In July 2017, Reactor issued a data ideation challenge: “Good Questions Meet Big Data.” Applicants were asked to identify a human health problem that might be resolved using big data and a computational solution. Entries were required to be problems that could be addressed with clinical and translational research in areas such as diagnostics, therapeutics, public health, technology, or outcomes.

A total of $10,000 in prizes were awarded to 11 winners.

**Optimizing Antibiotic Therapy through Machine Learning Models**
Sanjat Kanjilal

Antimicrobial resistance (AMR) is among the most challenging problems facing modern medical care and is associated with increased morbidity, mortality and economic costs. In addition to reducing unnecessary prescriptions, an important part of preventing AMR is optimizing the use of existing antibiotics. While national guidelines and antibiotic stewardship programs provide general guidance on the management of many infectious syndromes, they are not personalized to the history and exam of a given patient and do not explicitly account for the impact of their recommendations on the future development of AMR in the patient. We propose to apply machine learning algorithms to the problem of antibiotic treatment optimization on a cohort of patients who presented with bacterial infection over an 18-year period in Boston at the Massachusetts General Hospital and the Brigham & Women's Hospital.

Our first objective is to estimate the probability that a patient presenting with the signs and symptoms of an infection has a pathogen that is resistant to a set of common antibiotics. We generate a representation of each patient that captures their risk factors for AMR, immune status and most likely infectious syndrome and then train a set of machine learning models including deep neural networks to predict their likelihood of resistance given an infectious syndrome. We will compare our results to traditional logistic regression models.

Our second objective is to estimate the impact of antibiotic treatment on the likelihood of AMR at a future date. We limit our predictions to one, three and six months post-treatment. Using the patient representations generated by our first model, we will produce estimates that account for missing data through multiple imputation and use propensity scores to account for confounding by indication. Our outcome fills an important gap in clinical practice by helping providers choose the antibiotic treatment that is least likely to result in an antibiotic resistant infection in the future.

The activities described in this proposal represent the first attempt at applying machine learning models to this important clinical problem and have the potential to significantly change clinical
practice in outpatient and inpatient settings. Future work will focus on expanding our datasets and using our models as the basis of a clinical decision support tool to provide real-time personalized antibiotic treatment recommendations.

Innovative Tumor Response Forecasting Model to Predict Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Patients
Ramya Palacholla and Nils Fischer

We propose using a comprehensive clinical database to identify robust predictors of tumor response for a given neoadjuvant treatment and, then, use this knowledge to develop an innovative predictive model based on machine learning techniques to optimize and personalize treatment regimens for advanced breast cancer patients and prevent over-treatment. In this era of precision medicine, the proposed tumor response forecasting model has the potential to change the current paradigm of clinical care for advanced breast cancer patients and reduce healthcare costs. The most interesting implication of this model is the ability to predict tumor chemosensitivity and prognostic impact of a specific neoadjuvant treatment regimen in individual breast cancer patients. Furthermore, this model can be used within clinical oncology with the potential to expand to other disease areas of application including other types of cancers and diabetes.

The Challenge of Cancer Genomic Data Interpretation
Jeremy Warner

Genomic information pertaining to cancer diagnosis and treatment is made available in clinical settings thanks to established technologies such as fluorescent in situ hybridization (FISH), quantitative polymerase chain reaction (PCR), gene-expression and next-generation-sequencing (NGS) panels, among others. With improvements on these technologies, and the ever-expanding knowledge base, genomic information is growing in quality, amount, and complexity. The field of precision oncology relies on the management and interpretation of genomic data, such that current issues of interest include interoperability and definition of common standards, efficiency in workflow and ensuring the granularity of data, and the development of clinical decision support tools. Cancer panel NGS, in particular, encompasses a vast knowledge space that far exceeds individual cognitive capacity, and could benefit from publicly available knowledge to support decision-making by physicians and to answer patients questions. The fact that approximately two-thirds of NGS cancer panel results are Variants of Unknown Significance (VUS), illustrates the problem we face: there is a non-negligible amount of genomic information that remains uncharacterized. Thus, an important step in furthering precision oncology consists of drawing from publicly available knowledge bases that synthesize cancer variant interpretation information, with the goal of identifying and resolving discrepancies in genomic data interpretation.

Leveraging public gene expression data and machine-learning for efficient, high yield differentiation of iPSCs into therapeutic cell populations
Surojit Biswas

In vitro differentiation of patient derived iPSCs holds great promise for personalized disease modeling and cell replacement therapy providing a replete source of patient- and disease-
specific human cells. From deep public gene expression databases, we can machine learn which transcriptome modulating perturbants are available and how they can be used to modify the iPSC transcriptome. We aim to then propose succinct differentiation programs that apply these perturbants strategically in serial or parallel in order to drive differentiation of iPSCs into possible therapeutically relevant cell-types.

**Big data for characterization of brain tumors: A radiomics and radiogenomics-based approach**
Prateek Prasanna

Problem and Unmet Clinical Need: Two most challenging clinical problems in management of brain tumors are: 1) Estimating survival characteristics in newly diagnosed primary brain tumors and 2) Distinguishing radiation-induced treatment effects affecting 20-40% of all brain tumor patients, from tumor recurrence.

Clinical and translational impact: The total national cost towards patient care in 2010 for over 250,000 brain tumor patients was $4.47 billion and is projected to reach at least $6.2 billion in 2020. Identifying markers that are predictors of patient outcome will set the stage for future trials to further validate and potentially combine them with other prognostic markers such as age, stage, extent of resection, pathological, and molecular information to guide therapeutic decisions and patient management. Evaluating treatment response early can impact 75,000 brain tumor patients in the U.S. every year who may have tumor recurrence but are subjected to a wait and watch, in the absence of sufficient diagnostic information. Similarly, ability to reliably identify radiation necrosis in brain tumors will substantially reduce over 18,000 unnecessary biopsies every year (including 5,000 GBM) thereby mitigating healthcare costs by $720 million.

While interrogated in the context of brain tumors, the proposed work will also find significant implications in identifying imaging markers of infiltration across other aggressive cancers (breast, lung), as well as across metastatic, and pediatric brain tumors.

**Feverprints: Crowdsourcing temperatures in health and disease**
Jonathan Hausmann

Our study seeks to leverage modern technology, including continuous temperature monitoring and crowdsourcing, to collect temperatures from thousands of participants throughout the country to reassess what is “normal” and what temperature constitutes a “fever.” Second, we will use machine learning to discover unique fever patterns (“feverprints”) for febrile illnesses that can be used for rapid and accurate diagnosis. Finally, we aim to show whether the use of antipyretics improves or worsens the course of febrile illnesses.

**Incorporating the Patient’s Voice into Cancer Care and Research**
Charlotta Lindvall

Electronic Health Records (EHR) contain enormous amounts of data that may be used to facilitate cancer treatment discovery, guide quality and safety initiatives, and enhance patient satisfaction. Currently, data from clinical visits (i.e., the conversation between clinician and patient) are interpreted and compressed by clinicians who write it directly into the medical record. This process misses much that transpires in the clinical encounter, risks error, and
results in lost opportunities to deeply understand how our care impacts outcomes. We propose to use advanced computer science methods to embed audio-recordings of clinical encounters into the EHR, employ machine-learning analytics to harness this rich data source, ensure that what actually transpires between patients and clinicians in the exam room is included in outcome analyses, and to thus transform clinical care by capturing, measuring, and accounting for the full breadth of patient experience.

**Automatic classification of clear cell renal tumors using deep learning: implications for diagnosis and prediction of metastasis**

Jan Heng

*The abstract for this prize has not been published at the request of the award winner*

**Computational approaches to assessing breast asymmetry for early detection of breast cancer**

Rulla Tamimi

Wide-spread mammographic screening as well as improvements in breast cancer treatment have greatly improved 5-year survival rates for breast cancer. However, the sensitivity of mammography is still not ideal and is lower in younger women and women with dense breasts. It may be possible to use big data and computational solutions to improve the sensitivity of mammographic screening and improve early detection of breast cancer. We and others have observed that the patterns of breast density are very similar in comparing the right and left breast from the same woman. However, when women are diagnosed with breast cancer it is primarily in only one breast. We hypothesize that there are early divergences in breast density patterns between the breasts that are detectable through imaging that could be early predictors of breast cancer. If there were automated ways to compare asymmetry in breast tissue patterns (i.e., spatial variation in breast density and features) between breasts over time, we may be able to detect breast cancers even earlier. Additionally, comparing breasts may help to control for differences that influence patterns at different time points such as technical issues (e.g., compression, mammographic machine, positioning on machine), and individual differences over time (e.g., changes in weight, hormone use, menopausal status).

**Mental Health Related Visits in Pediatric Emergency Departments Over Time: Differences in age groups, socioeconomic status, and race/ethnicity**

Anna Abrams

*The abstract for this prize has not been published at the request of the award winner*