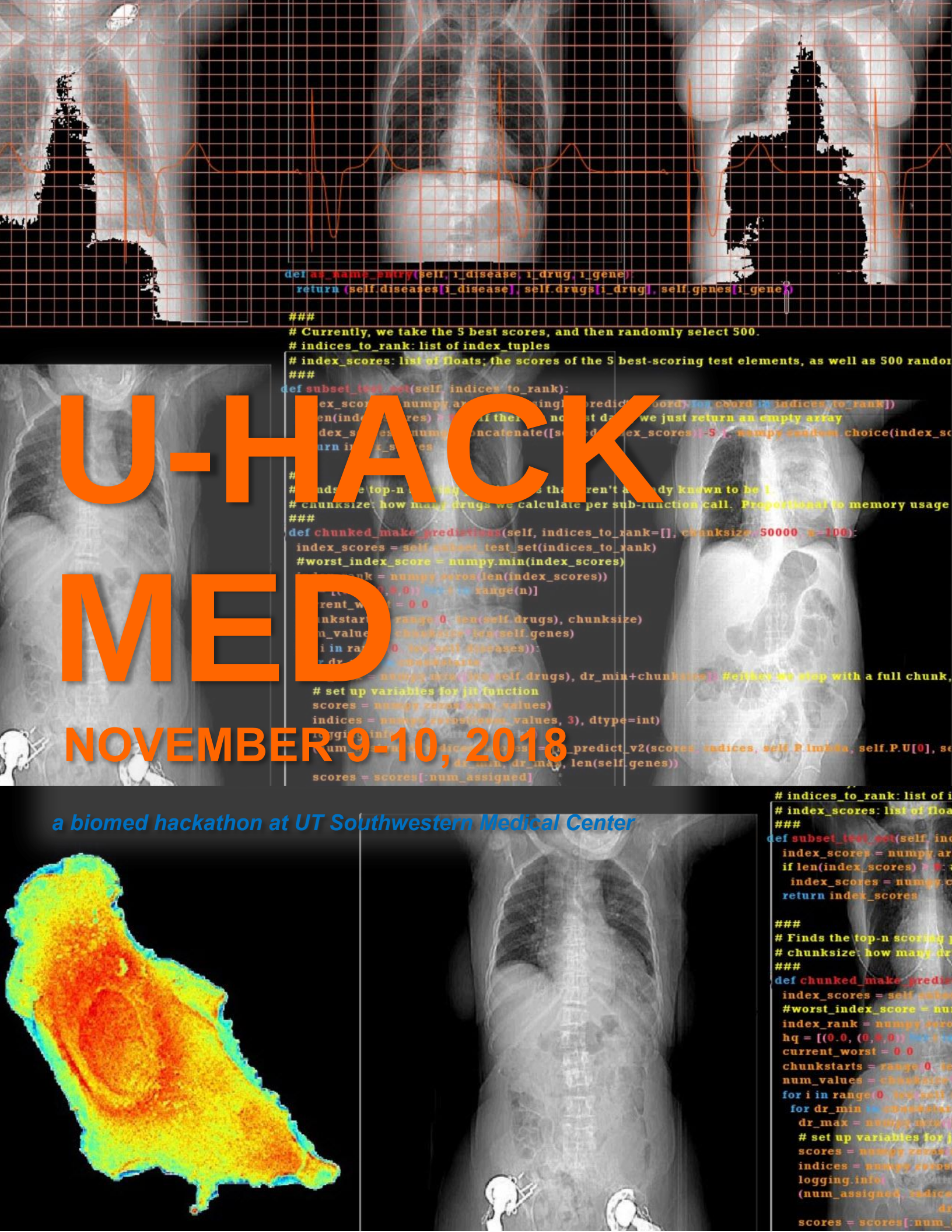


U-HACK MED

NOVEMBER 9-10, 2018

a biomed hackathon at UT Southwestern Medical Center

Contents

"This was a very stimulating, trailblazing event!"

–Carlos Arteaga, Director, Harold C. Simmons Comprehensive Cancer Center, UT Southwestern

Background

The Most Common Question

"What is a hackathon?"

That's the most common response encountered in preparing for this event, the first of its kind at UT Southwestern. A marathon of collaborative, exploratory computer programming best describes the occasion, in which teams of clinicians, programmers, software engineers, mathematicians and physicists – students and working professionals – came together not to run 26.2 miles, but to write thousands of lines of code to solve pressing clinical and biomedical problems in just over 26 hours.

By The Numbers

150	External applicants
12	Teams
55	Team members external to UT Southwestern*
96	Total attendees
$18,000	Total budget, including $10,000 earmarked from the Lyda Hill Foundation and $5,000 contributed by community partner Mark III Systems
$1,500	Cost per team, yielding biomedical innovation

* External participants were mostly graduate or undergraduate students, but also included research scientists and industry professionals
12 UT Arlington students & faculty
25 UT Dallas students
Other participants from Baylor, MIT, Rice, SMU, Texas Advanced Computing Center, UT Austin, University of Mississippi Medical Center, Virginia Tech, and even a local high school student

Team Formation

Rather than invite hacking teams to bring their own projects to pitch in a “Shark Tank” style as many hackathons do, the organizers met with clinicians and scientists across the university and curated 12 projects that lent themselves to computational solutions and showcased the breadth and quality of research at UT Southwestern. The event was widely promoted through a dedicated website and by our community partners, and “hackers” from around the nation were invited to apply to join these 12 teams. Participants were selected on the basis of their technical expertise, prior hackathon experience, and statement of interest. In addition to our UT Southwestern team leads (clinicians or scientists from across the university), we also assigned a Lyda Hill Department of Bioinformatics staff member to each team.

The event was supported by community partners Mark III Systems, the National Institutes of Health’s National Center for Biotechnology Information (NIH-NCBI), the Lyda Hill Foundation, and the Cancer Prevention and Research Institute of Texas (CPRIT). The NIH-NCBI has partnered with other institutions to host a few dozen hackathons around the nation over the past two years. In their experience, just in the weekend hackathon event, 80% of teams will produce an alpha or beta prototype (which can be further developed into a licensed or patented product) and at least 10% will publish their outcomes as a peer-reviewed manuscript. Our expectation is that this U-HACK MED 2018 event will produce such outcomes over the coming months and that this may be the beginning of an era where UT Southwestern, via such events, taps into a broad community of talent for solving some of its pressing biomedical data challenges.

Team Descriptions

Following are the twelve projects addressed by U-HACK MED 2018 teams, illustrating the diversity of scientific topics featured by the events. Full problem statements are available at: https://www.u-hackmed.org/2018teams/.

Team 1: Machine learning to distinguish T-cell receptor subregions

Mobilizing a patient’s own immune system to kill tumor cells is the current frontier in cancer treatment. At the same time, the early response of the immune system to the presence of a tumor in a patient can potentially become a powerful tool for cancer diagnosis. The design of such a reporter system hinges on identifying the specialized adaptive areas of T-cell surface receptors that controls the interaction of these immune cells with cells of a particular tumor. This team aims to develop a machine learning method to classify the adaptive receptor areas from about 20,000 receptor sequences of T-cells that have interacted with tumors and from about 10,000 receptor sequences of T-cells without interaction. Moving forward, this classifier may be able to detect the presence of tumors based on T-cell surface receptor sequences excised from patients with unknown cancer disease status.

Team Lead: Bo Li, Bioinformatics and Immunology

Background

Team 2: Optimizing a simulator of molecular interaction for a GPU pipeline

Graphical computer simulations of the interactions of molecules inside a cell are essential to understanding and manipulating these cellular processes. The general purpose “brain” of a computer is the Central Processing Unit (CPU), but a newer type of processor, originally developed for the computer animations and gaming industry, is the Graphics Processor Unit (GPU). GPUs can efficiently process large chunks of data, but require a different type of programming code. The goal of this project is to re-write and optimize a simulator of molecular interactions on a GPU instead of a CPU.

Team Lead: Khuloud Jaquaman, Biophysics

Team 3: Developing deep learning models in virtual reality

This team aims to develop a virtual reality (VR) environment for the ‘physical’ assembly of artificial neural networks. Deep Learning (DL) uses brain simulations (neural networks) to make transformative advances toward artificial intelligence. Training a computer in DL is not always intuitive, but the ability to visualize the computational process could help users rapidly develop “smarter” DL models. This team will design a virtual reality tool that will visually connect the user to the computer architecture.

Team Lead: Murat Can Çobanoğlu, Bioinformatics

Team 4: Simulating molecular dynamics of potential drug target regions

The goal of this project is to simulate the binding of cancer drugs to protein targets by identifying the shape and location of “pockets” on these proteins that are mutated in patients who have cancer. These pockets often come and go over time, but if identified, may be stabilized by the binding of drug molecules. Using computer simulations to understand the structural features of these pockets may lead to development of targeted drug molecules that stabilize the pockets and kill cancer cells.

Team Leads: Mike Trenfield & Milo Lin, Green Center for Systems Biology

Team 5: Unbiased computational strategy to identify distinct populations

Recent advances allow us to sequence the genetic material inside single cells and even within individual components of a cell, providing unprecedented level of detail for researchers. Thousands of cells are pooled and sequenced in this methodology, and then the data for single cells is identified by a genetic “cell barcode” inserted during processing. Human brain tissue includes several different cell types with more complexity than with other types of tissues. This team aims to develop a computational strategy that can distinguish brain cell types (neurons and glia) from other cellular debris and background noise in the sequencing data.

Team Lead: Fatma Ayhan, Neuroscience

Background

Team 6: Artificial neural networks to predict ECMO-related neurologic injury

The most critically ill children in pediatric intensive care units must often receive ECMO therapy, that allows a temporary bypass of the heart and lungs so that those tissues can rest and heal. However, serious neurologic injury, such as stroke or brain bleed, occurs in approximately 15% of children who receive ECMO therapy. Our goal is to develop a machine learning prediction tool that can identify in advance which patients are likely to experience neurologic injury as a result of ECMO therapy.

Team Leads: Neel Shah & Abdelaiziz Farhat, Pediatric Critical Care

Team 7: Deep learning cell annotation framework for tumor classification

Many research teams are currently working to automate the characterization of tumor biopsy images to speed up the process of diagnosis and clinical decision-making for cancer patients. However, the deep learning models being employed produce massive libraries of image annotations that must be properly stored, visualized and edited in order to be useful. We will build an efficient annotation framework to visualize and edit tumor biopsy images with >100,000 annotations.

Team Lead: Satwik Rajaram, Bioinformatics

Team 8: Differentiation of metastatic potential in melanoma cells

The prediction of whether a particular cell belongs to a tumor with high vs low metastatic behavior is very complex. We aim to develop a computational solution that will allow the computer to differentiate between melanoma cancer cells with high and low metastatic properties and to present the data in a graphical format. This could be employed in a diagnostic pipeline to complement existing genomic tumor characterization, which often offers limited precision in assigning a tumor to particular cancer subtypes to guide treatment decisions.

Team Lead: Andrew Jamieson, Bioinformatics

Team 9: Machine learning for brain and cardiac MRI feature extraction

The project centers on the development and application of a machine learning tool for the analysis of medical imaging data. This tool will automate the analysis of MRI features, making diagnosis or prognosis of a patient's condition more rapid and more accurate.

Team Leads: Albert Montillo, Bioinformatics & Ashoo Tandon, Pediatric Cardiology

Team 10: Visual dashboard and predictive analytics for pediatric endocrinology

Children with juvenile diabetes often wear blood glucose sensors for continuous monitoring. However, the data produced by these different brands of medical devices cannot be easily integrated into the electronic health record for analysis in the context of the patient's other medical data. For each patient clinic visit, up to 15 minutes of the 20-minute appointment may be devoted to gathering this data in a format that the physician can digest and make treatment decisions. But even this

does not allow a physician to compare data for their entire population of patients in order to identify trends. Our goal is to develop visual dashboards for patient population data and tools for predictive analytics to enable quick, efficient risk stratification of individual patients in a clinical setting.

Team Lead: Soumaya Adhikari, Pediatric Endocrinology

Team 11: Natural language processing tool for clinical trial inclusion criteria

At present, determining which clinical trials a patient may be eligible for entails an entirely manual process of matching patients with criteria for thousands of possible clinical trials. Our goal is to develop a natural language processing tool to automate this process. By training a computer model to scan the text of clinical trial descriptions and classify certain descriptors, physicians can then use the tool to search and filter the database of potential trials.

Team Leads: Jeffrey Gagan, Pathology and Brandi Cantarel, Bioinformatics

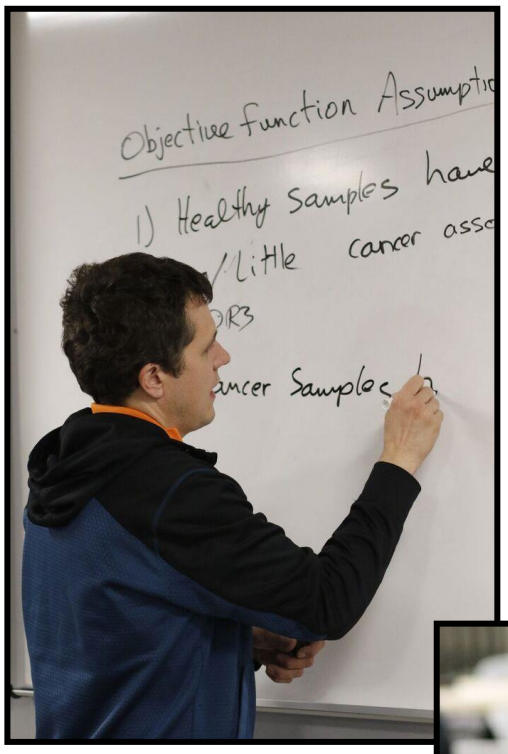
Team 12: Machine learning algorithm for staging of liver disease

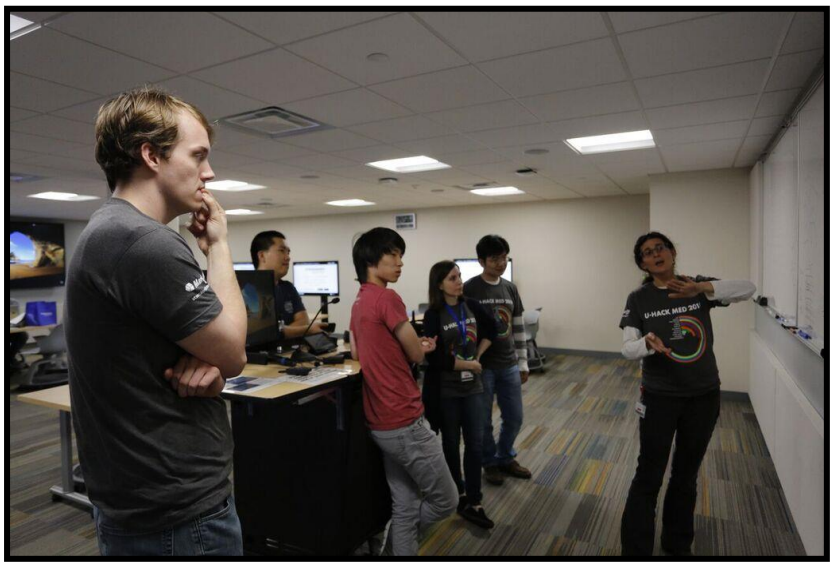
For patients with fatty liver disease, the current method of determining the disease stage and prognosis is by liver punch biopsy. This has a possibility of complications and because it takes only a small sample, the diagnosis can be inaccurate. The focus of this project is to develop a machine learning algorithm that can use the lipid data from a patient blood sample to provide more accurate staging for the disease, using a method that is much less invasive for the patient.

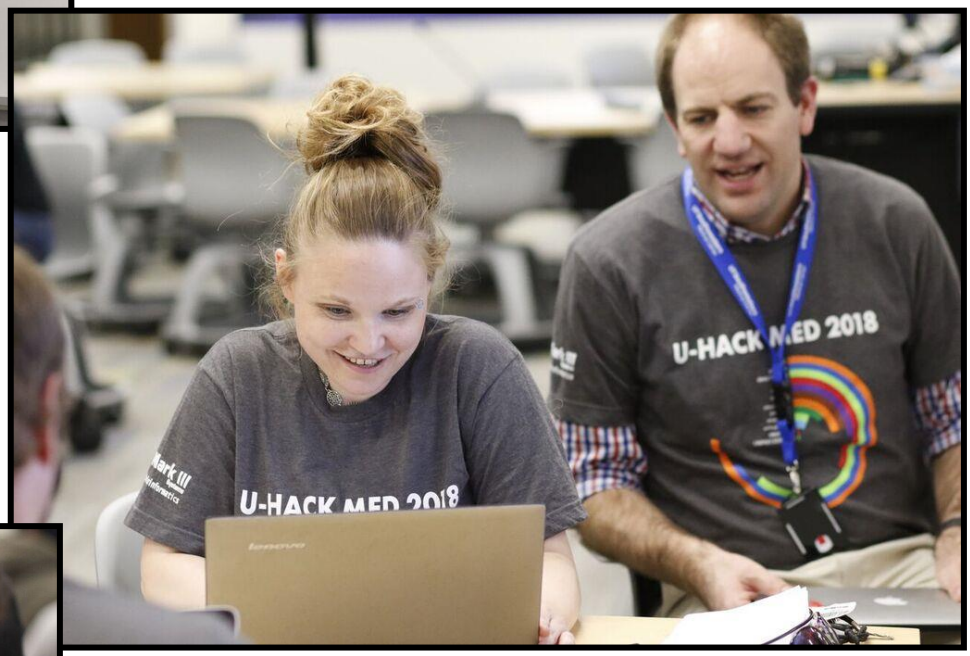
Team Lead: Jeffrey McDonald, Molecular Genetics and Alex Treacher, Bioinformatics

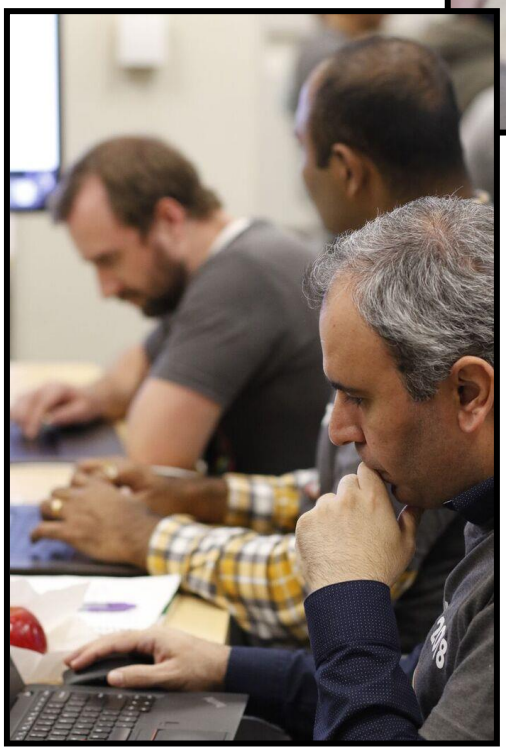
"U-HACK MED is a way to showcase the diversity of research questions we have at UT Southwestern and to bring interest from other universities and industry partners. We wanted to break down the silos at UT Southwestern and promote interdepartmental research and we were able to make progress towards that goal thru this event. It is also a way to highlight career paths for undergraduate and graduate students in the greater Dallas-Fort Worth area."

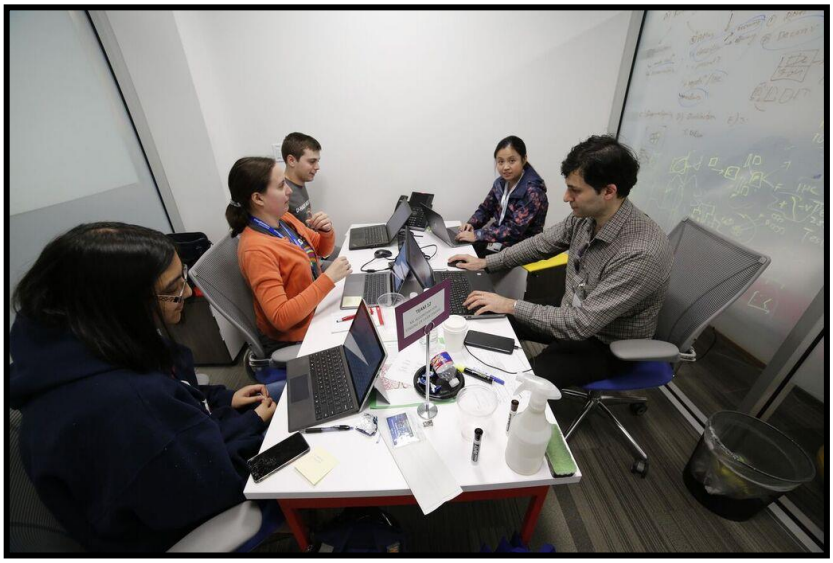
--Venkat Malladi, Director, Bioinformatics Core Facility


Objective function Assumptio
1) Healthy samples have
little cancer asso
cancer Samples




U-HACK MED 2018
U-HACK MED 2018





Judging & Awards

Judging Criteria

Technical Merit:	How elegant is the solution? Was a particularly difficult design challenge met? How polished is the Hack? How complete is the Hack, as presented? What work was actually accomplished vs what was envisioned?
Innovation:	Is the hack tackling a new problem? If an existing problem, is the hack approaching it with a significant new idea / technology?
Clinical Application:	Is there a clear potential clinical care application? Will a patient benefit from this product over the current standard of care? Does it contribute to patient care improvements in a clear and tangible way? How big is the potential benefit to patients if the project was developed?
Feasibility:	Is it realistic to develop the hack further? Is the hack usable in its current state? Is the hack easy to understand? Did the presentation convince you there is a viable product?
Wow Factor:	Did the outcome of the hack surprise you? Was the presentation exciting? Is there impressive use of cutting edge technology?
Visualization:	Is the visualization impressive? Does the visualization clearly present the data it is based on? Is the visualization accessible, using good design practices? *(e.g. easy to read text, consider color-blindness etc.)*

Awards

AWARD	CRITERIA
Lyda Hill Award	Highest Total Score, excluding Visualization
Most Out of the Box	Highest Innovation Score
Best Desk to Bedside Project	Highest Clinical Application Score
Most Likely to Become a Startup	Highest Feasibility Score
Next Sci-Fi Blockbuster	Highest Wow Factor
Best Use of Visualization	Highest Visualization score

Judging & Awards

Jury Members

Dr. Carlos Arteaga, Chair of the Jury
Professor and Director, Harold C. Simmons Comprehensive Cancer Center
UT Southwestern Medical Center

Dr. Robert (Bob) Toto
Professor and Associate Dean, Clinical and Translational Research
UT Southwestern Medical Center

James E. Canady
Sr. Software Developer
IBM Watson Health

Lauren Bui
VP Data Management and Analytics
CHRISTUS Health

Joe Riss
Solutions Architect
Hewlett-Packard Enterprise

Brandon Draeger
Director of AI Enablement for Americas and Europe
Intel

Dr. Gaudenz Danuser
Professor and Chair, Lyda Hill Department of Bioinformatics
UT Southwestern Medical Center

"I was very impressed with all the teams' presentations. It was easy for me to understand and "see" the many potential applications of the methods for discovery science, clinical and translational research, and patient care. The most exciting aspect of the hackathon was the interdisciplinary team approach to the challenges. Many of the teams included clinicians and computer scientists working together to solve important clinical problems. This is something that I don't see enough of on our campus." –Robert Toto, Assoc. Dean, Clinical & Translational Research, UT Southwestern

"The outcome of this event has been beyond my greatest expectations. Of the 12 projects, 8 to 10 emerged from the two-day hackathon with actionable outcomes. The clinical projects in particular did very well. We can now collaborate further with these teams to continue developing and polishing their solutions to the point of bringing these innovations to the patient bedside."–Gaudenz Danuser, Chair, Lyda Hill Department of Bioinformatics, UT Southwestern

Outcomes

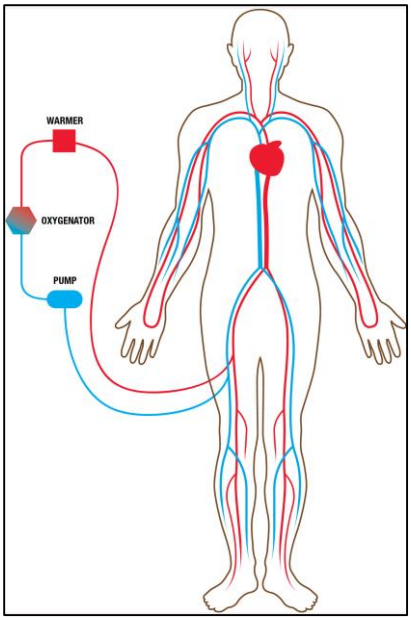
All teams achieved measurable progress toward their goals. Of these, three team outcomes are featured in this report. These three projects illustrate the types of tangible and immediately applicable outcomes that may result from events such as this: an alpha-prototype solution to a pressing clinical problem; significant progress on a computationally sophisticated problem; and a failed test that leads the team to pivot to a different long-term solution. Outcomes reports for the other nine projects will be published online (https://www.u-hackmed.org/outcomes/) as they are assembled. All software codes and project presentations will reside in a GitHub repository organized by the National Institutes of Health (https://github.com/NCBI-Hackathons) where they are accessible to the online community for further development.

Team 6: Winners of the Lyda Hill Award

Artificial Neural Networks to Predict ECMO-related Neurologic Injury

Extracorporeal membrane oxygenation (ECMO) is a form of cardiopulmonary bypass used as a life-saving therapy in the intensive care unit. ECMO involves drainage of a patient's blood from the central venous system, pumping it via an artificial pump through devices that allow for addition of oxygen and removal of CO2. Blood is then warmed and pumped back to the patient.

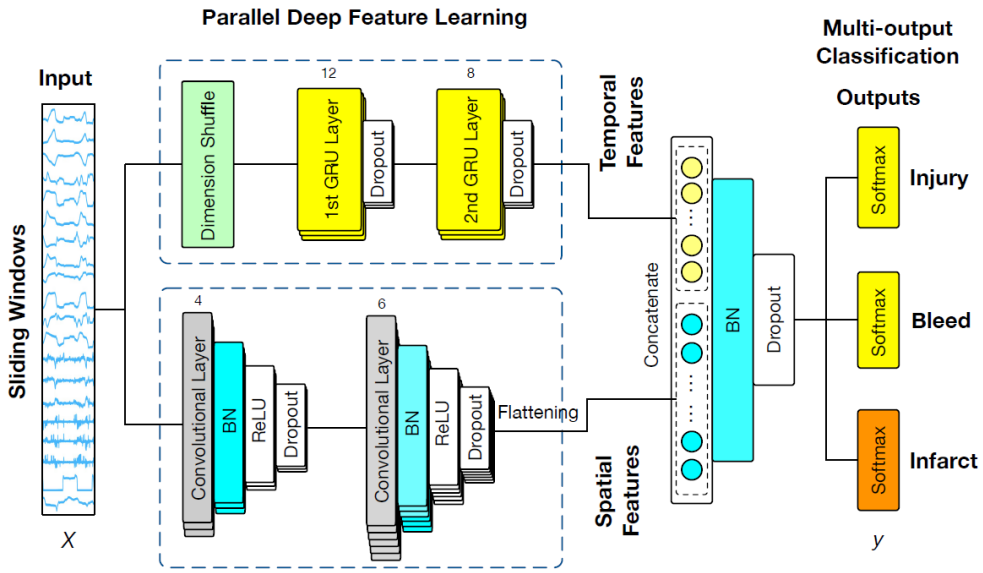
Neurologic injury is a significant burden in this field. Even with rapidly expanding use of ECMO, the occurrence of neurological injury has remained the same, with some estimating the incidence as high as 33%. Overall survival in ECMO is close to 60%, but survival drops by half in the setting of neurological injury. Survivors have multiple long-term morbidities, including seizures and global developmental delay. Because of the many intertwined factors involved in management of ECMO patients, it is hard to predict which patients develop these injuries. We hypothesized that an artificial neural network will be able to predict outcomes.



We retrospectively obtained hourly data of 141 pediatric ECMO patients treated at the Children's Medical Center with known outcomes (one of the largest such data sets in the U.S.). Over 40 different types of clinical and lab parameters were collected for each patient every hour. This included ECMO mechanical data, vital signs, acid-base homeostasis lab values, coagulation lab values, and laboratory markers of end organ perfusion. Data collection started at 24 hours prior to being started on ECMO, and lasted throughout the course of this therapy. The outcome analyzed was presence or absence of

intracranial injury documented on head imaging. A secondary outcome we sought to predict was type of injury, e.g., bleed vs infarct.

We analyzed 141 patients with head imaging results, totaling to more than 15,000 hours of clinical and laboratory data. The fast processing of such large data in short enough response times to allow development and testing of numerous variants of computational models predicting the relation between ECMO data and patient outcome was enabled by the powerful data storage and computing infrastructure provided by UT Southwestern Medical Center's High-Performance Computing team (BioHPC). Due to the varying lengths of ECMO duration across patients, as well as irregular sampling and missing data, it was initially difficult to build and train a predictive model. Our first task was to standardize some of the medication doses as well as clinical data by adjusting to a weight-based number for each patient. To fill in missing data points, we utilized linear regression and inserted the resulting average values in the gaps. On to the modeling - we first applied a recurrent neural network (RNN) to our data, but despite data clean-up, the overall predictive accuracy was merely around 70%.



In attempt to increase the predictive power, we adopted a hybrid deep learning model, integrating a convolutional neural network (CNN) and an RNN, a strategy often pursued in the field of Artificial Intelligence. We hypothesized that such a setup would simultaneously learn spatial information and temporal dynamics within a moving window, and improve our predictive accuracy. After we optimized model parameters, and experimented with different sizes of the moving window, our model showed outstanding prediction accuracies—90% for any injury, 93% for brain bleed, and 89% for infarct (stroke).

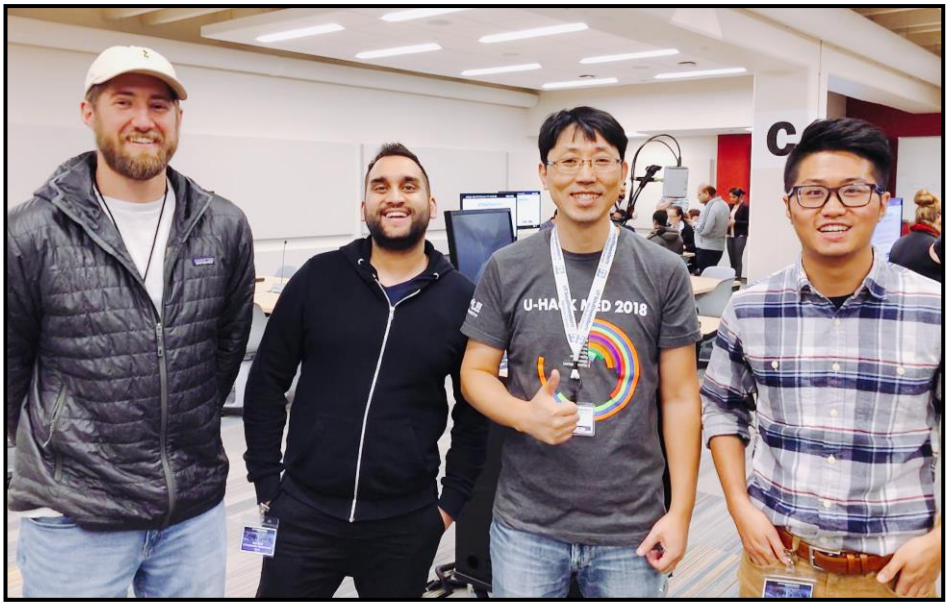
Outcomes

Additional Comments

For each pediatric intensive care patient, we are collecting the same amount of data as that collected for a Formula One race car. But unlike the immediate data analytics that lead racing teams to improve performance every weekend, at present no such predictive scoring tools exist for ECMO patients. We have masses of data but cannot make sense of it without a computational solution. Our team approached this problem asking whether the data paint a picture that allows the machine to discern and predict significant medical complications (like neurologic injury).

This team's solution showed, with impressive accuracy, that the machine can indeed forecast medical complications. The immediate next step for this team is to produce a high-impact publication in the next 3 to 4 months and then apply for grant funding for further study, so that this technology can be brought as an online decision support into the Intensive Care Unit of our hospital. To achieve this, the intermediate goal will be to collect more data and determine whether this analysis model can be applied prospectively versus the retrospective data analysis done in this hackathon. Another goal will be to evaluate patients who are in the "standby" phase of possibly, but not yet, receiving ECMO therapy. It may be possible to derive from the hackathon model an early prediction model that will guide the clinical decisions such as if and when a patient should receive ECMO therapy.

Team 6 (L to R):
Rafe McBeth, UTSW Medical Physics Resident
Neel Shah and Abdelaziz Farhat (not pictured), UTSW Pediatric Critical Care Fellows
Jeon Lee, UTSW Bioinformatics
Ziheng Wang, UT Dallas PhD candidate, Mechanical Engineering



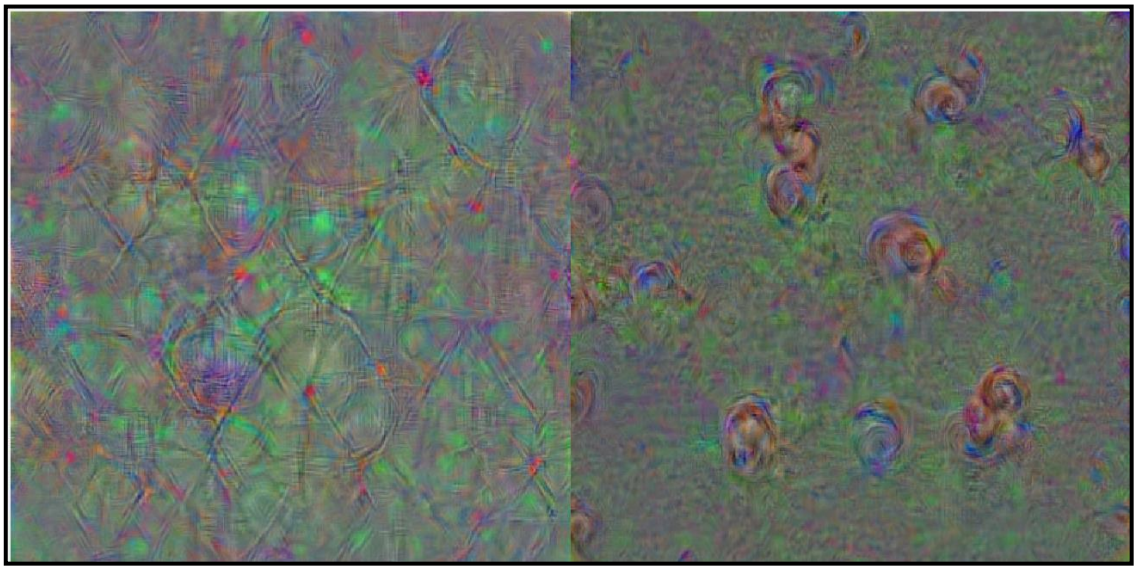
"This event showcased the compute capabilities of our BioHPC as a truly outstanding computing infrastructure in the academic space. Our teams did not experience any of the common technical roadblocks that are often encountered at other hackathons." – *Gaudenz Danuser, Chair, Lyda Hill Department of Bioinformatics, UT Southwestern*

As part of the *Lyda Hill Award*, each team member received: 1) one year of access to the UTSW BioHPC supercomputing infrastructure to continue working on innovative projects; 2) certificate for a mentoring lunch with UTSW Lyda Hill Department of Bioinformatics faculty 3) a Coursera course; 4) a $50 gift card for purchase of a textbook; and 5) a Lyda Hill Department of Bioinformatics hoodie sweatshirt.

Team 8: Award for Best Use of Visualization

Our hacking goal was to differentiate high versus low metastatic potential melanoma by observing live cell activity captured in 4-hour-long label-free phase contrast microscopy videos. With less than 30 hours to work, fueled by a passion to fight cancer, and undeterred by the massive 6 terabytes of data to crunch through, our determined team delivered the *RU Meta* (read: *Are you meta[static]*) software package.

Two key breakthroughs led the way to a prototype diagnostic pipeline: first, collapsing the massive raw wide-field 720 frame videos filled with multiple cells moving about into a single maximum projection. Second, deploying Deep Learning + Transfer Learning to leverage image feature knowledge learned from previously trained neural networks (AlexNet) to our problem. After training the adapted supervised convolutional neural network (CNN) on our images labeled as high or low, we achieved a surprising 92% prediction accuracy.



Unsatisfied with numbers alone, we sought to unveil the rich, hierarchical image representation encoded within the deep learning network, and demystify how, exactly, the network determines what is high vs. what is low metastatic potential. Mesmerizing visualizations capturing the intricate features learned by the model, were generated by an open source tool called *DeepDream*. More specifically, *DeepDream*, finds input images that, when fed into the model, most powerfully activate particular neurons throughout the deep learning network. By focusing on neurons at the last layer, i.e.,

Outcomes

those corresponding to the high and low metastatic cells respectively, we can essentially “ask” the network what it “thinks” high vs. low metastatic melanoma looks like. These visualizations led us to discover that the network “imagines” high and low cell images to appear (and potentially behave) very differently!

As seen by the stark crisscrossing repeating line patterns (on the left in image above), these “dream” images tell us that high metastatic cells are likely more active over the video time frame, creating long spiculated extensions across the time-projected images. Conversely, the visualization on the right for low metastatic cell images shows distinct and compact spiral/nautilus-like objects, implying cells are more stationary and limited in motion over the video time frames.

Visualization of these abstract behaviors, which escape any human observer, garnered the team the award for ‘Best Visualization’.

Team 8 (L to R): **Stefan Daetwyler**, UTSW Bioinformatics **Elizabeth Zou**, Massachusetts Institute of Technology, Computer Science undergraduate **Mahmoud Elgenedy**, UT Dallas PhD student, Electrical Engineering **Andrew Jamieson**, UTSW Bioinformatics **Colin Brochtrup (not pictured)**, UT Dallas PhD student, Electrical Engineering

Additional Comments

With this project we intended to benchmark the classification accuracy of an existing computational pipeline in Gaudenz Danuser’s lab with an alternative approach. In the existing pipeline, developed by team leader Andrew Jamieson and a postdoc over almost four years, the classifier relies on deep learning of individual cell behaviors. While this does account for the heterogeneity among cells from a single tumor, the prediction of whether a particular cell belongs to a tumor with high vs low metastatic behavior is much more complex. The best solution to this question currently offers 70 – 75% accuracy.

In contrast, the *RU meta* approach ignores the variation between individual cells and asks a simpler question, does the population of cells filmed in one movie belong to a tumor with high vs low metastatic behavior? In only 26 hours, an image processing

pipeline was assembled that answered this question with astonishing accuracy – no human observer could make this call with comparable consistency. Moving forward the best performance will likely be achieved by combining the two approaches. The resulting hybrid will benefit from the tumor-centric perspective of *RU meta* and the cell-centric perspective of the more complex existing pipeline. In parallel, the Danuser lab will seek to partner with local and international clinical collaborators to test drive the new hybrid classifier for actual application in a diagnostic pipeline. This would be particularly powerful for cancers with a broad mutational spectrum, like melanoma, where a genomic characterization often offers limited precision in assigning a tumor to particular cancer subtypes for a targeted treatment. This outcome also showcases how the creative space of a hackathon offers a platform for complementing a well-defined research project with fresh and disruptive ideas.

As winners of the *Best Use of Visualization* award, each team member received 1) Think Geek mechanical model kits; 2) certificate for mentoring appointment with Ryan VanAlstine, Chief Technical Officer and Development team lead at Mark III Systems.

Team 11: Valuable Lessons Learned

For this project we wanted to tackle the problem of clinical trial annotations. There are certain elements that are important to annotation of clinical trials including: 1) drug, 2) disease and 3) biomarker. The data source for this project is the NIH Clinical Trials Database (clinicaltrials.gov). The descriptions of clinical trials are mostly in free text with a few sections that contain key information including: Condition, Intervention, Eligibility Criteria, and Detailed Description. Condition usually describes the disease, however, these names do not follow a disease ontology and there may be multiple naming variations for the same disease or other confounding factors. Intervention will include the name of any treatment, including drug treatment, surgeries, devices, etc. Biomarkers, used to test for presence of a disease, might be contained in the Detailed Description or Eligibility Criteria. Using the simplest three elements, we decided to employ a natural language processing (NLP) method to predict the disease referenced in a clinical trial description. Using a database of diseases and synonyms, we used this model to predict the official disease name. This did not work as we hoped and was almost an entire day's effort.

Because the NLP approach did not work well with disease names, we decided to use exact matching to identify the biomarkers or genes that were targeted for treatment. We were much more successful with this approach, with a good accuracy, but with a relatively high number of false positives for genes with short names. In all, our team learned that in order to develop a good NLP model, we must have a large data set that is already "clean" on which to train the model. As a next step, we will create a number of manually curated clinical trial description data sets for use in developing automated annotations.

Participating in the hackathon was a useful experience in that it forced the team to sit down and really tackle this problem. On a normal day, we might try to work on it, get stuck and then decide to come back to it later because we have so many things on our plates. This allowed us to test our initial approach, realize that it was not going to work, and then pivot to engage a better long-term strategy. –Jeffrey Gagan, Medical Director, Clinical NGS Lab, UT Southwestern

Additional Comments

At present, determining which clinical trials a patient may be eligible for entails an entirely manual process. Most cancer centers have dedicated staff that are responsible for maintaining a list of clinical trials available at that center or partnering institutions. This is usually a spreadsheet with many columns describing the disease, treatment, and all eligibility criteria. A research nurse must then compare the individual patient's information with this spreadsheet and call each clinical trial coordinator to determine if they are enrolling patients at that time.

Natural language processing would be one method to automate this process, by training a computer model to scan clinical trial descriptions and classify certain descriptors so that they could be searched and filtered. There are tens of thousands of clinical trials, however, there is no universal standard language for clinical trial descriptions and it would be difficult to develop one. Disease ontology, or formal naming system, is complicated not only by synonyms and abbreviations, but also by different levels of granularity. For example, non-small cell lung carcinoma encompasses both lung squamous cell carcinoma and lung adenocarcinoma, and any of these names may be used in a clinical trial description. The same applies to biomarkers. Sometimes a description will not name a specific mutation in the bulleted criteria, but upon reading the description it becomes apparent they are only enrolling patients with a specific activating or loss of function mutation.

The magnitude of this problem led us to test this natural language processing approach in the hackathon. We learned from this test that we must first have a very large, already cleaned training data set. As a next step towards this goal, a curation scientist in the Department of Pathology's Clinical NGS Lab will attempt to take only the clinical trials that are active at UT Southwestern, "clean" the description data and input them into this model to train an NLP classifier. The expectation is that the classifier tool can then ingest clinical trial descriptions that it hasn't seen before (from the public repository, clinicaltrials.gov).

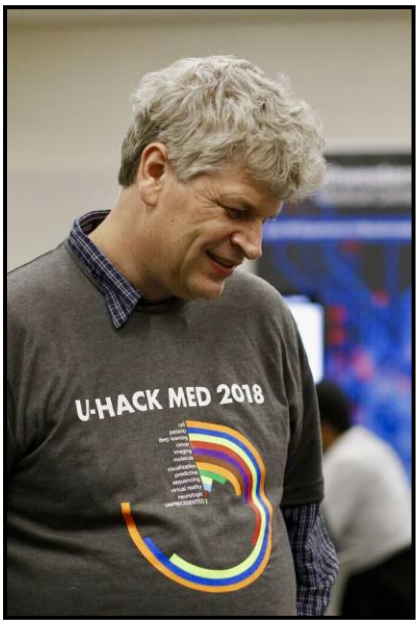
Team 11 Members: Jeffrey Gagan, UTSW Pathology; Karan Patel, UT Dallas Computer Science graduate student; Anant Ahuja, UT Arlington Computer Science graduate student; Benjamin Wakeland, UTSW Immunology; Brandi Cantarel, Qiuyan Shao & Daniel Moser, UTSW Bioinformatics

Concluding Statements

The rise of UT Southwestern to a leading academic medical center has been driven primarily by the excellence of individuals. With U-HACK MED we strive to convene a program for an era in which our institution's reputation will be complemented by excellence of the collective. For the first event in the U-HACK MED series, we had three goals: 1) Rally parts of the UTSW community around a set of data analytical problems in basic research and clinical practice, for which the solutions will depend on the inspiration of the collective; 2) Demonstrate the power of crowd science in spearheading such solutions; 3) Deliver for each of the analytical problems a robust thread that can be followed by the teams for the development of complete solutions. We believe we have reached every one of these goals. The level of motivation to continue and improve the U-HACK MED series with future events is high.

What remains for this report is to acknowledge the institution at large for a greatly supportive atmosphere, Drs. Podolsky and Malter for their presentations, the partnerships we forged within UTSW to implement the event (Information Resources, Medical School Team-Based Learning facility, Faculty Club), our community partners, including the Lyda Hill Foundation, Mark III Systems, NIH-NCBI and CPRIT (RP150596), who critically supported this event, the members of the award jury, and above all, all the participants for their enthusiasm and dedication.

Contact Information



Gaudenz Danuser
Professor and Chair
Lyda Hill Department of Bioinformatics
Gaudenz.Danuser@UTSouthwestern.edu

Venkat Malladi
Director, Bioinformatics Core Facility
U-HACK MED Technical Lead
Venkat.Malladi@UTSouthwestern.edu

Rebekah Craig
Business Analyst
U-HACK MED Logistics Lead
Rebekah.Craig@UTSouthwestern.edu