Statement of Purpose – Erin McMahon

Application for Fall 2014 Matriculation

I am applying for a Masters in Computer Science at Columbia University because Columbia's Institute for Data Sciences and Engineering (IDSE) promotes a holistic outlook on the emerging data sciences field. Like IDSE, I am holistic in my view of data science, triaging my experiences into two facets of data science: exploration and narrative.

Many Masters programs limit their definition of data sciences to exploration, a.k.a. learning advanced machine learning and statistical methods with little to no context. I like IDSE because it does not. With the assimilation of researchers from over thirty fields in the heart of New York, IDSE is the ideal setting to flesh out a narrative by involving non-quantitative specialists in how we both formulate and answer problems.

From internships at NASA and the Joint US-China Collaboration on Clean Energy, research in Pharmacology studies, and quantitative analysis at Financial Services, I've seen the gamut of data processes with varying levels of sophistication. I have interpolated flow patterns, spatially correlated population density and land use, developed regression models for loyalty behaviors, and forecasted transaction volume. As a Senior Quantitative Analyst at an international consulting firm, Corporate Executive Board (CEB), I use this rich backdrop to write the simplest models that are complex enough to be accurate. In every experience, I've pushed myself to incorporate new statistical methods and write stories to explain the findings.

NARRATIVE 1: *Spatial Exploration Re-Mapped*

Given my background in environmental science, I successfully integrated spatial modeling techniques into my current work in financial services. Through Columbia's Lamont-Doherty Institute, I worked with scientists to model arsenic concentrations at over eighty well sites at an EPA Superfund site using geo-statistical functions such as spline and kriging. From my research, scientists were able to assess the efficacy of their current methodology, gain insight into arsenic interactions at differing groundwater depths and soil composition, and forecast the time and budget needed for further remediation. As I moved to the financial services sector after college, I repurposed these skills at Corporate Executive Board. Through a study to evaluate US bank performance, I mapped over- and under-performing FDIC branches across the United States. Based on the distribution of age, income, education, and home prices within census tracts, I interpolated customer segments, predicted purchase likelihood, and compared potential deposits against real deposits. The output was zip code-level maps that easily explained findings useful for our client base.

NARRATIVE 2: *Building the Story with Data*

At SNBL, a clinical pharmacology firm, I helped launch a research study on anthrax packages in the case of bioterrorist attack. As an integrated member of the team, I worked on everything from writing the grant application to designing the research protocols to planning the survey analysis. I applied this same approach to all my projects at CEB, embedding myself into the qualitative teams. I helped our Wealth Management team whittle down its study to one basic question: what do customers most value in financial advisors? Rather than use a traditional approach, I employed Item Response Theory (IRT), a survey technique used at SNBL, to build a more robust scale of loyalty. Because the index was more exact, we were able to determine "killers" and "builders" in the customer experience, i.e. those that affected attrition versus those that solidified the relationship, respectively. By eagerly learning the full dimensions of the study, I chipped away at the team's skepticism, at the reluctance to stray from a tried-and-true method, demonstrating that IRT fit the function.

The data scientist must take on the role of scientist, economist, journalist, and so forth to model the relationship at a granular level and then aggregate. Donning this functional 'hat' is vital to the research. In both work and school, I have taken on various functional areas.

From these experiences, I quickly learned that design carries as much weight as theory. Independent of the field, all of my projects have required a narrative of the data findings. As analysis becomes more complex, the stories must mirror this complexity while still conveying a clear message. In my environmental research, I achieved this through maps. Environmental science is dependent on place: it describes physical, chemical, and biological interactions that are united only by location and time. Thus, the best visualization is a map, preferably an interactive one. My research internships and senior thesis utilized Geographic Information Systems (GIS) for this reason. While this concept was nebulous for me as an undergrad, I sensed that the form of visualization impacted how my research was consumed. This became a tenet of my job at CEB. Sophisticated models lost meaning without concise, intuitive graphs. Findings were missed; actions stifled. At first, this was hard to digest: doesn't everyone know what a beta coefficient means? Now, I craft each data visualization as I would an essay. I cycle through various drafts, seek out feedback from qualitative teams, and edit, edit, edit. Some of these visualizations fail; many are good; and a few hit that delicate balance of 'creative but simple.' In particular, I created a channel orbit graph to illustrate how customers' activities revolve around specific channels such as website or branch and orbit from there. As proof of my effort and dedication, one of my graphics has been displayed at the firm-wide competition each year since I started at CEB.

These goals – exploration and narrative – are interdependent. From its inception, mathematics evolved with problems arising from human complexity: fractions were invented for trade, matrices for astronomy, calculus from gravity. The "how" and the "why" have always been linked. Even now, data are reinventing mathematics, statistics, and computer science, illustrating how function can influence process. Especially – but not exclusively – through the medium of technology, new forms of data exploration and visualization have the power to establish a narrative by reorienting how people understand and interact with data.

With latent data available in unstructured social media, audio, and video multimedia, there is an untapped opportunity to extract and parse data in creative ways. These new resources raise a number of questions: i) what is the best methodology to analyze the data and can it be improved? ii) what story do the data tell us? iii) what visualization tools will tell this story most clearly? With Columbia's wide array of top-notch academic fields from economics to design, IDSE will foster symbiotic relationships among fields. I hope to use this ecosystem of resources to explore these unstructured data by adapting methodologies from other fields, broadening the uses and impact of these resources, and designing new visualization tools for how we play with these data.

Columbia University's Masters in Data Science is progressive in approach, balanced in its weight on application and theory, and significant in its ability to recast the way we, as a society, solve problems. My academic and industry experience prompts me to value these things. My future research into the methodologies, uses, and tools pertaining to unstructured data will depend on these things. Whether by garnering a deeper understanding of statistics, translating methodologies from other functional areas, or illustrating results in a cogent manner, I strive to fit exploration into narrative and vice versa, a goal I hope to continue through a Masters in Computer Science at Columbia University.