**Question 1**

**Describe your research to a non-expert \***

**Please limit the description of your work to one paragraph (4-8 sentences). If possible, include a description of a key result, accomplishment, or breakthrough you made that is representative of your work. In addition to your description, feel free to include links about your research, which may include URLs to your academic website, your research papers, videos of talks you’ve given, your Github repository, or press coverage of your research.**

My research focuses on improving the treatment of bladder and colorectal cancer by aiding the detection of flat or “invisible” lesions, which often are overlooked using current surgical methods. I am developing a new technology that uses nanoparticles to bind to cancer proteins, and these nanoparticles give off an optical signal that can be used to identify cancer tissue for removal. During surgery, doctors would detect the signal using an endoscope by employing the signal processing algorithms that I have optimized. My technology would allow doctors to surgically remove tumor tissue more accurately and thoroughly while protecting healthy tissue, which should prolong the patient’s survival and improve their quality of life. This project is an extension of my previous work developing novel cancer imaging technologies, for which I have received *Nature* news coverage and a young investigator award from the *International Journal of Hyperthermia.*

Coverage in *Nature* News: <https://www.nature.com/nphys/journal/v8/n11/full/nphys2462.html>

Academic website, which includes publication list: <https://profiles.stanford.edu/70801>

Matlab repository for this project: <https://github.com/ryanmdavis/MolecularEndoscopy>

**Question 2**

**What programming / scripting / statistical languages, data analysis software and/or databases have you worked with and in what capacity? Please make a detailed list. For each language, database or software, list how much time you spent working with it and briefly describe the work you did and what results you obtained to give us a sense of your proficiency.**

1. Matlab (10-20 hrs/week from 2007-present)
   1. Wrote comprehensive image-processing software that converted raw MRI data (see 2b below) into images of bone temperature (2011-2015). Showed for the first time that MRI can measure the temperature of bone marrow, published two peer-reviewed papers.
      1. <https://github.com/ryanmdavis/GE-MSE-HOT>
      2. <https://github.com/ryanmdavis/MSE-HOT-thermometry>
      3. <https://github.com/ryanmdavis/MRM---red-marrow-thermometry>
   2. Wrote image and signal processing software that displays images of protein expression in human tissue. The endoscope/software combination that I use can detect 4 cancer-associated proteins on tissue at the same time, compared to previous technologies which can only detect one at a time. This approach should enable accurate localization of cancer, and my software is used by 3 other lab members. <https://github.com/ryanmdavis/MolecularEndoscopy>
   3. Wrote code to interface fiberoptic thermometer to a serial port, generated timed measurements of temperature. <https://github.com/ryanmdavis/LuxtronControl>
   4. Wrote object-oriented software that controlled a research MRI machine.
2. C Language (10-20 hrs/week from 2011-2015)
   1. Design, programmed, built a fully-functional quad-copter without using kit. Programmed ATmega microcontrollers using C language (2013-2015) link: <https://github.com/ryanmdavis/QuadCopter>
   2. For PhD thesis, designed and implemented MRI-control software that acquired temperature images of human bone marrow (2012-2015). (C code is propriary, see 1.a above for explanation of project).
3. Python (~150 hours, 2014-present)
   1. Wrote code to communicate via Bluetooth to a quadcopter
   2. Wrote an application called WhosWhoBioTech which identifies key players in the biotechnology industry. <https://github.com/ryanmdavis/WhosWhoBioTech>. More below on question 3.
4. R (50 hours, Sept 2015 – March 2016)
   1. Used R packages to query The Cancer Genome Atlas, performed variable (gene) selection using sparse discriminant analysis to select genes predictive of cancer. I identified small panels of genes that are predictive of cancer.

**Question 3**

**Please describe any side projects relevant to data science you have completed (unrelated to coursework) - feel free to include links. [Optional]**

**Overview:** As Vice President of a student-run consulting organization that focuses on the life science industry, I often need to quickly identify the prominent companies and thought leaders in a given field. Obtaining this information by hand involves a time-consuming process of combing through industry reports and business articles. I overcame this challenge by writing an object oriented algorithm in Python called “WhosWhoBioTech”. This program takes a user query such as “antibiotics” and outputs a list of the most important companies and individuals relevant to that term.

**Methods:** WhosWhoBioTech uses a corpus of biotechnology business articles downloaded from FierceBiotech.com using import.io. The algorithm operates in three steps. First, the user submits a query, and a tf-idf approach is used to retrieve the articles most relevant to the user’s query. Second, keywords are extracted from the top ten articles. WhosWhoBioTech can use either Textrank or tf-idf-based keyword extraction algorithms. Finally, keywords that correspond to named entities (organizations or people) are identified using nltk’s named entity recognition functions and returned to the user. The algorithm output provides the most important named entities in the corpus related to the user query.

Algorithm performance was evaluated in two separate ways. First, the keyword extraction performance was evaluated by calculating the precision, recall, and f-measure of keyword extraction as a function of number of keywords returned by the algorithm. To calculate these metrics, extracted keyword lists were compared to gold-standard keyword lists provided on the FierceBiotech.com website. Second, the prominence and business activity of individuals and companies returned by the algorithm was assessed qualitatively to gauge their importance in the field (see below).

**Results and example:** For the first stage of evaluation, the keyword extraction performance was relatively low (recall = 20% when 50% of potential keywords in document are returned). One major reason for this low recall was that many of the gold-standard keywords were words or phrases not contained in the main text body. Also, many gold-standard keywords were not people or organizations, and were thus not relevant to the goal of identifying important companies and individuals related to a search query. For these reasons, in spite of the low recall, the tf-idf and TextRank algorithms were deemed sufficient for keyword extraction.

WhosWhoBioTech successfully identified important companies and individuals related to test queries. For example, querying the algorithm for “antibiotic” returned results including the Antimicrobial Resistance Center (ARMC, a private-public initiative in the UK to support/accelerate the development of new antibiotics), Motif Biosciences (a clinical stage biopharmaceutical company specializing in developing novel antibiotics), and Roger Pomerantz (the CEO of Seres Therapeutics, which is making a microbiome therapeutic intended to replace antibiotics for treatment of *Clostridium difficile* infection). Thus, with a simple query, WhosWhoBioTech allowed me to discover some of the important individuals and organizations in the field of antibiotics. To try it yourself, download the content of <https://github.com/ryanmdavis/WhosWhoBioTech>, and run the file wwbt\_example.py.

**Question 4**

**If you have done any coursework relevant to data science please let us know which courses and any projects etc. done during the course. [Optional]**

**Please include a short description of the courses and any in-course related projects that you may have completed.**

During graduate school I’ve taken courses that use many principles of data science including:

Course using optimization –

Advanced photon beam radiation therapy: Used gradient descent to optimize radiation dose distribution to tumors using intensity-modulated radiation therapy.

Course using Linear Algebra –

Statistical Thermodynamics: Simulated the dynamics of a nuclear magnetic resonance spin system using the von Neumann Equation. This formulism describes the magnetic state of an ensemble of organic molecules as a matrix.

Course using Signal Processing -

Advanced medical imaging systems: Used fast Fourier Transforms (FFT) and filtering to reconstruct raw magnetic resonance imaging (MRI) data.

**Question 5**

**What excites you about becoming a data scientist and working in the healthcare and life sciences industries? \***

I believe that my PhD-level engineering training combined with my consulting experience strongly positions me to be an effective data scientist in the life science and healthcare industries. With both technical and business experience, I would be able to solve applied business problems by using [NLP?] – for example, when pharmaceutical companies need to mine electronic health records to identify eligible patients for their clinical trials.

I am excited about a career in health data science because it will allow me to apply my love of data analysis to important problems in the life science and healthcare industries. My experience in scientific research is diverse: I have designed contrast agents to target tumor proteins, quantified the pharmacodynamics of contrast agents being metabolized by the body, and used quantum mechanics to improve MRI images of bone cancer. Even though these projects cover a range of subjects, my favorite part of each was writing customized signal processing algorithms to extract useful insights from raw data (see Github), and I see a career in health data science as a way to focus on the quantitative aspects of my work that I enjoy the most. Furthermore, my consulting experience has reinforced that my love of data analysis can be applied to important business problems in the healthcare industry. For example, companies developing novel drugs need to identify patients for clinical trials, which can be accomplished by systematically mining large numbers of electronic health records. I am excited to apply my experience with NLP and my knowledge of the life science industry to applied problems such as this, with the goal of promoting patient health and driving new drugs and medical technologies into the marketplace. Looking forward, I foresee significant changes in how the healthcare industry operates based on the rapidly increasing amount of data that is being generated per individual. Companies who are able to harness this vast amount of data will outcompete those who are not, and the Insight Health Data Science fellowship would help equip me with the skills needed to be at the forefront of this exciting paradigm shift.

**OLD**

I believe that my PhD-level engineering training combined with my consulting experience strongly positions me to be an effective data scientist in the life science and healthcare industries. My experience in consulting has highlighted issues that can be answered by data science in order to answer crucial business questions. For example, companies developing novel drugs need to identify patients for clinical trials, which can be accomplished by systematically mining large numbers of electronic health records. I am excited to apply my experience with NLP and my knowledge of the life science industry to applied problems such as this, with the goal of promoting patient health and driving new drugs and medical technologies into the marketplace.

I am excited about a career in data science not only because I will get to work on exciting problems in the life science industry, but also because I love working with data. My experience in scientific research is diverse: I have designed contrast agents to target tumor proteins, quantified the pharmacodynamics of contrast agents being metabolized by the body, and used quantum mechanics to improve MRI images of bone cancer. Even though these projects cover a diverse set of subjects, my favorite part of each was writing customized signal processing algorithms to extract useful insights from raw data (see the MolecularEndoscopy, MSE-HOT-thermometry, and MRM---red-marrow-thermometry repositories on my GitHub account). I have therefore seen first hand that I can apply my technical and coding skills to drive progress in scientific research, and with the help of Insight I want to apply this skillset to important problems in the healthcare and life science industry. I’m enthusiastic about a career in data science because I am confident in my ability to drive progress with data analytics, and because I will be able to focus my efforts on the mathematics and programming aspects of my work that I love most.

I foresee significant changes in how the healthcare industry operates based on the rapidly increasing amount of data that is being generated per individual. Companies who are able to harness this vast amount of data will outcompete those who don’t, and I want to equip myself with the skills to be at the forefront of this exciting paradigm shift. Thus, I am confident that the combination of my biomedical research experience and the training provided by the Insight Health Data Science fellowship will allow me to harness the power of data for the benefit of companies and patients alike.