

# Approximate String Joins

**Divesh Srivastava**  
AT&T Labs-Research  
divesh@research.att.com

## Abstract

String data is ubiquitous and is commonly used to correlate (or join) entities across autonomous, heterogeneous databases. The main challenge is to effectively deal with the noisy nature of string data, due to, for example, transcription errors, incomplete information, and multiple conventions for recording string valued attributes. Commercial databases do not support approximate string joins directly, and it is a challenge to implement this functionality efficiently. In this talk, I'll present techniques for performing approximate string joins, based on a variety of string similarity metrics, including variants of edit distance and cosine similarity. These techniques are scalable, and can be formulated to execute efficiently in a relational database management system.

This is joint work with Luis Gravano, Panagiotis G. Ipeirotis, H. V. Jagadish, Nick Koudas, and S. Muthukrishnan.