ALGAEORITHM

APRIL 6, 2022

Algaeorithm

Bioenergy for the Next Generation

WRITTEN BY

Ashwin Mukherjee

Rohan Chanani

PROJECT ADVISOR

Jake Valenzuela, Ph.D.

Table of Contents

List of Figures

Introduction

The grand challenge of the AlgaePrize competition is to develop novel solutions to algae production and processing that ultimately lower the cost of bioenergy products. Our team has taken an unconventional yet innovative approach to address this challenge. We believe that significant changes to the algal bioenergy and bioproduct industry require a generational shift in student accessibility to bioenergy information and a deeper understanding of the potential of machine learning to unlock the biological complexities from the boom of “big data”. Therefore, we have focused our efforts on developing and deploying a tool, Algaeorithm, that is a lens for high school students to explore the possibilities of the bioenergy industry and see the vision of BETO's education mission. Specifically, our project directly aligns with OPERATION BioenergizeME’s mission to support formal and informal education by engaging future scientists, engineers, and entrepreneurs about the bioenergy industry, while introducing how a powerful tool like machine learning can help solve the world’s energy problems. We believe our outside-the-box approach can have a major impact on the human capital that will drive the bioenergy industry of the future.

The Algaeorithm project originated out of the curiosity of two high school students, Ashwin Mukherjee and Rohan Chanani, who took the initiative to seek information about the algal biofuel industry and how they could incorporate machine learning into a useful tool for researchers. They sought out the expertise of systems biologist, Dr. Jacob Valenzuela, who has extensive experience in characterizing the physiological state shifts between growth dynamics and lipid accumulation (Valenzuela et al. 2012, 2013; López García de Lomana et al. 2015; Imam et al. 2015). Dr. Valenzuela has also developed an algal-based curriculum, *Bioengineering a Sustainable World* (BSW), as part of the Institute for Systems Biology (ISB) Systems Education Experiences (SEE) program. The curriculum teaches students how bioengineering can address the world's bioenergy and bioproduct challenges. The team, led by Ashwin and Rohan, has constructed a prototype application that uses machine learning to quantify cell concentrations from microscope images. In this synopsis, we provide a summary of how we aim to use Algaeorithm to engage students, educate them on the algal bioindustry, and introduce them to machine learning concepts. The novelty of this project is in its vision to invest in the human capital at the student level to achieve the generational shift required to disrupt the current petroleum-based economy. Our team has already constructed a working prototype of the Algaeorithm application proving its technical feasibility and how we can expand its versatility and refine its accuracy. We believe this project will make a significant impact in the classroom by providing substantial time savings, allowing students to access Algaeorithm’s statistics, and creating multiple instructional and explanatory videos and resources.

Novelty

The challenge of shifting the US economy to a bioeconomy is a generational problem. Large shifts toward sustainable technologies take time and face resistance from the status quo. For example, electric vehicles and solar-powered rooftops are now commonplace but were inaccessible to the general public 20 years ago. The catalyst for this movement will be the next generation of students, who can turn promises of a green future full of sustainable biopower, biofuels, and other bioproducts into their reality. Using a machine-learning-based web application and a diverse offering of educational content, our proposed project aims to support this movement by engaging current high school students in the world of bioenergy. We predict two major requirements to make the shift to a bioeconomy, both of which will be addressed by our project:

1. Increased education and awareness around bioenergy processes
2. The generation and analysis of “big data” using machine learning techniques.

The immense necessity for algae technology has only recently become mainstream, and as such there are not currently tools that adequately address these challenges. Furthermore, machine learning’s popularity and utility have both soared in several fields, including bioenergy, but much of the current educational resources for machine learning are directed at professionals, which leaves most students with little or no exposure to machine learning in high school. The broad accessibility of resources in established industries such as finance and graphic design has proven to be extremely effective at increasing student interest and eventual matriculation to related careers. An algae-focused application designed specifically with bioenergy and machine learning education in mind would yield the same positive effects: providing resources for the next generation of professionals to develop essential skills for a career in algae technology while serving as a platform for wide-reaching bioengineering curriculums.

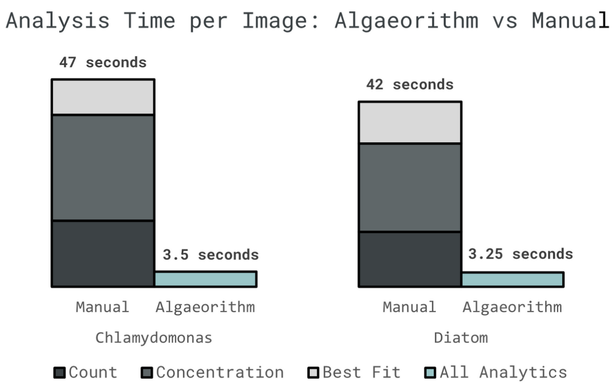


Figure 1. Analysis Time per Image: Algaeorithm calculates the cell count, concentration, and growth curve around 13.4 times faster than manual counting.

In classrooms and academic labs across the country, the primary means of obtaining culture dynamics from a time-course experiment is manually counting cells one by one in a microscope. A significant amount of time is spent both by the teacher explaining how to use a hemocytometer and by the untrained students trying to comprehend a difficult technique.

Based on our experience both as students and educators, we believe that learning about bioenergy practices and discussing challenges would be a more productive use of this time. For this reason, we constructed Algaeorithm, an application that uses machine learning to consolidate these processes into an easy-to-use application and capitalizes on the ability of computational tools to perform repetitive, predictable tasks. Algaeorithm already calculates relevant data much faster than manual analysis for *Chlamydomonas* and *Phaeodactylum* algae strains (Figure 1), which are both at the forefront of biofuel development. This improvement both allows students to process more images within a specific time and frees up class time that would have been spent manually counting cells (Figure 2). We plan to make significant improvements to the application by further reducing computation time, increasing the algorithm's accuracy, and introducing additional species, all of which are both fundamentally different from existing approaches and well within what we can accomplish with machine learning.

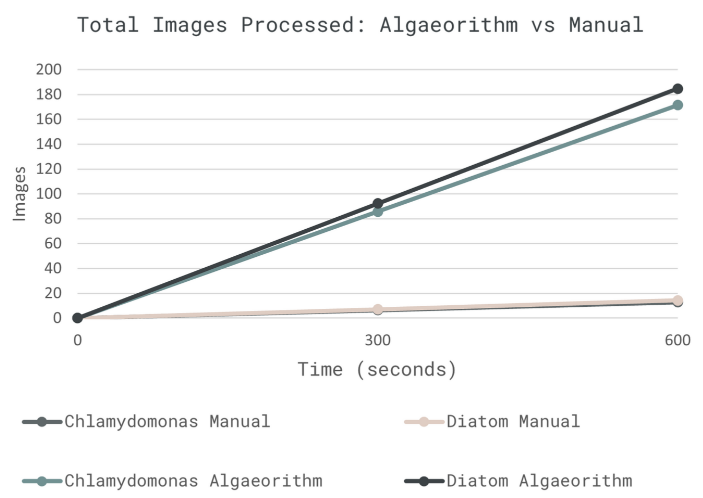


Figure 2. Total Images Processed: Within a given time period, the Algaeorithm prototype is able to process far more images than manual counting.

To improve the accuracy of Algaeorithm, we will construct mini photobioreactors, which can be made from laboratory supplies available in most science classrooms (Figure 3). We will cultivate algae ourselves to both optimize the application and demonstrate the viability of extracting high-quality information from algae samples using accessible tools. We also plan to provide the plans and instructional videos for setting up mini photobioreactors in the classroom. Algaeorithm is accessible from any web browser, so any smartphone with internet access, which is available to over 90% of the U.S. population (Pew Research, 2021), can upload images if the user can grow algae and has access to a microscope, which is available in most high schools. Thus, our project would improve both the quality and accessibility of algae research for students who are the next generation of scientists and engineers - the foundation of the algae value chain.

Currently, the fields of algal bioenergy or bioproducts and machine learning face a significant lack of exposure among high school students. In comparison with nuclear, solar, and wind energy, very few students are familiar with bioenergy, and most students know very little about machine learning beyond digital assistants and pop culture depictions of artificial intelligence. Our project will help remedy these issues with the following initiatives along with the actual Algaeorithm tool:

- Video tutorials and demonstrations that allow middle and high school students of all experience levels to use our application and gain valuable bioenergy experience working with algae
- Educational content about the basics of machine learning and how we use it in our application to expose students to this field
- A live, interactive workshop with students and teachers to teach them how to use the application in the classroom
- If applicable, collaboration with the BioenergizeME team to integrate some of Algaeorithm's proposed content into BETO's Bioenergy Education Resources

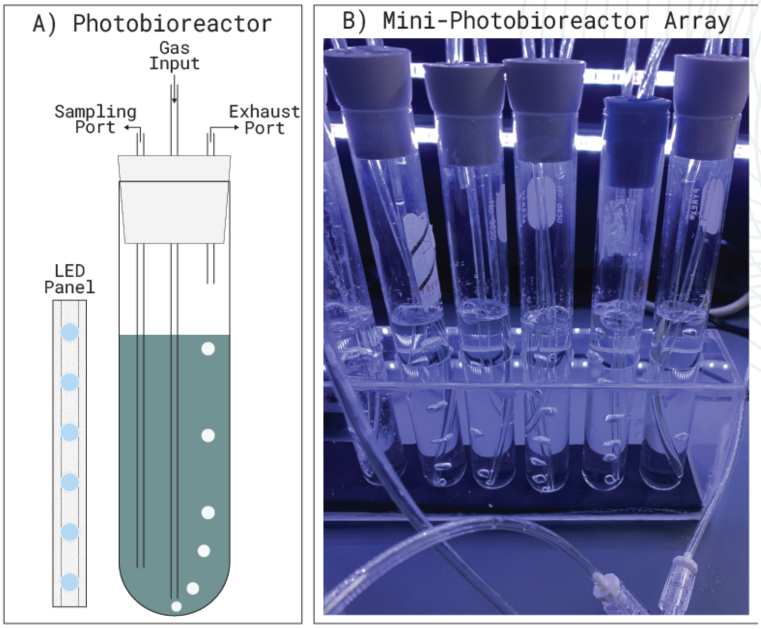


Figure 3. Mini Photobioreactors made by ISB: Small scale versions of the equipment used by scientists in research and commercial settings will allow students to replicate a professional environment when growing algae.

Machine learning has already become and will continue to be a major factor in the bioenergy industry. As such, ensuring that the next generation of bioenergy scientists and engineers are well versed in the field of machine learning will be critical to enhancing the commercialization potential of algae. From systems biology for strain improvement to understanding abiotic control of large outdoor ponds, the bioenergy applications of machine learning and big data are endless. Our application's use of machine learning shares the same fundamental concepts as the machine learning tools that are necessary to address issues like these. Thus, exposing students to these concepts with our application and educational content would better prepare them for a career in algae biotechnology, improving and increasing the human capital in the algae value chain.

Increasing education about algae's capabilities and the R&D process would spark interest in bioenergy among students across the country. It would also lead to dynamic participants in the bioeconomy who understand the importance of sustainable production and leverage tools like machine learning to directly enhance all domains of the algae value chain. Altogether, our application would contribute to a robust talent pool, educate participants in the bioeconomy, and ultimately streamline development, moving algae closer to becoming cost-competitive in global markets.

Technical Feasibility

The project consists of three major areas of technical focus: development of Algaeorithm, cultivation of algae for testing, and broader impact through educational tools.

Utilizing Neural Networks

To computationally extract useful data from uploaded images, the algorithm will include multiple convolutional neural networks. The neural networks will improve upon those already included in the application prototype, which is currently being beta tested by students and educators. A convolutional neural network sequentially applies linear transformations to inputs, in this case, microscope images of algae cells, and minimizes the difference between the predicted output and the actual output, in this case, the locations and classifications of the algae cells. For the cell counting algorithm, the neural networks will serve the following purposes:

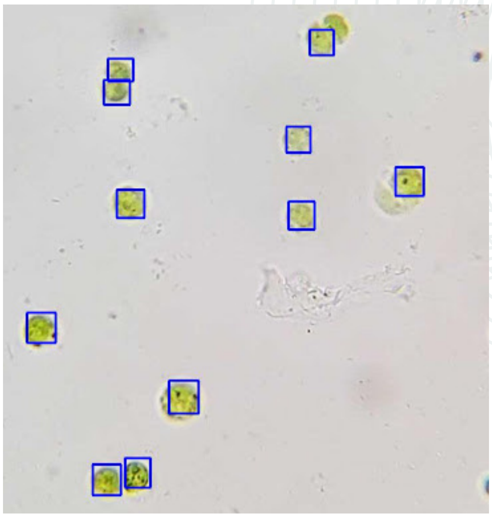


Figure 4. Neural Network Output: Image of Chlamydomonas cells annotated by the Algaeorithm neural network. The dark blue boxes represent which cells the algorithm identified.

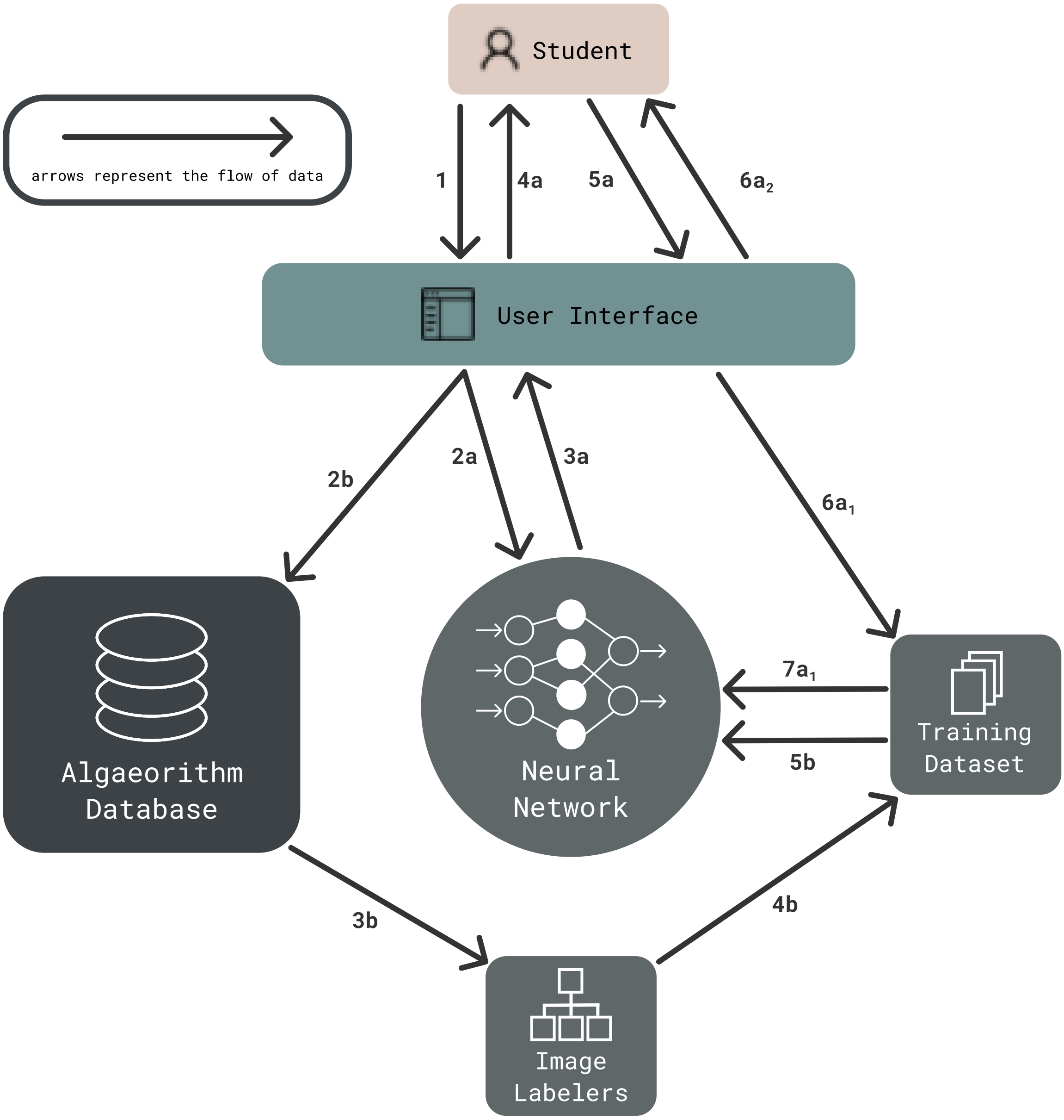
- Locating bounding boxes for all of the cells in each image, with the coordinates of each box for visualization purposes as well as the number of boxes and areas of each box for calculation purposes (Figure 4)
- Classifying the species of each located cell to provide information about the algal population and indicate possible invasions from external species
- Quantifying visual metrics from individual cells and the collective population, including chlorophyll level, lipid count, and overall algae health

The algorithm will also use the number of pixels in the image, known measurements from the species classified in the image, and the depth provided by the user or inferred from the image to calculate the cell concentration in each image, and it will ultimately return this concentration along with the corresponding cell count.

Analytical Processing

The algorithm will compute statistical metrics from the returned cell count and concentration to benchmark against existing data, which will come from both the Algaeorithm database of user-uploaded photos and from external databases. If users provide time-sensitive data, the application will use predictive models to forecast growth curves and output metrics about the macrostate of the population, such as carrying capacity, growth rate, and total biomass accumulation. These overall metrics will be combined with the information from the neural networks regarding chlorophyll, lipids, and cell health to provide a clear picture of the algae population state.

Technical Overview



Process Flow Diagram 1. Data Pipeline: After the user uploads an image (1), two pathways are initiated (a & b). First, the image is sent to the neural network (2a), which returns the corresponding output through the interface (3a) to the user (4a). The user can then modify the results and resubmit the edited output to the application (5a) to be used to both provide the user with more accurate information about their samples (6a2) and increase the training dataset (6a1), which will be used to retrain and improve the neural network (7a1). Second, the image is uploaded to the database (2b), where it will be made accessible to image labelers (3b). Each labeled image will then be added to the training dataset (4b), which will be used to once again retrain the neural network (5b).

User Interface

The application will concisely display the information computed by the algorithm through the following aspects of the user interface (Figure 5):

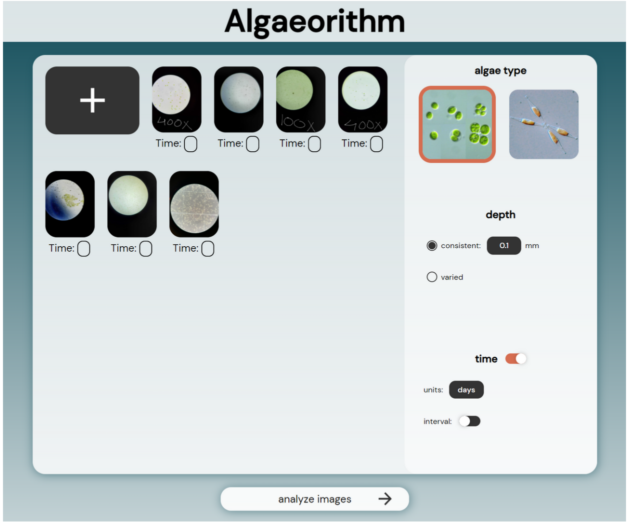


Figure 5. User Interface: The Algaeorithm prototype's user interface makes it easy for students to input images and associated information.

- Distribution graphs, such as box plots and histograms, to visualize frequency data
- Growth curves to visualize time-sensitive data in multiple forms, including linear, exponential, and logistic growth (Figure 6)
- A data table, which can be downloaded as a CSV file or exported to applications such as Microsoft Excel for further analysis
- Specific algorithmic output for each image, with bounding boxes around the identified cells and labels for the corresponding species and visual metrics

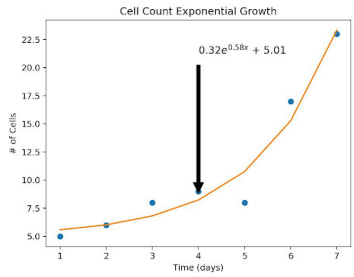


Figure 6. Exponential Growth Curve: The graph and equation were generated by the Algaeorithm prototype from a time series of 7 Chlamydomonas images.

The bounding boxes and labels on each image will be adjustable to allow users to manually improve the accuracy of their output data and to create additional training data for improving the core algorithm and neural networks. Users will also be able to compare their images and statistics with the application's broad database of results.

To test and improve Algaeorithm and its core algorithms, algae will be cultivated and compared with the output of the application. *Chlamydomonas reinhardtii* and *Phaeodactylum tricornutum* will be used as sample species to test the performance of the application with various types of microalgae. Algae cultivation will require controlling the environment in which the algae grow and collecting data from the algal samples, which can be used to retrain the neural networks.

The two main goals supporting this project are creating a tool that consistently outperforms traditional methods of manually analyzing algae samples and introducing the subjects of bioenergy and machine learning to a broader audience. The associated metrics for the application are accuracy, computation time, and information quality. For student exposure, the associated metrics include images uploaded, video views, and student survey data. The project will also be aided by numerous databases and resources with information about the species the application will analyze, such as the Chlamydomonas Resource Center, DiatomBase, and existing research from Google Scholar. Manual testing of the application and sources like these will provide a baseline to compare against the application results and assess the aforementioned goals and metrics.

The risks associated with the application are primarily due to the automation of cell analysis, which can cause unexpected behavior and analytical inaccuracies when faced with edge cases. However, these risks can be mitigated with the following strategies:

- Increasing the size of the training set with user uploaded and manually collected images and modifying the model architecture to improve the neural network
- Allowing users to manually correct any inaccuracies in the image labeling and growth projections, thus improving each user's data as well as providing additional labeled images to train the algorithm itself.
- Continuously testing and updating the algorithm with known accurate samples

The main risk associated with student exposure is the possible unpopularity of the application or additional content due to difficulty of use, obscurity, or inapplicability. These risk factors can also be addressed in several ways, including the following:

- Hosting workshops and Q&A sessions to explain the benefits and capabilities of the application along with how to use it
- Leveraging ISB's existing network of students and teachers to maximize outreach
- Including application support for multiple species and use cases

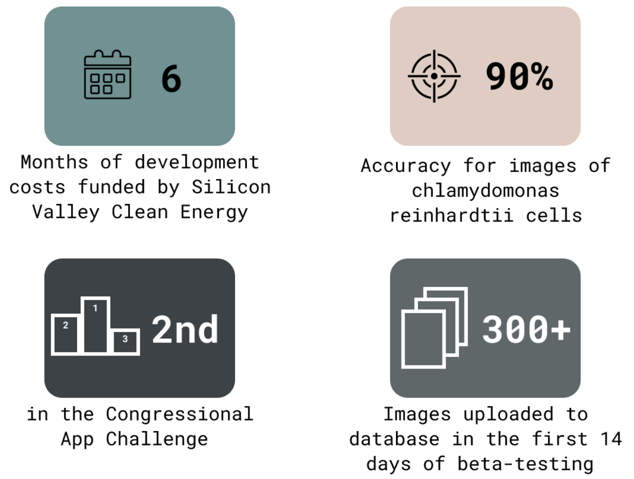


Figure 7. Accomplishments: Algaeorithm has already achieved success in several relevant metrics and accolades.

The project has already reached several significant milestones (Figure 7), and both students involved with the proposal have significant experience which will be relevant to the project, primarily the development of the Algaeorithm prototype. Rohan Chanani has worked on multiple web applications and implemented several machine learning models, including neural network architectures. Ashwin Mukherjee has experience creating explanatory content and intuitive user interfaces, and he has also manually worked with algae samples. The project's advisor, Dr. Valenzuela, has extensive experience in the field of biofuel research and bioenergy education.

The completion of this project will require cloud computing resources for application development, web hosting, and image storage, as well as laboratory resources for algae cultivation and data collection. The cloud computing resources are available through the Amazon Web Services (AWS) platform. The laboratory resources available include mini photobioreactors recently developed by ISB specifically for education, microscopes to collect data, and samples of various algae species, and the project will also have access to the ISB Baliga Lab, which hosts one of the top biofuel research facilities in the country.

Impact

Our team anticipates that Algaeorithm would expand access to and awareness of the bioeconomy, enable more collaborative academic endeavors, and drastically cut R&D resource usage. We anticipate the largest impacts in education by potentially allowing millions of students to participate in real-world algae research. Additionally, students would be part of a rich network that could use our application as an organizational framework to collaborate on large-scale projects. In ISB's words, these projects help "bridge the distance between the professional lab and the classroom." On the technical end of our project, one of the most crucial improvements is the ability to accurately model algae growth, which offers long-term insights using short-term data. This broadly applicable feature would also include image-based cell health diagnostics. Our team estimates that this could cut the cost of experimentation by 15% and reduce water, growth media, and nutrient usage by 65% leading to an overall increase in sustainability.

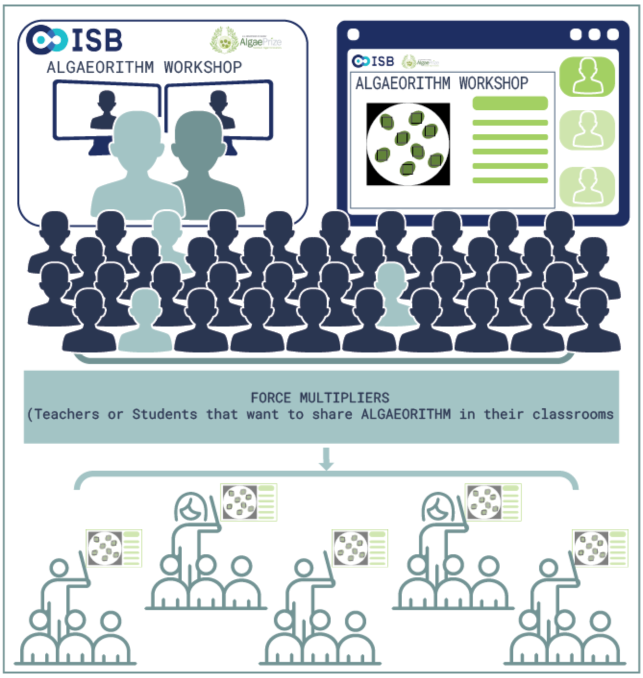


Figure 8. Algaeorithm Student Impact: By hosting live workshops for students, Algaeorithm will force multipliers who can share this information with their communitites.

In aiming to achieve cost-competitiveness with fossil fuel markets, far more students will need to become engaged with bioenergy and develop career-relevant experience. Our project would contribute to this labor pool by introducing students to the field of bioenergy and providing them with the knowledge, ability, and interest necessary to pursue bioenergy in a career (Figure 8). From industries ranging from agriculture to financial services, machine learning models have helped create more lean processes that increase profitability and cost competitiveness. Our team believes that a developmental approach that incorporates all steps of the production cycle would be able to realize these gains in the algae industry as well - a crucial step towards broad global bioproduct adoption. Our project would also generate both student interest and experience in machine learning, which are crucial for the development of these models.

Today, a barrel of petroleum typically costs around $80-$120 a barrel, while the same barrel of biofuels can cost anywhere between $300-$2600. However, as countries raise carbon taxes, oil reserves diminish, and algae development gets cheaper with an increased labor pool and predictive algorithms like ours, algae biofuels are moving closer to becoming a viable alternative energy source in world markets.

Our concentration on education allows our team's work to naturally incorporate diversity, equity, and inclusion. Despite its clear benefits, scientific research has traditionally been off-limits to millions of students across the country, excluding them from countless educational benefits. By specifically targeting middle schools and high schools across the country, our application will foster a demographic makeup of users that more accurately reflects the diverse makeup of the nation, an asset often overlooked in science.

As students ourselves and working with ISB's education branch that has nearly 2 decades of experience, we understand what works and what doesn't in scientific learning. As such, we've implemented and planned features - such as a public algae data repository and integrated experiment tutorials - that would be especially useful for those without prior research experience. Algaeorithm has already begun beta testing with 60+ students across two high schools in Snohomish County, Washington (Figure 9). These schools, using Chlamydomonas algae and multiple diatom strains, have so far adopted Algaeorithm as part of a bioengineering curriculum that reveals the applications of algae. These tests are just the beginning of a widespread push by our team to spread access to our platform in the 60,000 schools with over 2.6 million students in all 50 states that ISB's education team has already reached (Table 1).

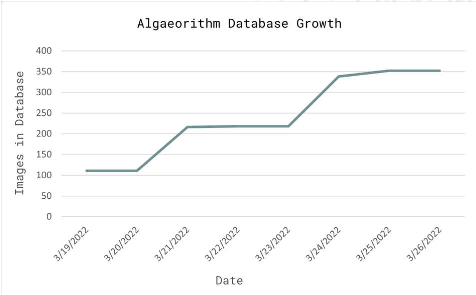


Figure 9. Algaeorithm Database Growth: Over 300 images were uploaded to the Algaeorithm database during the first week of beta testing.

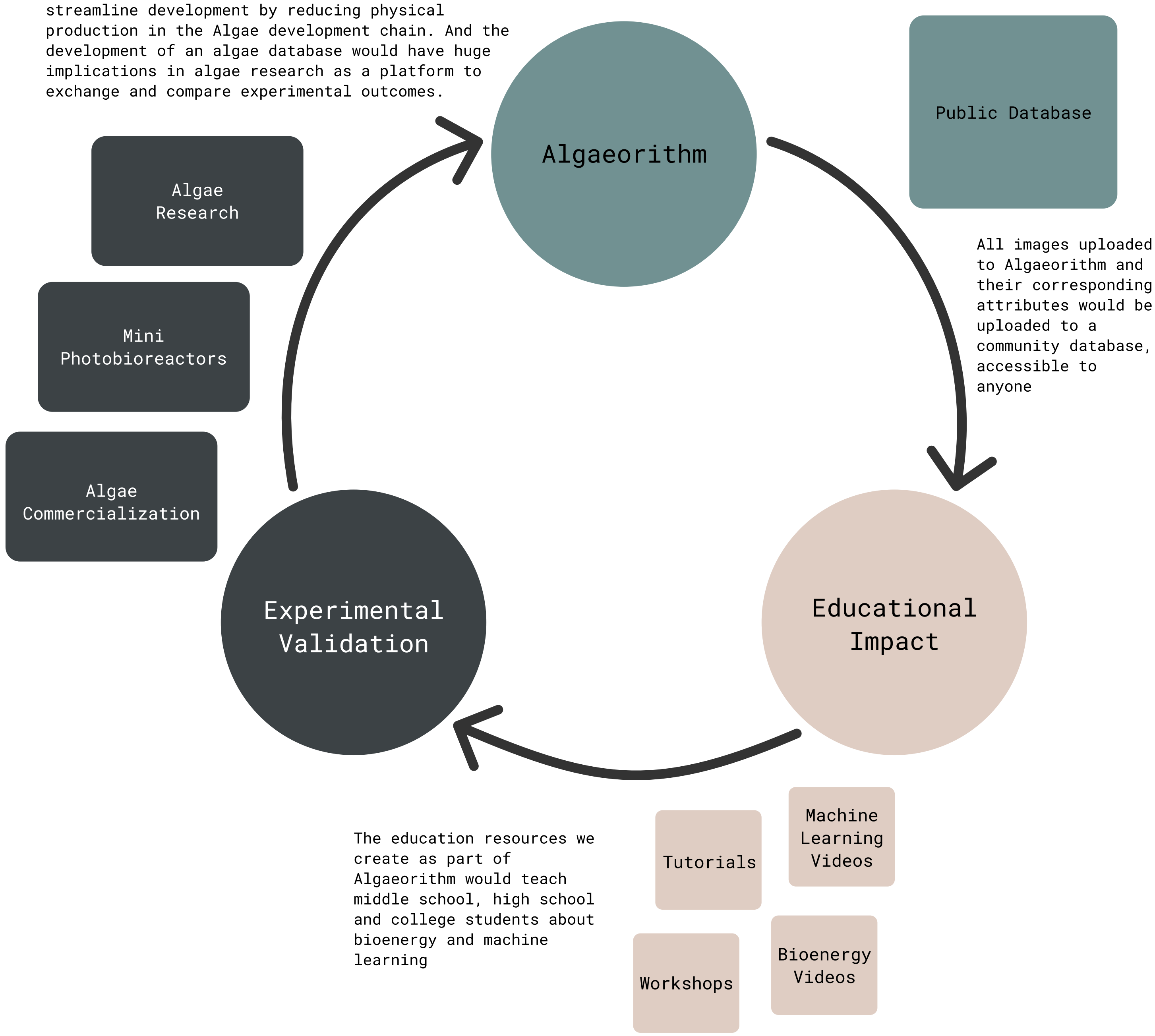
Project Engagement	Students	Algae Species	Schools	Database Images
Spring 2022	**60-100**	**2**	**2**	**350+**
Total 2022*	**200-400**	**3**	**6**	**1000+**
Total 2023*	**3,000 - 5,000**	**5**	**25**	**15,000+**

Table 1. Educational Impact: Current and projected application usage statistics. *Estimated future data using current teacher interest, growth trends, and ISB's education connections with schools and students from across the country

Algaeorithm is fundamentally a force multiplier; we take a set of tools and knowledge used by a select set of people and make it available to anyone for free. By multiplying access to these resources, our project would also multiply the number of opportunities to solve sustainability challenges. At the same time, it would also increase public engagement with algae, helping to incorporate everyone into the bioeconomy. Increasing the number of people learning about and working on algae research by multiple orders of magnitude would vastly improve the inclusion and diversity of the field, offer more equal opportunities to pursue experimentation and development, and result in a skyrocketing pace of change - this is our team's core vision.

Project Framework

With the help of experimental validation with mini photobioreactors, Algaeorithm could be expanded for use in Algae Research and Commercialization. More accurate predictive models could be used to streamline development by reducing physical production in the Algae development chain. And the development of an algae database would have huge implications in algae research as a platform to exchange and compare experimental outcomes.



Process Flow Diagram 2. Value Creation Cycle: The project consists of three elements, which together form a robust cycle. The application itself will contribute to educational impact through its use by students and teachers. Students and teachers using the application will contribute to the experimental validation by providing images for the training dataset. Finally, the experimental validation will contribute to the application by improving the algorithm with additional data. As depicted in the diagram, each element will also generate various offshoots which add to the algae value chain.

To give students in underserved communities the opportunity to participate in the frontlines of algae science, our application bypasses the myriad constraints and requirements of algae research that currently exist, which include prohibitively expensive equipment and production costs, access to Ph.D. faculty, and a well-thought-out organizational framework. In contrast, the sole resources necessary to utilize our team's tool are an internet-connected device and standard high school lab equipment. We categorically designed the application hand in hand with education programs, and because of this the resulting platform design includes numerous collaborative features, an easy-to-use user interface, and highly adaptive algorithms. These qualities are all very helpful when considering implementation in underserved communities, particularly in schools, and they would extend the ability to perform algae research from a small number of schools to nearly any community in the country.

Compounding this explosion of access to tools is access to a world-class network of scientists, academics, and peers. This network would include algae educational resources and curricula along with data repositories for analytical comparisons on a macro level. These are all resources traditionally out of the reach of underserved communities. However, the advent of more inclusive tools such as Algaeorithm would create interest in bioenergy within these communities and allow underserved students to become part of a dynamic network of scientists and community leaders at the forefront of the algae industry.

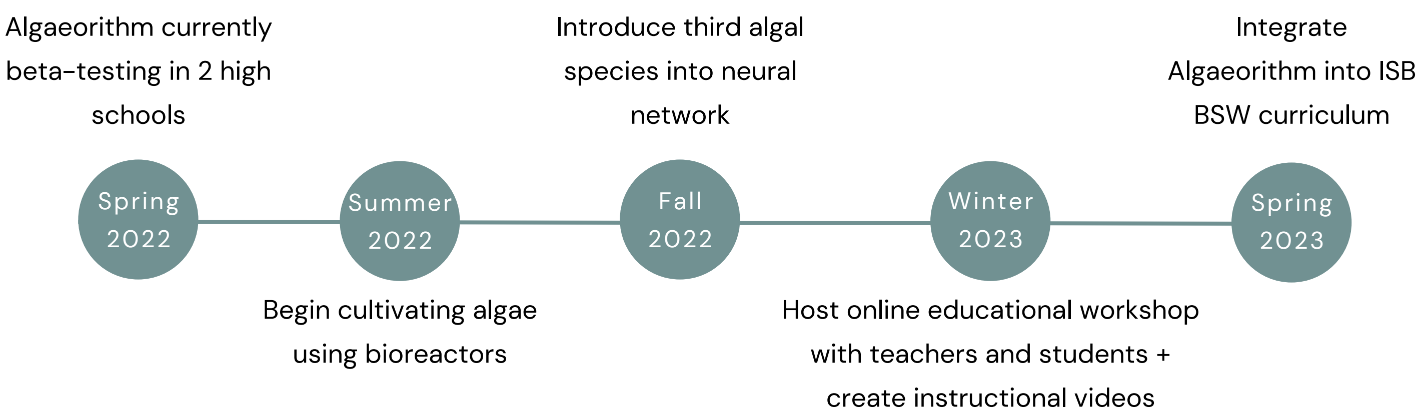


Figure 10. Timeline and Deliverables: Projected milestones for Algaeorithm over the course of the AlgaePrize competition.

Conclusion

Improvements in machine learning technology and bioenergy awareness will be crucial to initiate a multi-decade generational shift to a bioeconomy. As outlined in the synopsis, our project aims to address the challenges this transformation will present by heavily investing in the human capital that will form the foundation of this movement: students. We will support this goal through the achievement of several deliverables over the course of the project, which includes novel application features, various educational resources, and integration with existing curricula (Figure 10). These deliverables will play a key role in inspiring the next generation of students to realize a sustainable world, and ultimately help the world transition towards a future powered by biology.

Citations

Gottschalk, U. (2011). *Downstream processing*. Downstream Processing - an overview | ScienceDirect Topics. Retrieved April 4, 2022, from https://www.sciencedirect.com/topics/engineering/downstream-processing

Institute for Systems Biology. (2022, March 2). *Home*. Baliga Systems Education Experiences. Retrieved April 4, 2022, from https://see.isbscience.org/

Institute for Systems Biology. (2021, May 28). *Curriculum modules*. Baliga Systems Education Experiences. Retrieved April 4, 2022, from https://see.isbscience.org/modules/

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). {TensorFlow}: A System for {Large-Scale} Machine Learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283).

Pew Research Center. (2021, November 23). *Mobile fact sheet*. Pew Research Center: Internet, Science & Tech. Retrieved April 4, 2022, from https://www.pewresearch.org/internet/fact-sheet/mobile/

Pew Research Center. (2021, November 23). *Internet/broadband fact sheet*. Pew Research Center: Internet, Science & Tech. Retrieved April 4, 2022, from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Renotte, N. [Nicholas Renotte]. (2021, April 9). *Tensorflow Object Detection in 5 Hours with Python | Full Course with 3 Projects* [Video]. YouTube.

Valenzuela, J., Mazurie, A., Carlson, R. P., Gerlach, R., Cooksey, K. E., Peyton, B. M., & Fields, M. W. (2012). Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum. Biotechnology for biofuels, 5(1), 1-17.

Valenzuela, J., Carlson, R. P., Gerlach, R., Cooksey, K., Peyton, B. M., Bothner, B., & Fields, M. W. (2013). Nutrient resupplementation arrests bio-oil accumulation in Phaeodactylum tricornutum. *Applied Microbiology and Biotechnology*, *97*(15), 7049-7059.

López García de Lomana, A., Schäuble, S., Valenzuela, J., Imam, S., Carter, W., Bilgin, D. D., ... & Baliga, N. S. (2015). Transcriptional program for nitrogen starvation-induced lipid accumulation in Chlamydomonas reinhardtii. *Biotechnology for biofuels*, *8*(1), 1-18.

Imam, S., Schäuble, S., Valenzuela, J., López García de Lomana, A., Carter, W., Price, N. D., & Baliga, N. S. (2015). A refined genome-scale reconstruction of Chlamydomonas metabolism provides a platform for systems-level analyses. *The Plant Journal*, *84*(6), 1239-1256.

Acknowledgments

We extend our appreciation to the Institute for Systems Biology (ISB) & Silicon Valley Clean Energy for their support of Algaeorithm.