ACTIVE DISCOVERY IN BIG DATA
With the explosive growth of the Internet and availability of increasingly cheap large-scale storage, amassing huge datasets is becoming commonplace. In many real-world scenarios, however, conducting detailed analysis of an identified data point is very expensive, requiring human intervention or a costly experiment. For example, in intelligence analysis or fraud detection, collecting and storing social and/or financial transactions of an agent is trivial, but extracting intelligence value or conducting an in-depth investigation can require hours of a human expert's time. In such situations, it is critical that we allocate limited resources effectively.

"Active" machine learning considers how to selectively choose observation locations so as to accurately predict the outcomes of an experiment on the remaining points. Oftentimes, however, our ability to make accurate predictions is only of secondary importance. We will describe effective and theoretically motivated active learning policies for several important problems with radically different objectives, including discovering points of interest, conducting targeted opinion polls, and learning low-dimensional embeddings of expensive, high-dimensional functions. These policies offer effective solutions with numerous diverse applications, including intelligence analysis, drug discovery, black-box optimization, and automated model selection.

Dr. Roman Garnett is a postdoctoral researcher in the Knowledge Discovery and Machine Learning group at the University of Bonn. Previously, he was a postdoctoral fellow in the Robotics Institute at Carnegie Mellon University and an applied research mathematician at the National Security Agency (NSA). He received a Ph.D. in Machine Learning from the University of Oxford in 2010. His research interests include active learning (especially with atypical objectives), Bayesian optimization, and Bayesian nonparametric analysis.