Condor as a tool to study the functions of plant genes

Logan Johnson & Edgar Spalding

Department of Botany
University of Wisconsin

A Bit of Genetic Background

- A major goal in biology is to learn the function of each gene in an organism.
- A proven approach is to compare the behaviors of individuals possessing different versions of that gene.
- Organisms have on the order of 10^4 genes so that makes for a lot of comparisons.

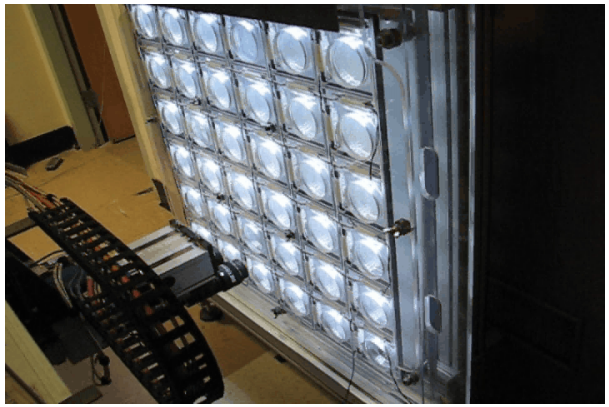
24,999 genes

25,000 genes

Root Gravitropism

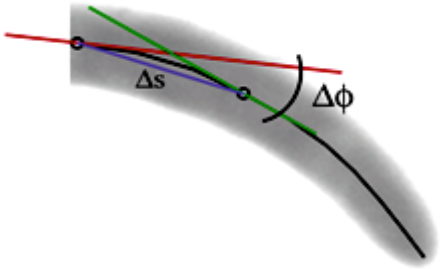
A process we could study with high resolution, high accuracy, and high throughput

Our Automated Workflow

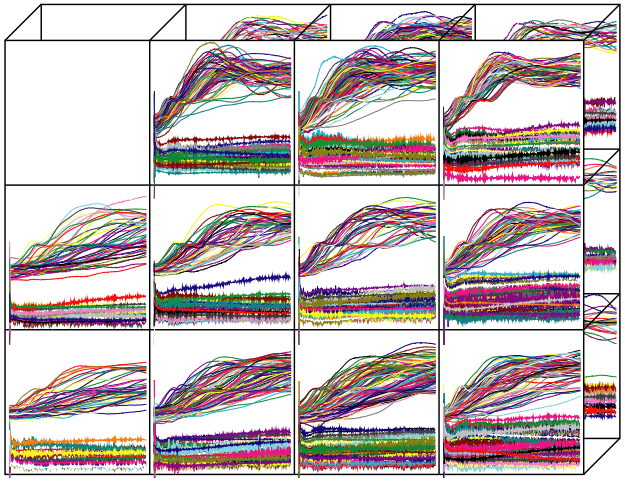


automated image acquisition generates movies

algorithmic feature extraction and quantification operates on each object…



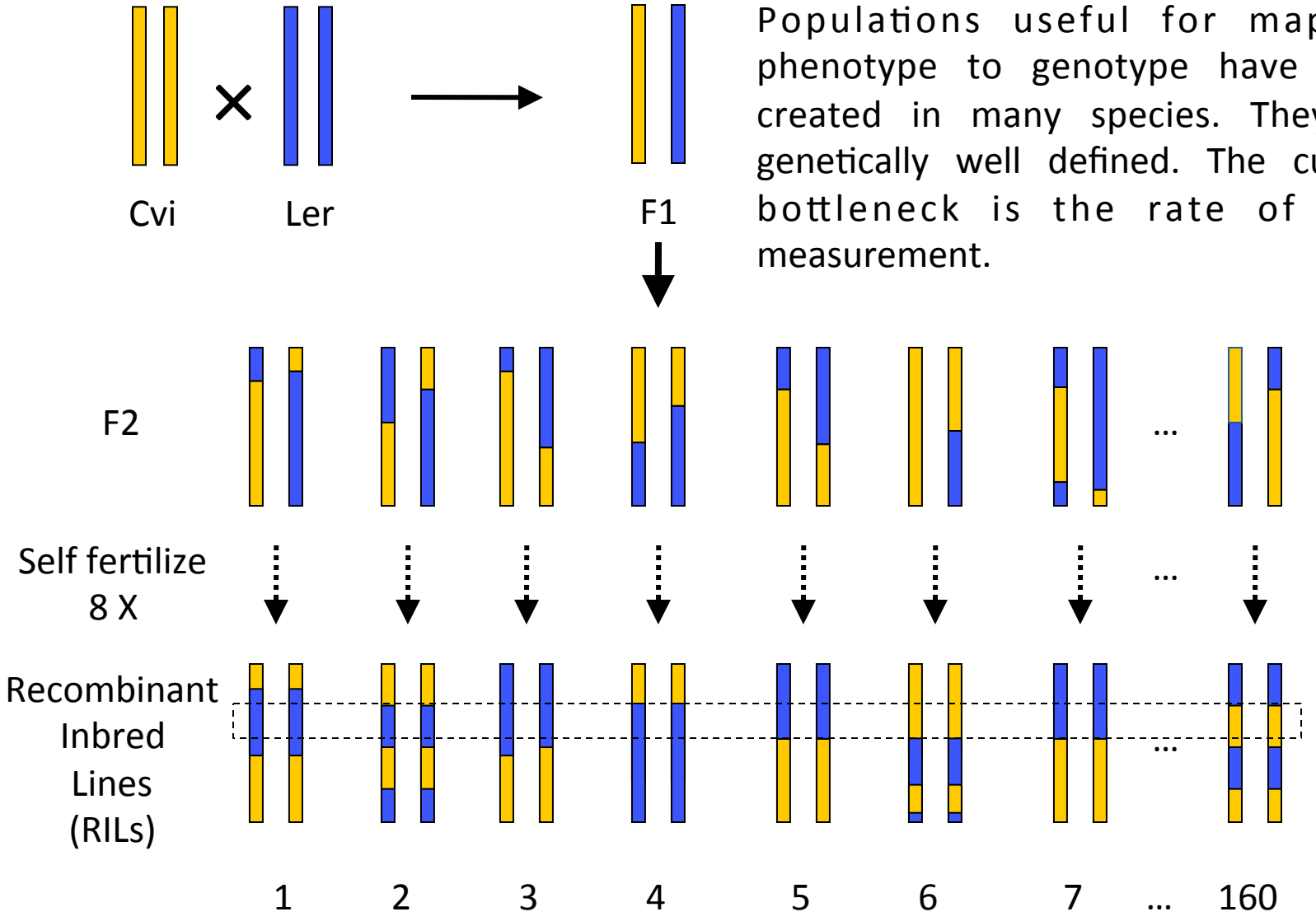
…resulting in large, high dimension data sets…



data mining

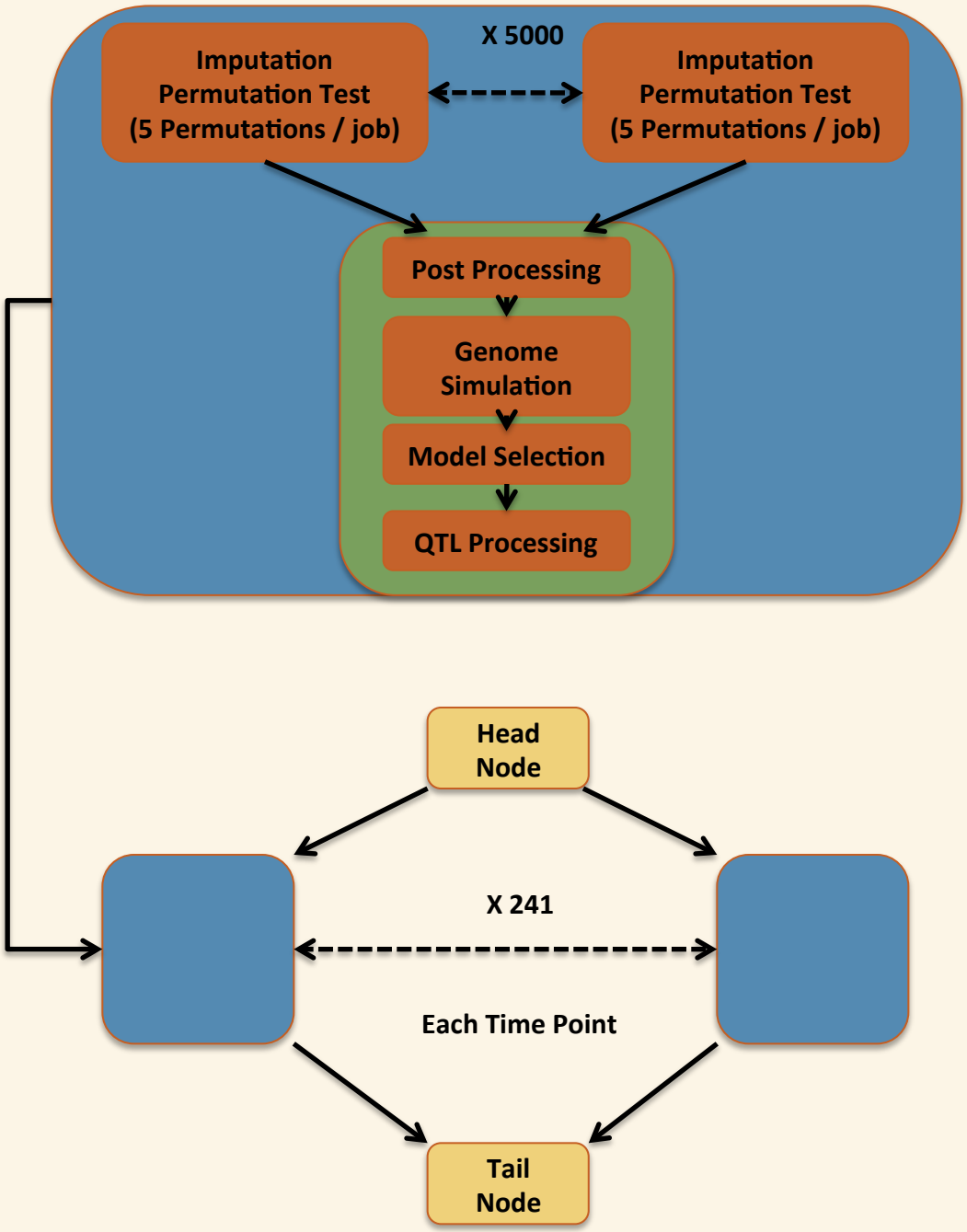
test new hypotheses

Why not / How to scale up to the whole genome level?



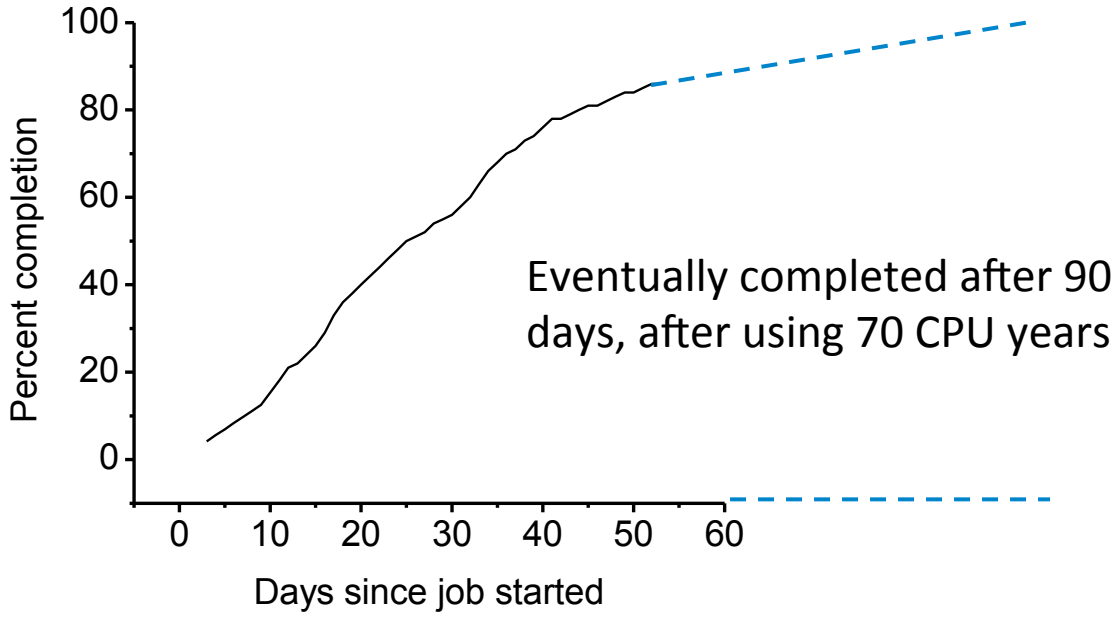
Populations useful for mapping phenotype to genotype have been created in many species. They are genetically well defined. The current bottleneck is the rate of trait measurement.

Alonso-Blanco *et al. The Plant Journal* **14**, 259-257 (1998)

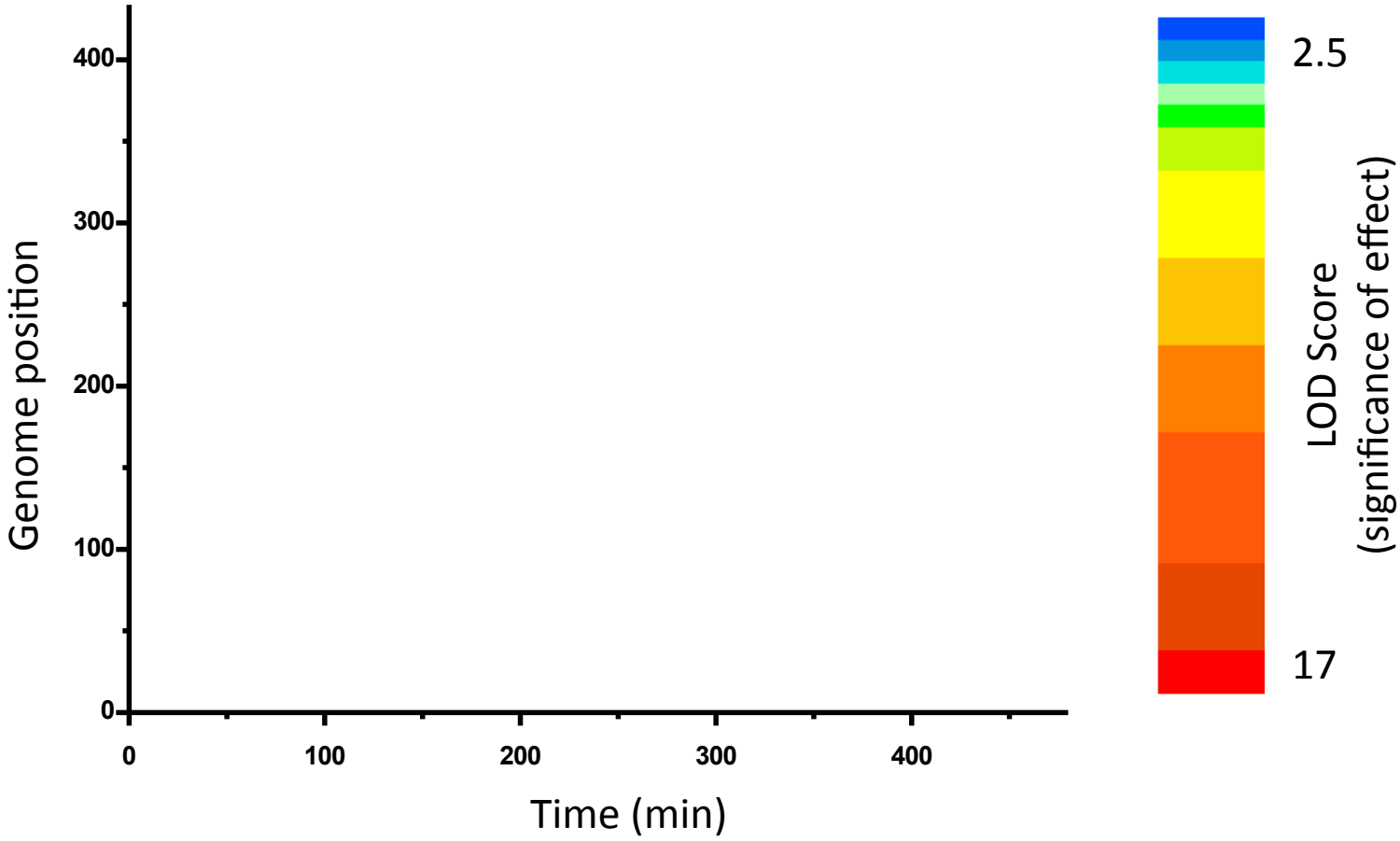

X 5000
Imputation
Permutation Test
(5 Permutations / job)
Imputation
Permutation Test
(5 Permutations / job)
Post Processing
Genome
Simulation
Model Selection
QTL Processing
Head
Node
X 241
Each Time Point
Tail
Node

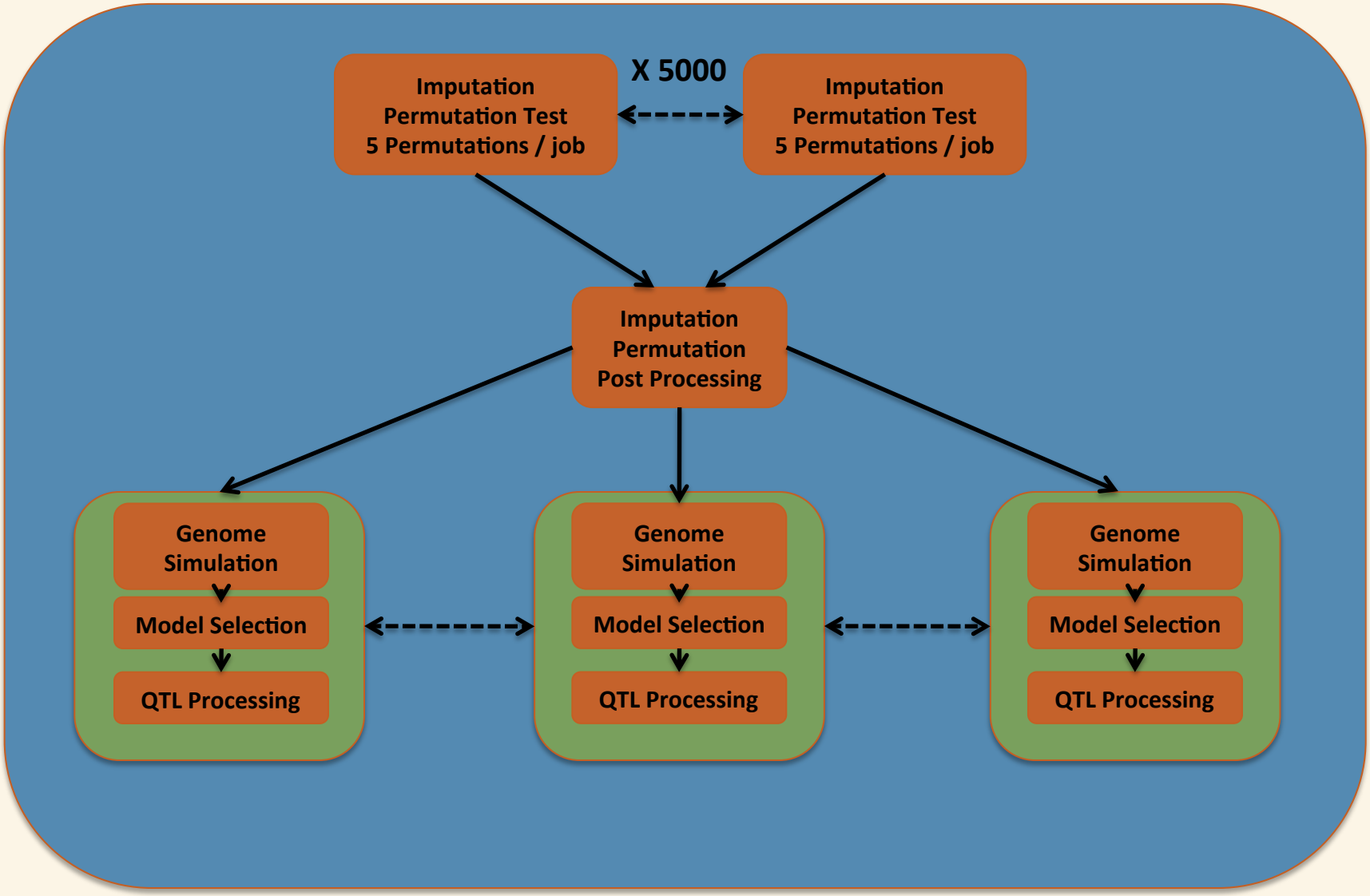
Even more useful than the instigating image-processing has been statistical genetic modeling of the results.

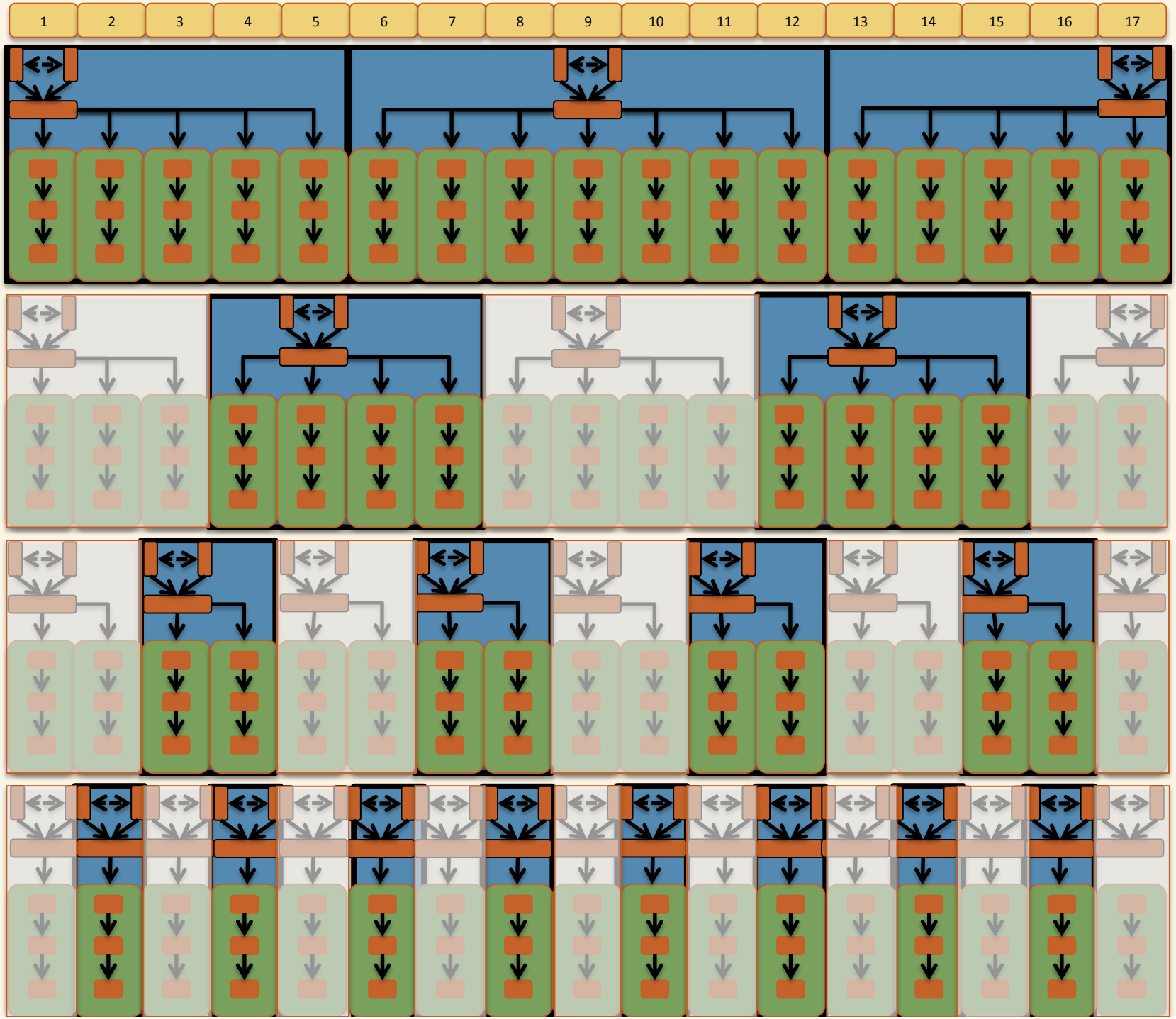
In the summer of 2010 we launched 1.1 million Condor jobs on CHTC as part of a genetic trait mapping experiment. This was far from trivial. My understanding is that this created learning moments for the Condor team. How do you get a condor to swallow a python?

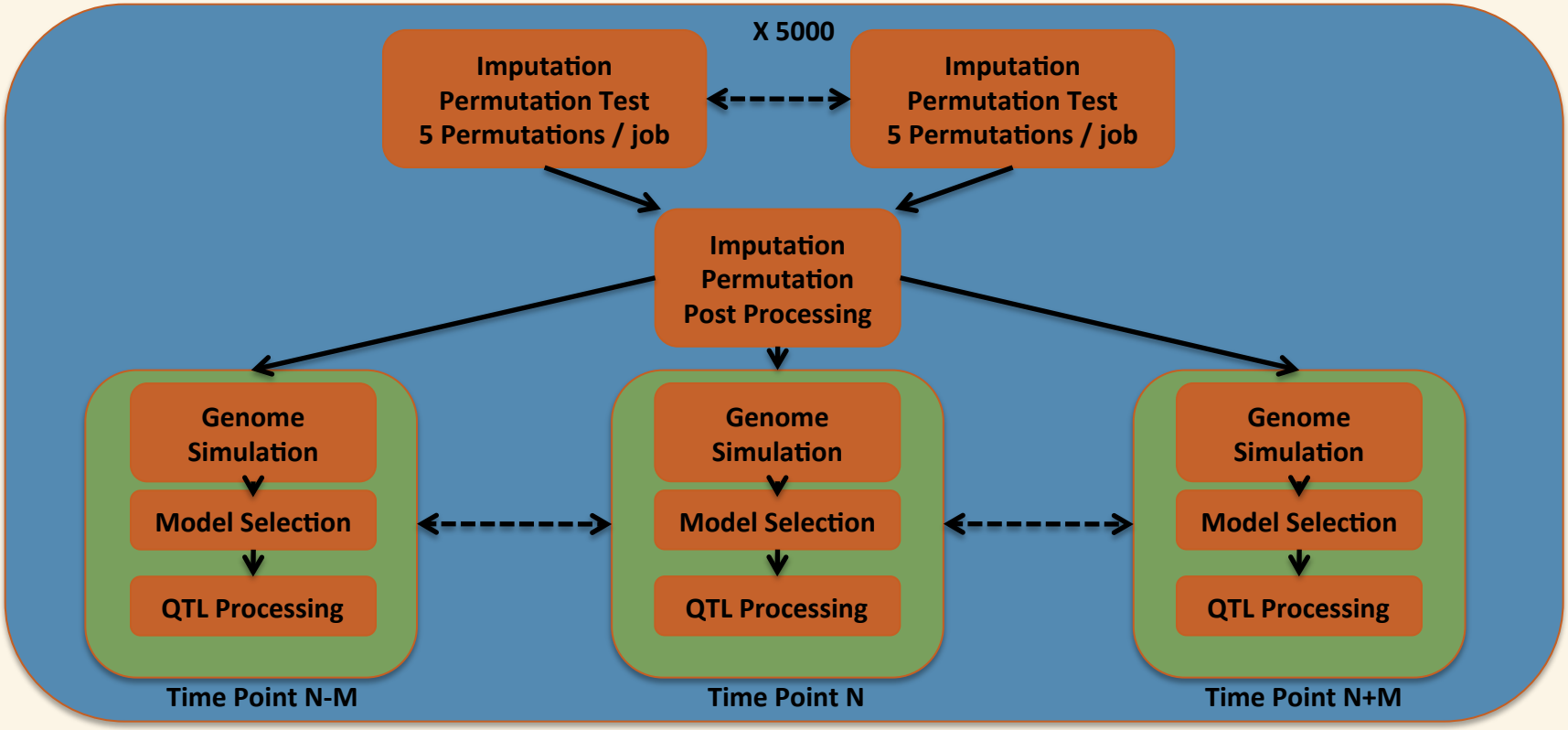


A new kind of Quantitative Trait Map was the Result



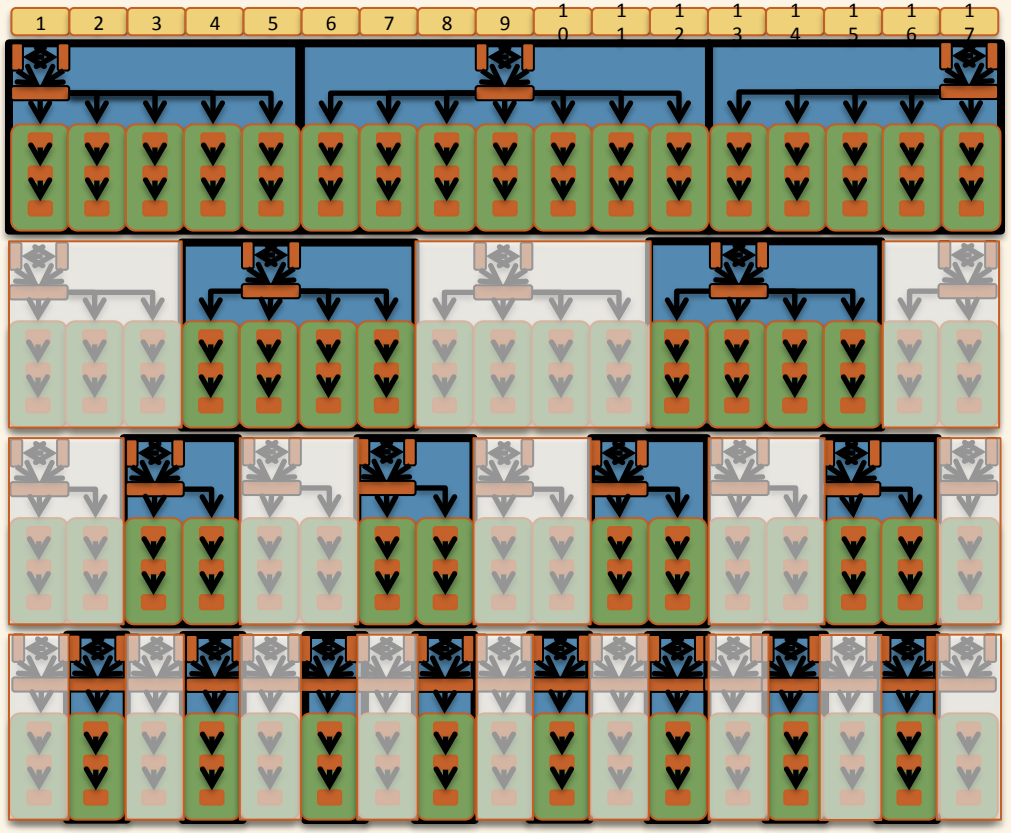

Imputation
Permutation Test
5 Permutations / job
X 5000
Imputation
Permutation Test
5 Permutations / job
Imputation
Permutation
Post Processing
Genome
Simulation
Model Selection
QTL Processing
Genome
Simulation
Model Selection
QTL Processing
Genome
Simulation
Model Selection
QTL Processing


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17



K = Number of Time Points
L = Iteration Level
M = Threshold Matching Distance
M = K / (2 ^ (L + 1))

Time points run at level L
1, K, round(1 + (M * O))
Where O is an integer such that
(1 + (M * O)) > 1 & (1 + (M * O)) < K



Q = Number of Iterations to compute exhaustively
Q = ceiling($\log_2(K-1)$)

P = Number of time points complete at level L
P = (2 ^ L) + 1

N(L) = Number of time points computed at level L
N(1) = 3
N(L > 1) = 2 ^ (L – 1)

Large DAG / Many Files

- Sub Dag External
- Super Dag - DAGMAN_MAX_JOBS_SUBMITTED
- Sub Dag - DAGMAN_MAX_JOBS_IDLE

Priorities

- Job – Control which jobs in queue get run first
- Node - Control submission of sub dags
- Lack – Priority knowledge between instances of dagman

Issues with Large DAG's

File Descriptors

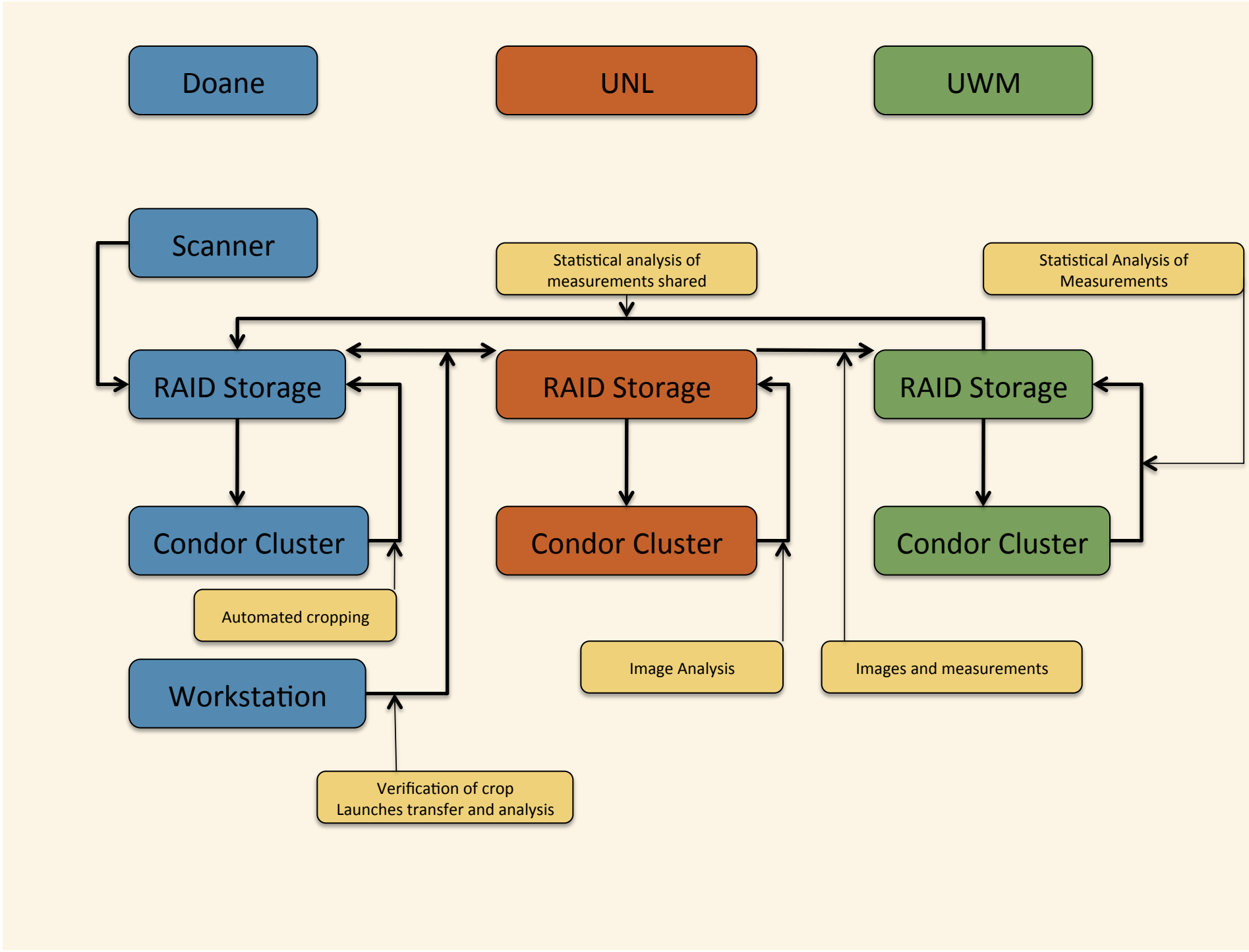
- **Fast file system**
- **Upgrade Condor – Lazy Logging**

Debugging

- **Prescript to determine whether a job is done based on what files exist,**
  - **edits the submit file and changes it to a noop job**
- **Exit Codes – Retry Subdag based on exit code of post script – AbortDagOn**

Retries

- **Want retries at top of**
  - **ready queue - DAGMAN_RETRY_NODE_FIRST = TRUE**
  - **condor queue – Prescript to inflate priorities based on retry number by editing submit file**


Doane
UNL
UWM
Scanner
Statistical analysis of measurements shared
Statistical Analysis of Measurements
RAID Storage
RAID Storage
RAID Storage
Condor Cluster
Condor Cluster
Condor Cluster
Automated cropping
Image Analysis
Images and measurements
Workstation
Verification of crop
Launches transfer and analysis