# A data-driven ontology of brain function

## Engineered, interrogated, and clinically applied

**Ellie Beam**
MD/PHD Candidate
PhD in the Neurosciences
Stanford University

In modern healthcare, there is no accepted system for diagnosing disorders of mental function based on the altered brain systems that cause them. In neuroscience, there is no consensus on what a brain system *is.* This thesis takes a data-driven approach to mapping domains of brain function across the human neuroimaging literature. Neural circuits and associated mental functions were mapped from the brain coordinate data and full texts of nearly 20,000 studies. The resulting domains characterize several novel brain circuits that are absent from the conceptually dominant expert-determined frameworks in the field.

Having engineered a framework for brain function, the next aim was to interrogate whether its domains adhere to basic organizational principles. Are the domains *reproducible*, meaning that their circuits and functions predict one another in held-out articles? Are they *modular*, partitioning the literature into homogeneous and separable subsets? Are they *generalizable*, serving as prototypes of circuits and functions observed in single studies? It is demonstrated here how competing knowledge frameworks can be evaluated against a common set of standards. The data-driven framework is consistently seen to offer leading performance.

The second aim of the thesis was to apply the data-driven framework to predict clinical outcomes. The notes from over 20 million healthcare visits were rated by semantic similarity to domains, and these dimensional phenotypes were found to outperform binary diagnoses in predicting future psychotropic prescriptions, hospital visits, and all-cause mortality. Though the data-driven framework was not designed to have clinical relevance, its prognostic performance is on par with frameworks crafted specifically for psychiatric applications.

Computational approaches to ontology engineering are expected to mitigate certain forms of bias, while others may be inherent to the neuroimaging literature. The third and final aim of the thesis was to systematically quantify discrepancies between article texts and brain coordinate data. The results point to the influence of the publication incentive structure on how neuroimaging results are interpreted by neuroscientists.

The collective results illustrate the scientific, clinical, meta-scientific interest in engineering a data-driven framework for brain function. They suggest a potential for biological knowledge frameworks to go beyond guiding diagnosis and to generate the prognoses that have long eluded the standard of care for disorders of brain function.