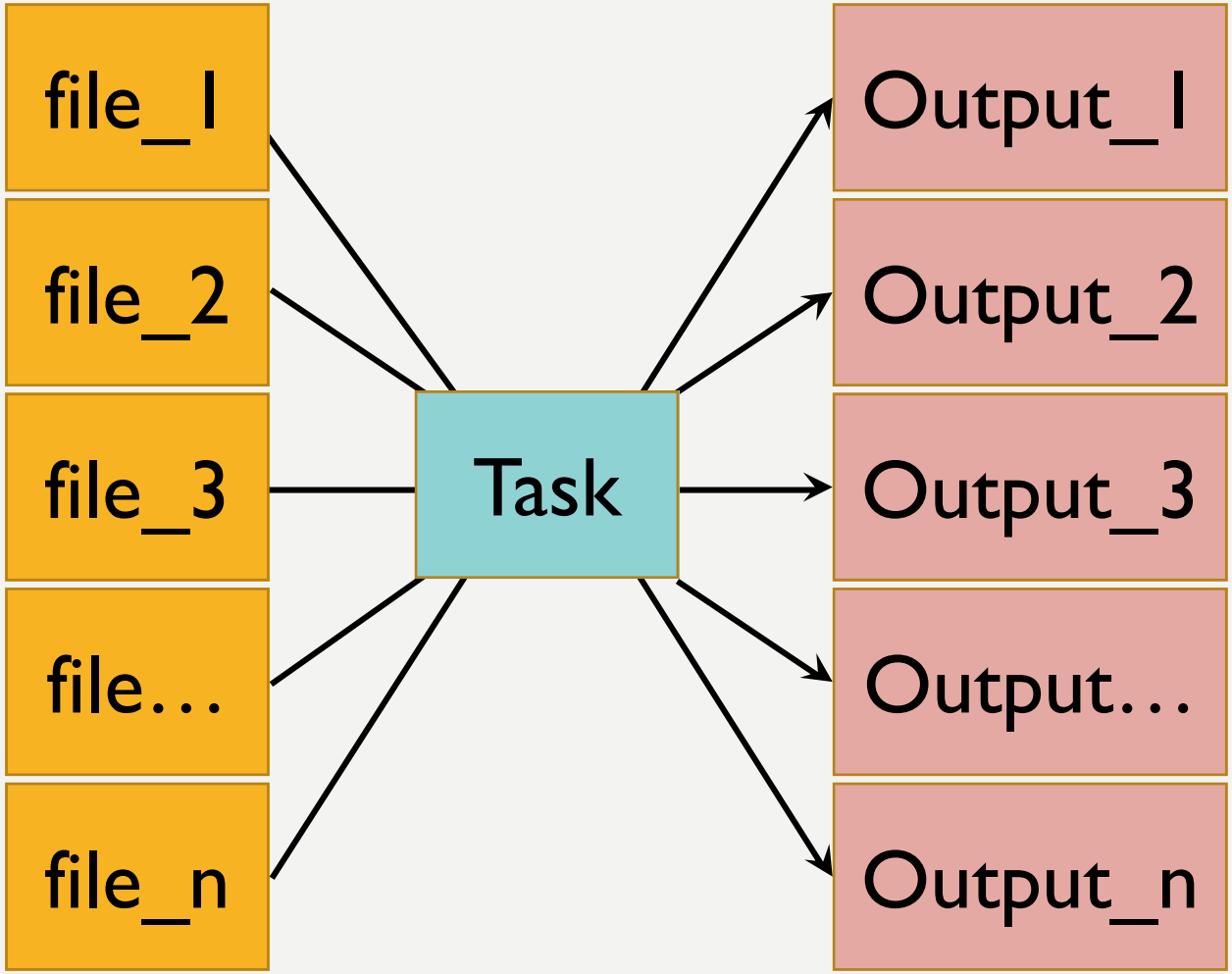
One file = One job


file_1
Task1
Output_1
file_2
Task2
Output_2
file_3
Task3
Output_3
file…
Task...
Output…
file_n
Task-n
Output_n


file_1
file_2
file_3
file…
file_n
Task
Output_1
Output_2
Output_3
Output…
Output_n


file_1
file_2
file_3
file…
file_n
Task
Output_1
Output_2
Output_3
Output…
Output_n

for f in file_*; do task > ${f}.output; done



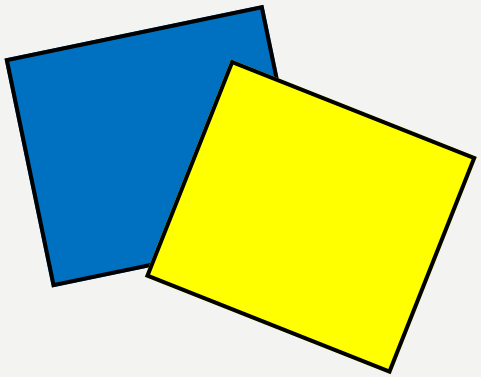
```
FOR DOG IN PUPPY_*; DO
   PLAY ${DOG} > ${DOG}.HAPPY;
DONE
```

BACK TO AUGUSTUS

- We will create one job (augustus.job) and will submit it using a for loop that will iterate over each scaffold - all at the same time.

TASKS

- Create the augustus job and save it in the jobs folder.
- You will submit the job from the folder scaffolds.

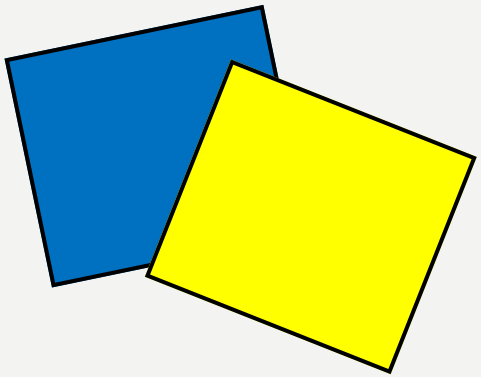


AUGUSTUS RESULTS

- List the files in the folder outputs. You should have one gff file per scaffold.
- Use cat or less to visualize file contents.

COMBINING THE RESULTS

- Now we want to combine all the gffs into a single gff file. Dhydei_augustus_all.gff3
- We will use the script `join_aug_pred.pl` from AUGUSTUS to combine the results.



FROM THE FOLDER AUGUSTUS

cp -r /data/genomics/workshops/Gaworkshop/augustus/outputs outputs2

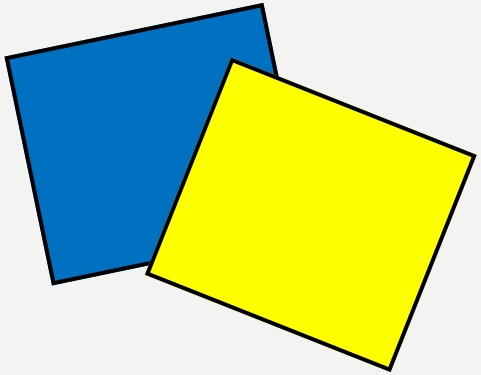
JBROWSE

JBROWSE


Available Tracks
filter tracks
BAM 8
BAM - paired-read test pattern at 28kb
BAM - volvox-longreads.bam
BAM - volvox-sorted SNPs/Coverage
BAM - volvox-sorted.bam
BAM - volvox-sv.bam
CRAM - volvox-longreads.cram
CRAM - volvox-sorted.cram
CRAM - volvox-sv.cram
Miscellaneous 15
BEDTabix - volvox.bed in-memory adaptor
BigBed - volvox genes
ChromHMM
FromConfig - Features in Configuration
GFF3 - sub-subparts
GFF3 - volvox.gff3 in-memory adaptor
GFF3Tabix - volvox.gff3 in-memory adaptor
GTF - volvox.gtf in-memory adaptor
HTMLFeatures - ESTs
HTMLFeatures - Example Features
HTMLFeatures - Example motifs
HTMLFeatures - Features with right-click menus
HTMLFeatures - Fingerprinted BACs
HTMLFeatures - Name test track has a really long track label
HTMLFeatures - SNPs
Quantitative 4
Density 2
BigWig Density - volvox_microarray
BigWig Density - volvox_sine
XY Plot 2
BigWig XY - volvox_microarray
BigWig XY - volvox_sine
Reference sequence 1
Reference sequence
Transcripts 5
Genome
Track
View
Help
Volvox Example
Share
ctgA
ctgA:4411..6492 (2.08 Kb)
Go
Reference sequence
Zoom in to see sequence
CanvasFeatures - transcripts
BigWig Density - volvox_microarray
BigWig XY - volvox_microarray
mean
VCF - volvox-sorted variants
SNV C -> G
BAM - volvox-sorted SNPs/Coverage
BAM - volvox-sorted.bam

TASKS

- Run the JBrowse scripts on your files. You need:
  - Assembly
  - GFF (final)
- Compress the file and copy it to your Desktop.



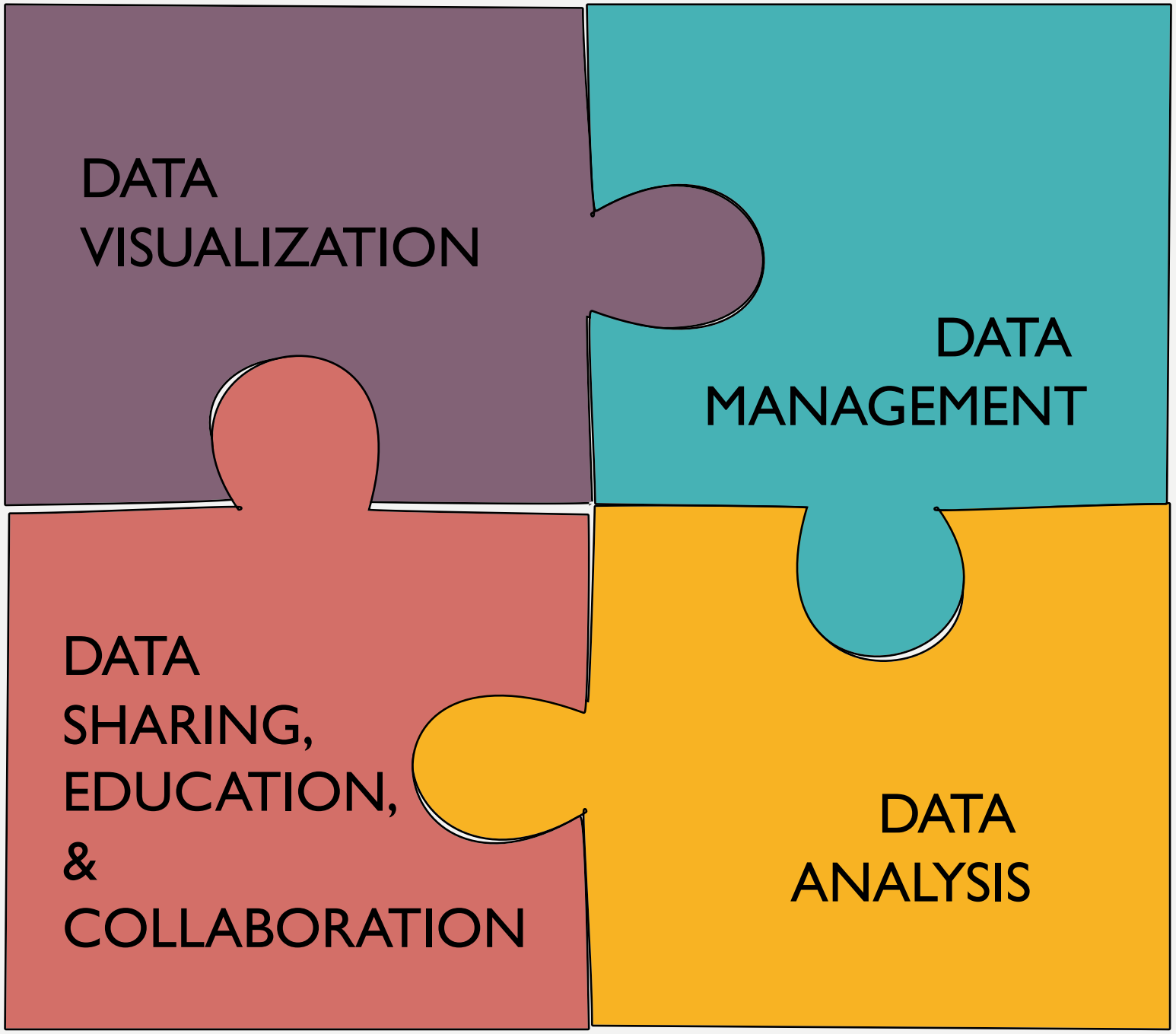
JBROWSE

- You can deploy an instance locally (from your computer) or you can use a cloud service (AWS, Azure, etc)

SMITHSONIAN GENOME HUB

A CLOUD BASED COLLABORATIVE TOOL FOR BIODIVERSITY SCIENTISTS

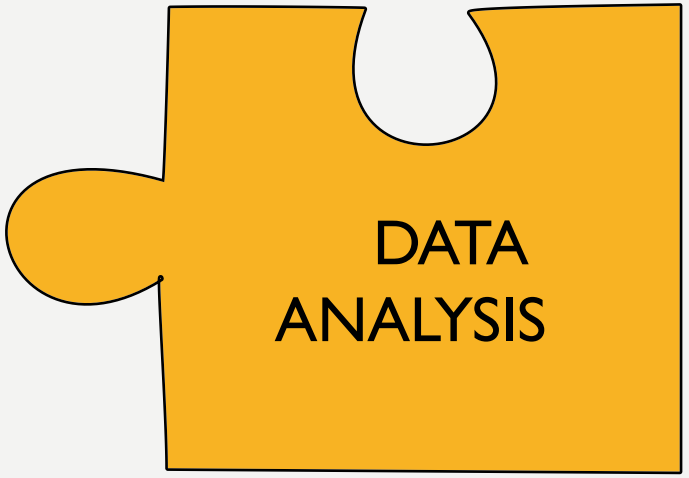
species
flies evolution fungi
endangered
birds lichens
conservation
biodiversity
invertebrates
plants butterflies fishes
amphibians
non-model
lichens
mammals


DATA VISUALIZATION
DATA MANAGEMENT
DATA SHARING, EDUCATION, & COLLABORATION
DATA ANALYSIS

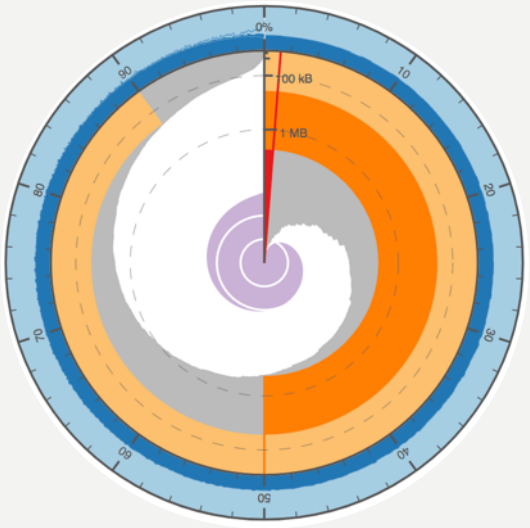
- User-friendly web interface for data analysis
- Pipelines/workflows are easy to create and share
- Widely used software = Galaxy tools
- WebApollo for manual annotation

Galaxy PROJECT

Web Apollo

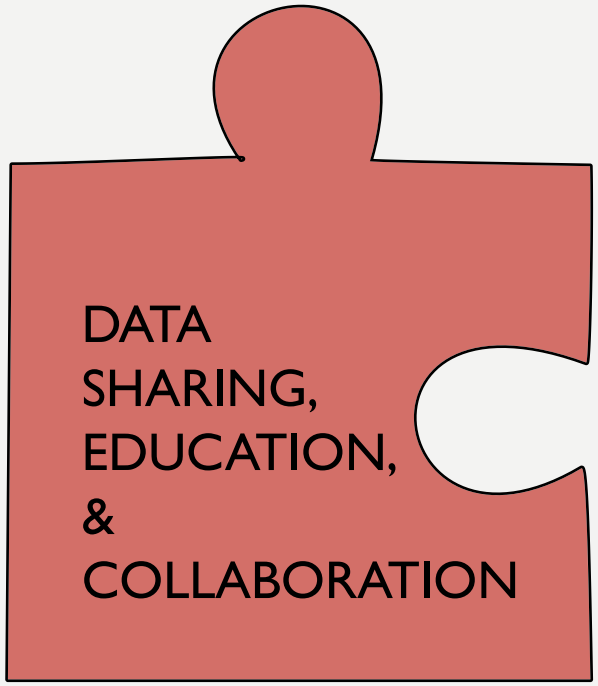


DATA VISUALIZATION



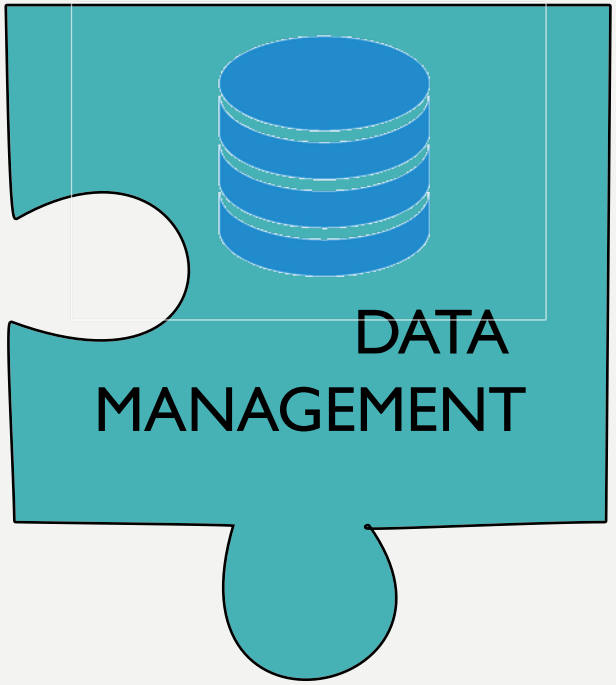
- Individual and institutional project dashboards
- Genome browsers
- Assembly stats and other graphs useful for publication

- Facilitates collaboration and data sharing among researchers
- Creates learning opportunities for students and researchers

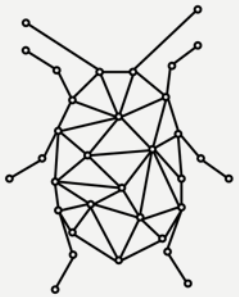


- Promotes diffusion of knowledge to the general public

- Repository of all *de novo* genome assemblies and projects
  - Reporting
  - Managing
  - Organization




Smithsonian's
MISS PEREGRINE'S
HOME FOR
PECULIAR CHILDREN
genomes



OCIO
DATA
SCIENCE
LAB

@SIDatascience

datascience.si.edu

tsuchiyam@si.edu

@MirianTsuchiya

Hydra help: SI-HPC@si.edu

Bug #415

Bug #416

Bug #417

Bug #418

Bug #419

Bug #420