Abstract:
An unprecedented volume of biomedical evidence is being published today. Indeed, PubMed (roughly, a search engine for biomedical literature) now indexes more than 600,000 publications describing human clinical trials and upwards of 22 million articles in total. This volume of literature imposes a substantial burden on practitioners of Evidence-Based Medicine (EBM), which now informs all levels of healthcare. Systematic reviews are the cornerstone of EBM. They address a well-formulated clinical question by synthesizing the entirety of the published relevant evidence. To realize this aim, researchers must painstakingly identify the few tens of relevant articles among the hundreds of thousands of published clinical trials. Further exacerbating the situation, the cost of overlooking relevant articles is high: it is imperative that all relevant evidence is included in a synthesis, else the validity of the review is compromised. As reviews have become more complex and the literature base has exploded in volume, the evidence identification step has consumed an increasingly unsustainable amount of time. It is not uncommon for researchers to read tens of thousands of abstracts for a single review. If we are to realistically realize the promise of EBM (i.e., inform patient care with the best available evidence), we must develop computational methods to optimize the systematic review process.

To this end, I will present novel data mining and machine learning methods that look to semi-automate the process of relevant literature discovery for EBM. These methods address the thorny properties inherent to the systematic review scenario (and indeed, to many tasks in health informatics). Specifically, these include: class imbalance and asymmetric costs; expensive and highly skilled domain experts with limited time resources; and multiple annotators of varying skill and price. In this talk I will address these issues in turn. In particular, I will present new perspectives on class imbalance, novel methods for exploiting dual supervision (i.e., labels on both instances and features), and new active learning techniques that address issues inherent to real-world applications (e.g., exploiting multiple experts in tandem). I will present results that demonstrate that these methods can reduce by half the workload involved in identifying relevant literature for systematic reviews, without sacrificing comprehensiveness. Finally, I will conclude by highlighting emerging and future work on automating next steps in the systematic review pipeline.

Bio:
Byron Wallace is an assistant research professor in the Department of Health Services, Policy & Practice at Brown University; he is also affiliated with the Brown Laboratory for Linguistic Processing (BLLIP) in the department of Computer Science. His research is in machine learning/data mining and natural language processing, with an emphasis on applications in health. Before moving to Brown, he completed his PhD in Computer Science at Tufts under the supervision of Carla Brodley. He was selected as the runner-up for the 2013 ACM SIGKDD Doctoral Dissertation Award and he was awarded the Tufts Outstanding Graduate Researcher at the Doctoral Level award in 2012 for his thesis work.