TMCity

Senior Data Engineer
Job Description

Neuroinformatics Project

Phase I: Review of current state of data sharing in the field of brain health with a focus on digital data sets.

Background

TMCity Foundation supports efforts that will advance our understanding of the brain, with a view to accelerating the discovery of treatments and cures for brain-related disorders. We believe one key to unlocking this knowledge is to be found in data collection, sharing, and analysis. We want to apply machine learning to potentially validate digital biomarkers and digital phenotyping for neurological conditions, and thus find more effective solutions that can be applied early and easily. Our goal is to use digital data to make significant progress in how we care for people's brain and mental health. This work will include evaluating the validity of creating a digital health data repository and sharing platform for brain health.

Some of the questions we are asking include:

- How to ensure the optimal structure, quality, security, and compatibility or standardization of data from varied sources.
- How to achieve and promote interoperability across data platforms, including access and sharing restrictions or requirements.
- How to address concerns related to privacy, security, consent, de-identification/anonymization and data ownership.
- How to incentivize data sharing.
- How to use both retrospective and prospective data sets.

In order to determine how best to proceed with this work, the Foundation seeks to hire a data scientist to undertake a review of the current state of data sharing in the field of brain health with a focus on digital data sets, and to propose a framework for the Foundation's ongoing engagement in the field.

Objectives

- To understand the current state of data formatting, sharing, and analysis in the brain health space, including a review of current data repositories related to Alzheimer's Disease and other neurodegenerative conditions (Parkinson's, etc.)
- To identify current relevant data sets and their availability/accessibility, including analysis of formats, requirements for access and use, and current state of interoperability across platforms.

- To test the feasibility of building an algorithm based on these data sets related to the validation of digital biomarkers for brain health.
- To identify gaps/opportunities in the field in general that will inform the Foundation’s data initiatives and recommend next steps

Required Qualifications/Expertise

- Strong problem-solving skills with an emphasis on product development.
- Experience using statistical computer languages (R, Python, SQL, etc.) to manipulate data and draw insights from large data sets.
- Experience working with and creating data architectures.
- Knowledge of a variety of machine learning techniques (clustering, decision tree learning, artificial neural networks, etc.) and their real-world advantages/drawbacks.
- Knowledge of advanced statistical techniques and concepts (regression, properties of distributions, statistical tests and proper usage, etc.) and experience with applications.
- Excellent written and verbal communication skills for coordinating across teams.
- A drive to learn and master new technologies and techniques.

We’re looking for someone with 5+ years of experience manipulating data sets and building statistical models, with a graduate degree in Statistics, Mathematics, Computer Science or another quantitative field, and who is familiar with the following software/tools:

- Coding knowledge and experience with several languages: C, C++, Java, JavaScript, etc.
- Knowledge and experience in statistical and data mining techniques: GLM/Regression, Random Forest, Boosting, Trees, text mining, social network analysis, etc.
- Experience querying databases and using statistical computer languages: R, Python, SQL, etc.
- Experience using web services: Redshift, S3, Spark, DigitalOcean, etc.
- Experience creating and using advanced machine learning algorithms and statistics: regression, simulation, scenario analysis, modeling, clustering, decision trees, neural networks, etc.
- Experience analyzing data from 3rd party providers: Google Analytics, Site Catalyst, Coremetrics, Adwords, Crimson Hexagon, Facebook Insights, etc.
- Experience with distributed data/computing tools: Map/Reduce, Hadoop, Hive, Spark, Gurobi, MySQL, etc.
- Experience visualizing/presenting data for stakeholders using: Periscope, Business Objects, D3, ggplot, etc.

Please send your resume and a role-specific cover letter to Belen Paley Cox at bcox@tmcgroup.com. Applications will be considered on a rolling basis until the position is filled.