Nagy (adat)halmazok véletlen feldolgozása a Google-ben

Fogaras Dániel

A Google-ben nagy adatok feldolgozására és tárolására használt elosztott számítási környezet néhány építőkockájáról fogok mesélni: MapReduce-ról illetve BigTable-ről. Ezeket olyan algoritmusokkal illusztrálom majd, amelyek közös tulajdonsága, hogy nagy halmazokat tömör mintavételezéssel reprezentálnak. A különböző reprezentációk segítségével becsülhető a halmazok mérete (Probabilistic counting), átfedése (MinHash) vagy maga a halmazba tartozás (Bloom filter).

Az előadás a "Nagy adathalmazok kezelése" című tárgy és a SZIT tanszéki szemináriumum közös rendezvénye.