

CSE 550: *Systems for all*

Au 2021

Ratul Mahajan

“Big Data” problems

Lots of data

Semi-structured data

- Web logs
- Documents

Ad hoc computations

- But naturally parallelizable

Throughput-oriented (not latency-oriented)

Consistency not so important

- Or, writes are uncommon

Why not DBs?

Design space for data processing systems

	Latency	Throughput
Parallelizable (Data parallel)	Search, KV lookup (NoSQL)	Word counting (MapReduce)
Intertwined	Transactions (Traditional DBs)	Drug simulations (HPC systems)

Over to Prashant and Wentao