

The Bisque-Phytomorph Image Analysis Project

Edgar Spalding (spalding@wisc.edu) and B.S. Manjunath (manj@ece.ucsb.edu)

A major goal in plant biology is to learn how the genome produces and controls the functions of an organism. Phenotypes due to induced mutations or natural genetic variation are rich sources of information about the genomic basis of form and function. This decade and the one previous brought enormous, transformative progress in the way genomes and their proximate levels of expression are characterized and manipulated. The next decade will bring more of the same. On the other side of the equation, organism form and function, the methods of study have not undergone a similar transformation in precision or capacity. Thus, there exists today a large imbalance between the sophistication with which genotypes and phenotypes are quantified. Making progress on the phenotype side of the equation will result in better use of the genotype information, and most efficient progress toward the overall goal will be made when the two sides of the equation are in a technological balance.

Currently, two general classes of technical impediments limit progress in learning about genome functions through phenotypes. One is that measurement of the feature of interest is not sufficiently sensitive to detect the influence of a genetic effect. The other is that the method may not have sufficient throughput to make genome-level investigations feasible. Image analysis has the potential to raise the capabilities of plant phenotype studies by reducing both types of impediments. In a broad sense, image analysis is the science of extracting information embedded in digital images. The images could be of molecular distributions in a cell, or texture distributions in a crop field, or at some biological level in between. The images can be of one sample at one time or multiple samples in one frame, or frames at a series of time points. Image processing algorithms if appropriately sophisticated can extract and quantify information suspected of being important by the biologist or of unknown utility. In some instances such algorithms can be completely automated. Depending on the resolution of the acquired images, the level of extracted information can be much more detailed and objective than a manual measurement. Enough progress has been made recently on understanding how to perform image-based plant growth and development experiments, on feature-extraction algorithm design, and incorporation of the enabling computer sciences to propose an activity to iPlant that will have broad impact on how certain aspects of plant biology research are performed.

The Phytomorph project at the University of Wisconsin (<http://phytomorph.wisc.edu>) focuses on enabling plant research by creating and disseminating tools for quantifying plant phenotypes in popular experimental systems, mostly Arabidopsis and maize. The emphasis has been on the early stages of seedling development. The Bisque project at the University of California, Santa Barbara (<http://www.bioimage.ucsb.edu/bisque>) presents and advances a sophisticated database environment for storing and manipulating biological images. At its core is a flat-structured database designed for images and metadata. The key feature of Bisque is that it can incorporate analysis-specific algorithms to create a workflow that takes images and the metadata and returns processed results required by the biologist. The following proposal prepared by Edgar Spalding (PI of Phytomorph) and B.S. Manjunath (PI of Bisque) and their colleagues describes a year-long activity that would integrate aspects of their two projects into a cyberinfrastructure-based tool taking the shape of a portal that enables the plant biologist to perform quantitative, image-based phenotype studies. The longer-term vision is to integrate the output of a Bisque-based image analysis pipeline with the outputs of activities closer to the genotype side so that plant biology research begins to build true multi-dimensional quantitative representations of the genotype-to-phenotype relationship.

Starting Points

As starting materials, the Phytomorph project will provide image analysis algorithms capable of measuring the size and morphology of Arabidopsis seeds (the same for maize kernels are in progress), and root growth. These methods already exist. Some are fully automated (e.g. Arabidopsis seeds, some root growth and shape algorithms). Some require bare minimum input (e.g. maize and rice coleoptile growth and shape, Arabidopsis root branching) and some require user supervision (e.g. Arabidopsis hypocotyl tracing). Thousands of experiments, each typically consisting of hundreds of images already exist in a database. Some of these experiments have already been processed from image acquisition all the way through the quantitative trait locus-mapping stage. The Bisque project has immediately available a user environment in which the existing algorithms can be exposed to the user and through which the results will be presented.

Next Steps

The next steps will be to incorporate 2 or 3 well-tested algorithms into Bisque, adapting both until a user-friendly tool with optimal performance is achieved. The first algorithms selected will be those for which a lot of raw images and a set of processed and well-mined data exist.

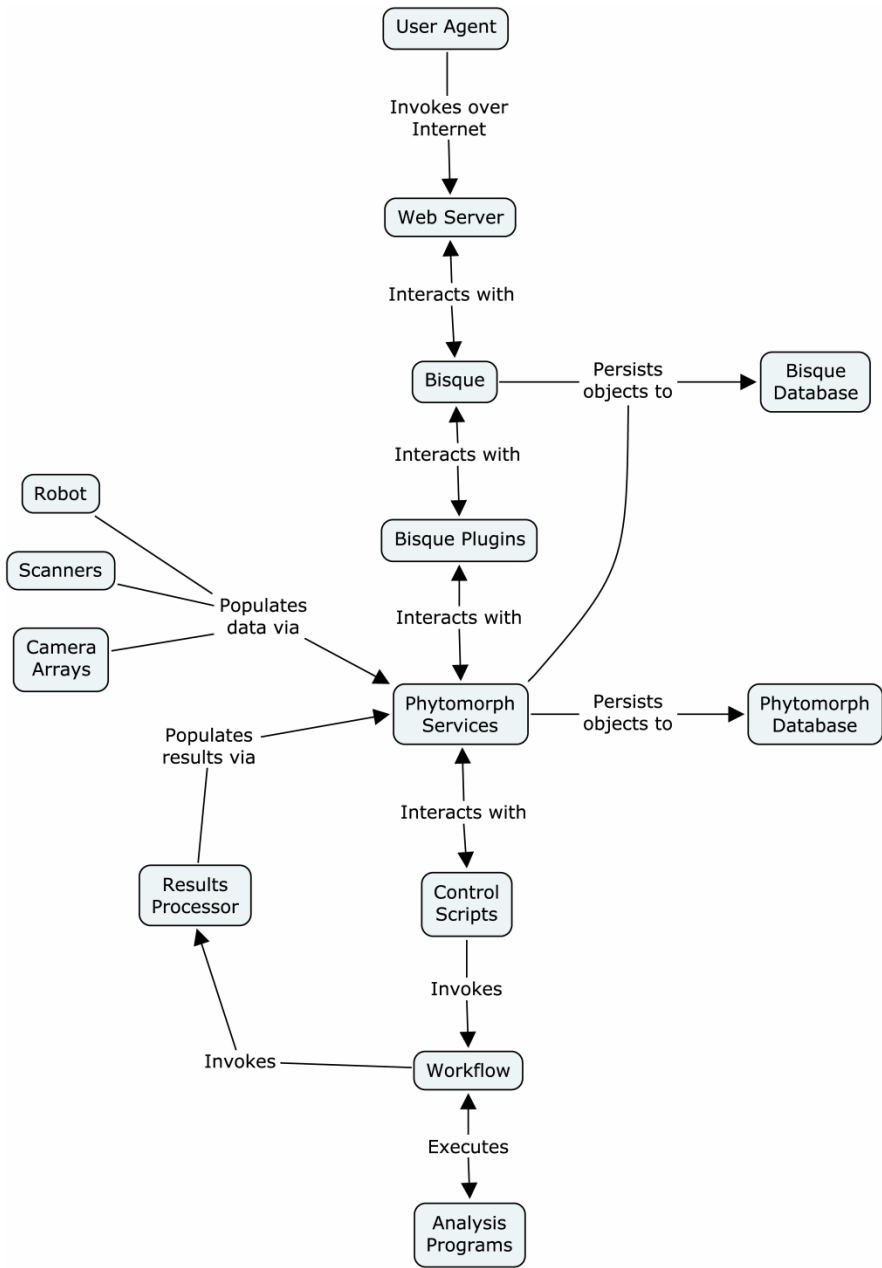
Creation of a relational database and a workflow that populates it as well as the xml-tagged Bisque database to provide users alternative methods of interacting with the data. In some cases, the 'flat' structure of the Bisque database will be ideal and for other purposes a more rigidly defined relational database may be advantageous, especially when integrating with other types of 'omic' data.

Additional research and development directions include incorporating the Condor distributed computing tool at the back end of Bisque. Essentially, the goal will be to make a biologist-friendly tool capable of converting a user request and image set into a Condor job. We will also be looking at the cloud computing paradigm that is becoming increasingly popular for large scale computing.

The Year Goal

At the end of a year we are confident that a robust tool for efficient analysis of images of plant structures relevant for phenotype studies will be in operation. It will feature algorithms that we know from experience are helpful to multiple labs that acquire compatible image data. Roots and shoots of monocot and dicot seedlings will be represented. As the community sees what can be accomplished by this way of quantifying phenotypes, we expect interest in algorithms capable of analyzing specific phenotypes at scales ranging from subcellular to field plots will drive path-breaking research. Any such algorithms could be incorporated into the tool.

The Figure below shows a preliminary flow chart for the proposed tool. In general, the Bisque team will focus on the databasing and workflow aspects and Phytomorph team will focus on algorithms, their integration, and connection to a Condor grid. There will be substantial areas of overlap between the teams. For example, the Bisque team is also expert at image processing algorithms and the Phytomorph team will need to work at some times deeply with the Bisque source code and databases. A Powerpoint presentation outlining some ideas of how the project could be structured and the work distributed (appended) served as the basis of a recent 2 h meeting between the Condor and Phytomorph teams. Miron Livny, PI of the Condor project is committed to the project.



The Necessary Support

To achieve this goal in a year, a person on the Phytomorph team and 2 on the Bisque team should focus full time on it (the project manager of Bisque, Kris Kvilekval, will spend 75% time and another full time project scientist will be helping Kris). The below budget reflects their salary and fringe benefits, some travel funds for personnel exchange, and institutional overhead (at its full rate which may not apply in this case). We are not currently expecting the Condor grid component to require resources but that could conceivably change if the project grows quickly.

Phytomorph side

Salary for scientific programmer	60,000
Fringe	23,700
Travel	5,000
Indirect costs (overhead)	42,500
Total	\$131,200

Bisque side

Salary for Kris Kvilekval (75%)	90,000
Salary for a Project Scientist (100%)	60,000
Academic Time (faculty)	10,000
Fringe (estimate)	45,000
Travel	05,000
Total Direct Costs	210,000
Workstation	05,000
Indirect costs (UCSB overhead, 51.5%)	108,150
Total:	323,150

Condor side

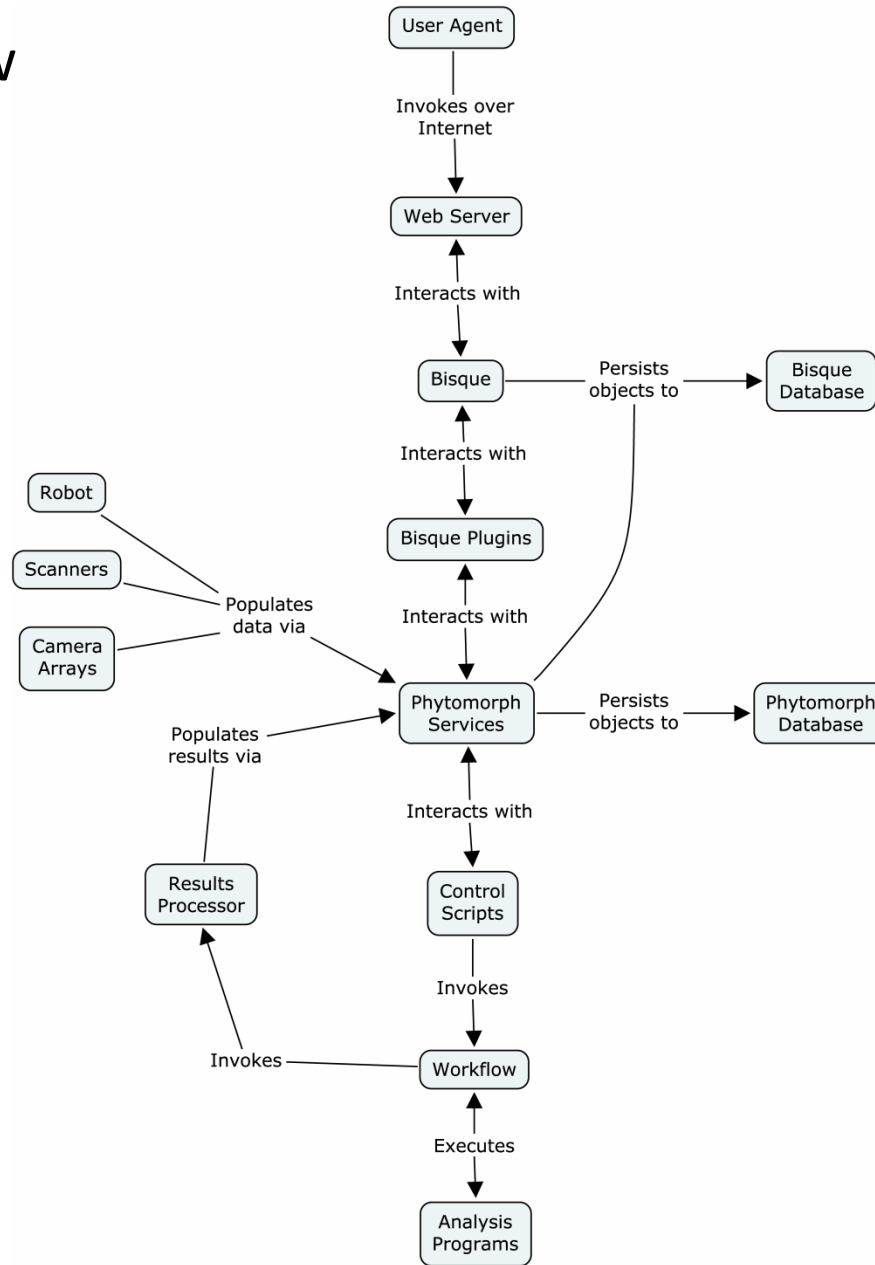
The Condor team is not requesting support.

Bisque, Phytomorph, Condor Collaboration:

Project Description

This presentation was used in a meeting between members of the Phytomorph and Condor groups on August 20.

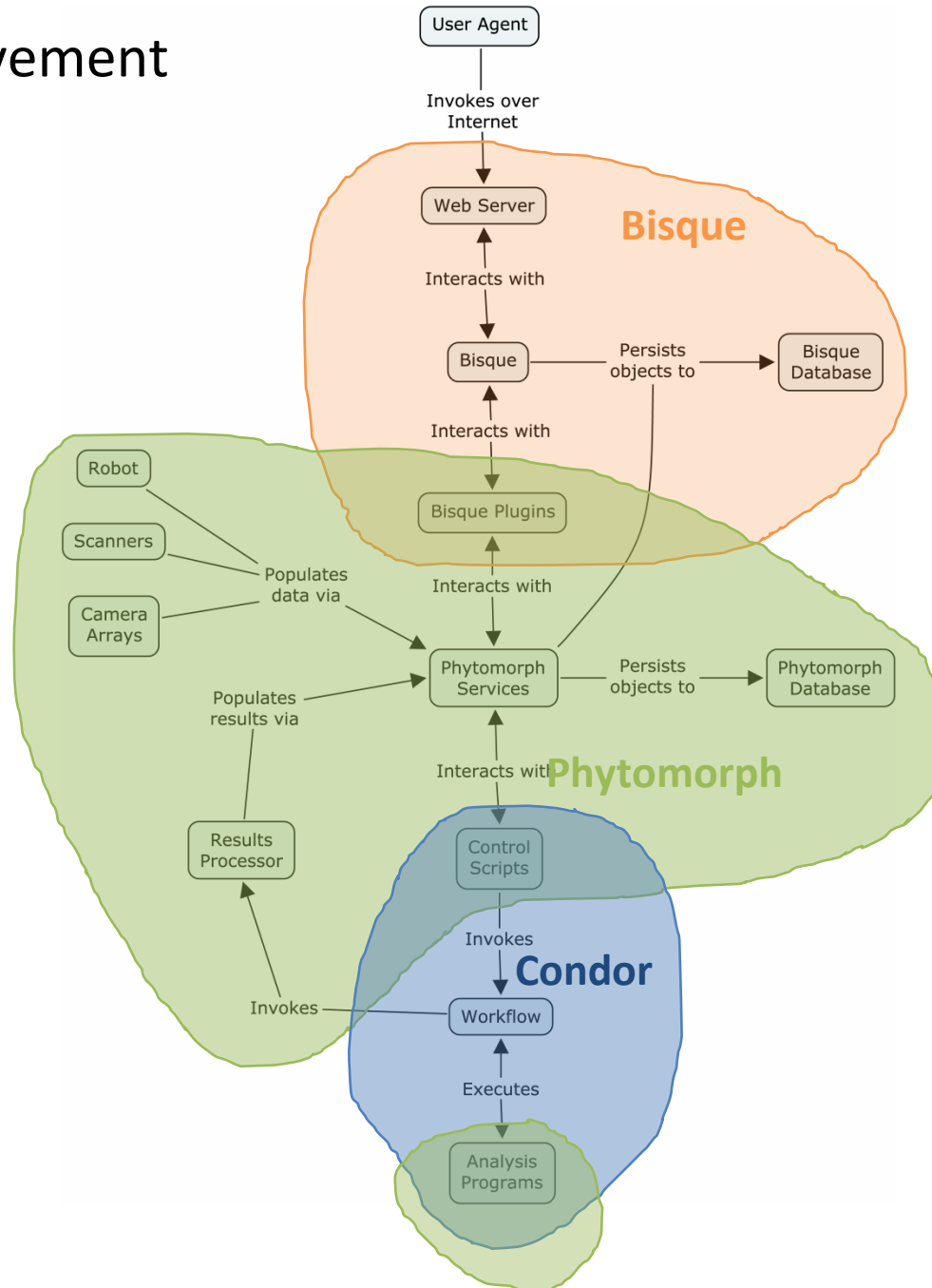
Project Overview



Project Overview

- The project follows a very standard multi-tiered architectural pattern
- Primary project concerns:
 - User interface (Web tier)
 - Domain logic (middle tier)
 - Persistence logic (middle tier)
 - High-throughput computing (computational tier)

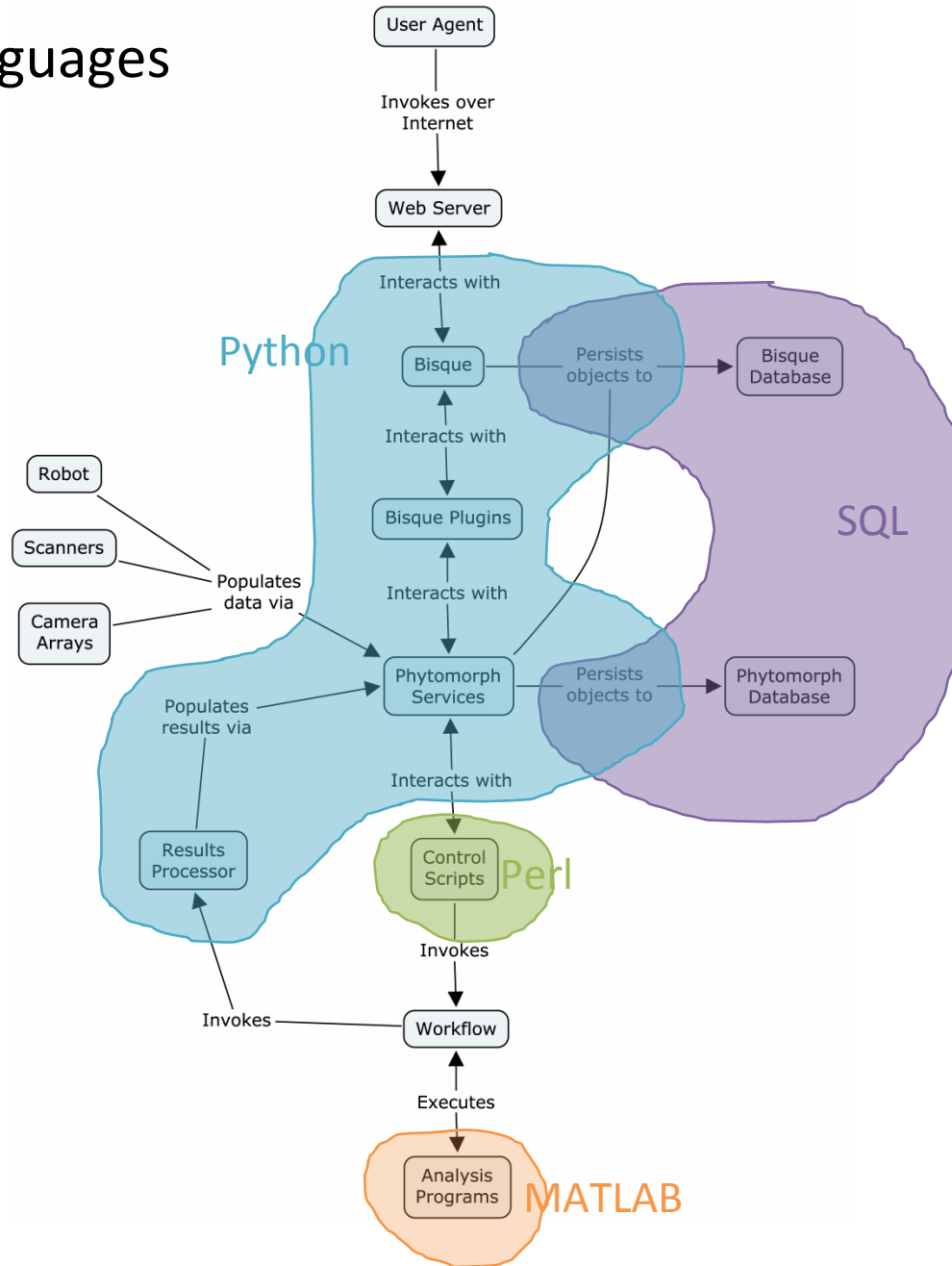
Team Involvement



Team Involvement

- Primary duties of each team are partitioned approximately across functional lines, with overlap:
 - Bisque is used to implement the user-facing Web tier
 - Condor enables the high-throughput computing functionality within the application
 - Phytomorph will implement the middle tier, integrating the stack of components. Phytomorph also provides plant sciences expertise and the fundamental analysis routines.

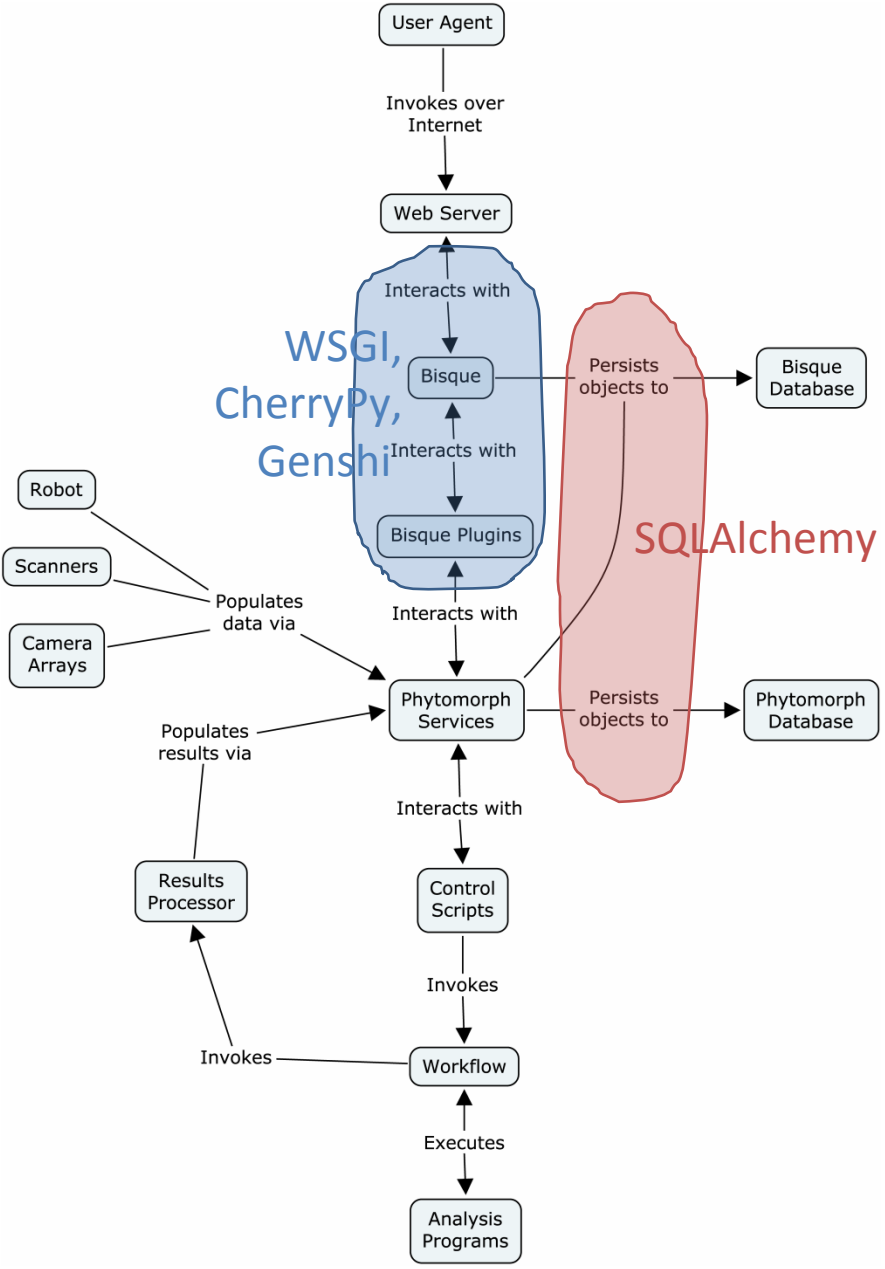
Programming Languages



Programming Languages

- This diagram shows the primary languages from the perspective of the Phytomorph project/middle tier
- Python is a solid general-purpose language, and is well suited for the middle tier's needs
 - Bisque is implemented in Python
- MATLAB is used by the Phytomorph group to develop image processing algorithms but the tool should accommodate compiled algorithms written by Bisque or others in any language.

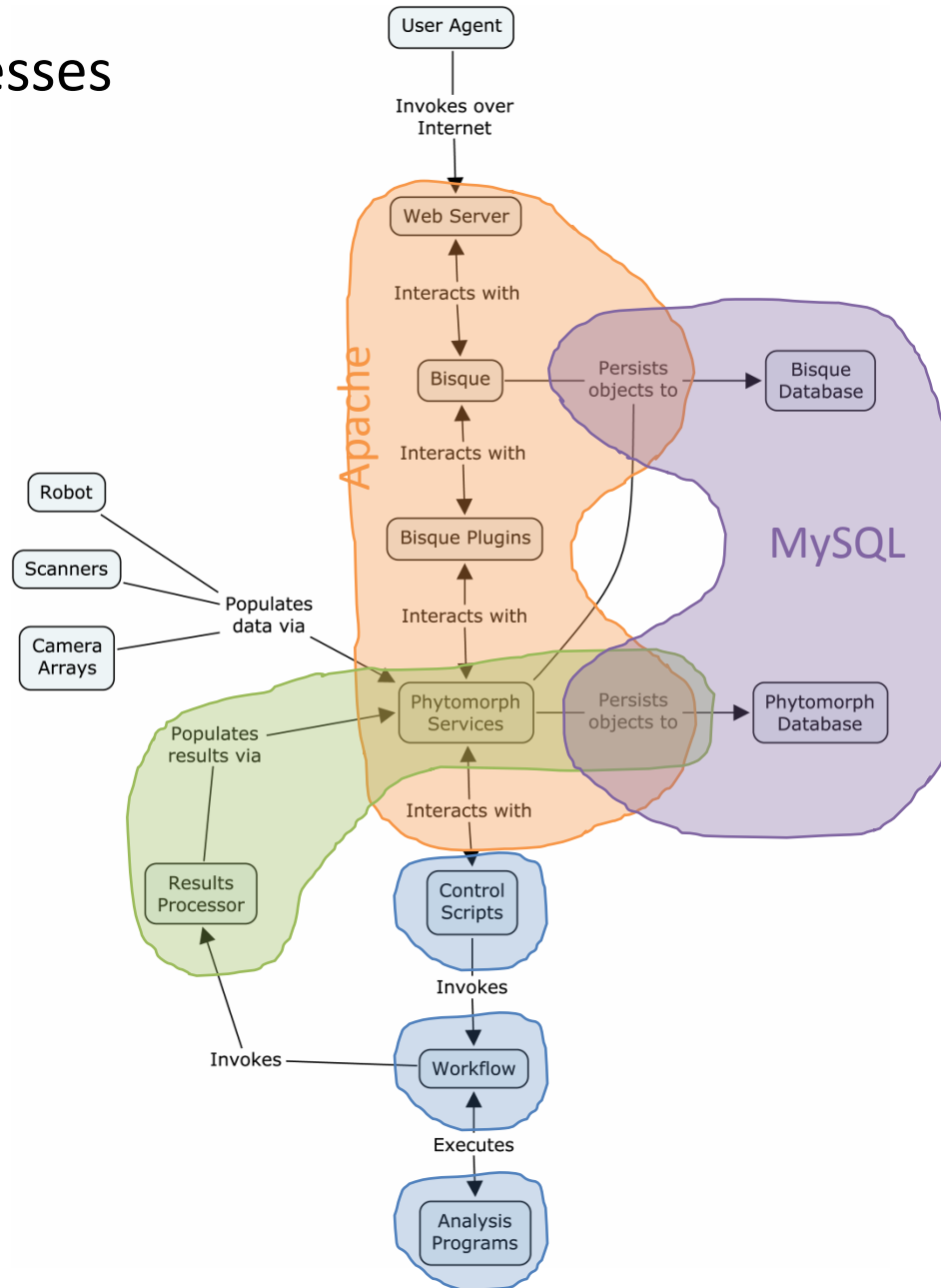
Major Libraries



Major Libraries

- Bisque is implemented using TurboGears, which is a stack of smaller Python libraries; we will inherit these wherever possible
- The most important library is SQLAlchemy, which maps Python objects to relational database tables)

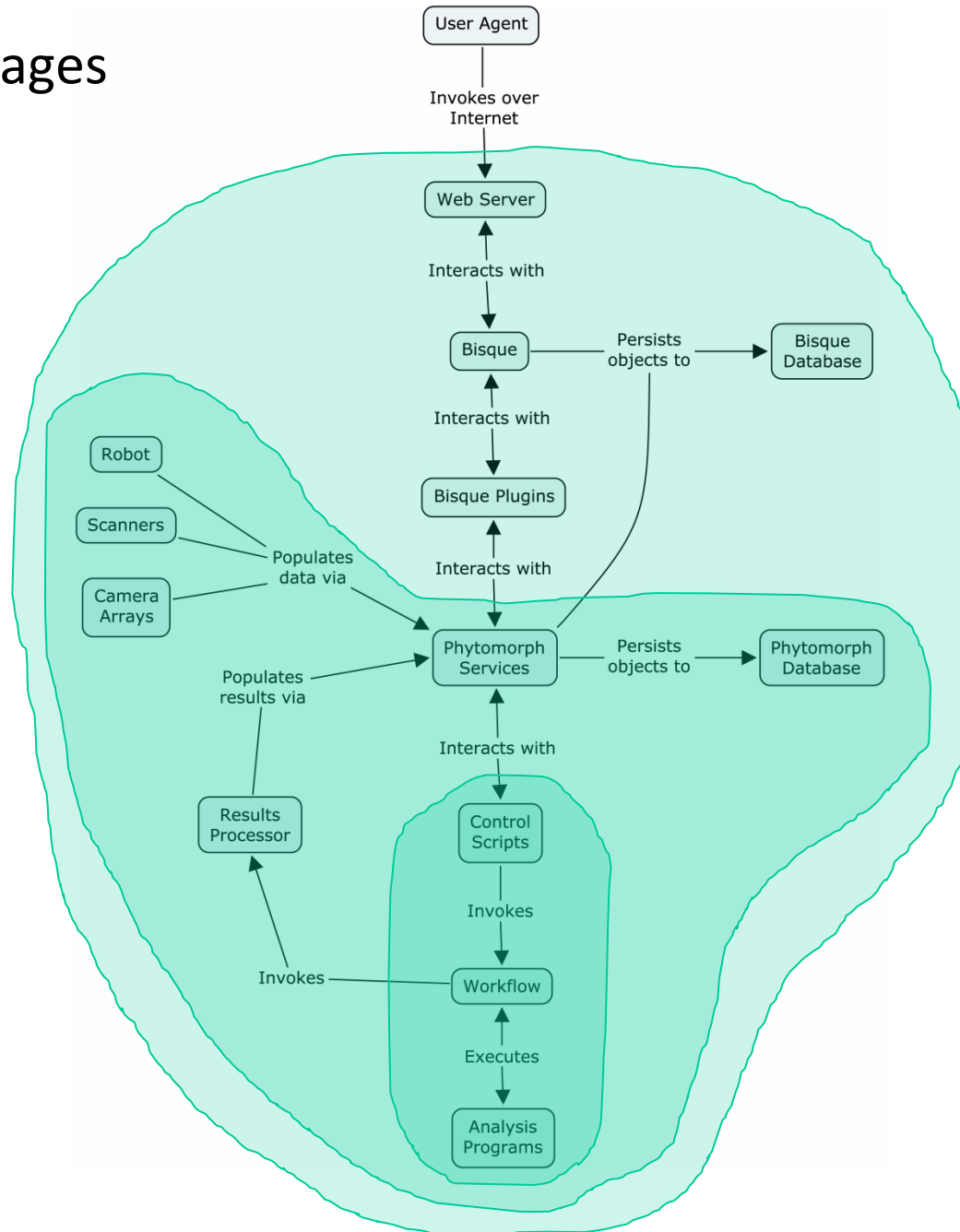
Significant Processes



Significant Processes

- A LAMP-based implementation stack is used
- MySQL and Apache are already in production within the Spalding lab
- Request handlers within Apache delegate control to Condor via SOAR as is appropriate

Project Stages



Project Stages

- Development will proceed in a bottom-up fashion, successively building atop more fundamental components
- While conceptually distinct, the stages are likely to be developed largely simultaneously
 - For example before it is possible to integrate Bisque, the system itself must be deployed
- Much of the first stage (indicated by the most heavily shaded area) has already been completed