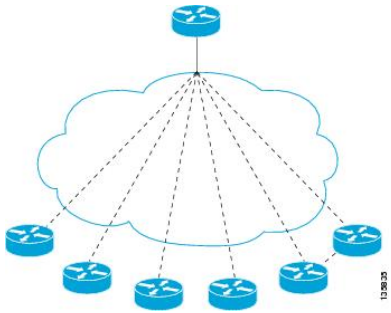

Qin Zhang
Communication-Efficient Distributed Computation
INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

Big data computation model: the coordinator model

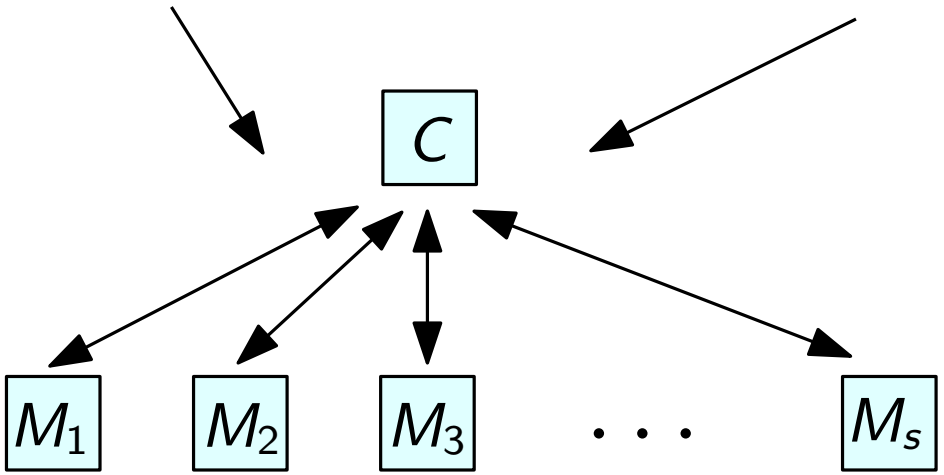
Big data is often collected/stored distributedly, while we want to compute on the global dataset

sensor networks



network routers

cloud computation

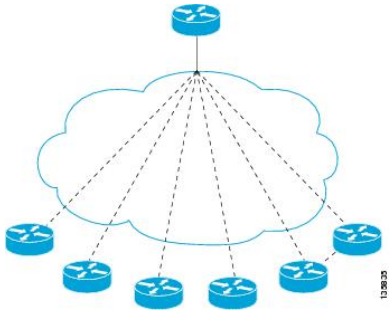


the coordinator model

Big data computation model: the coordinator model

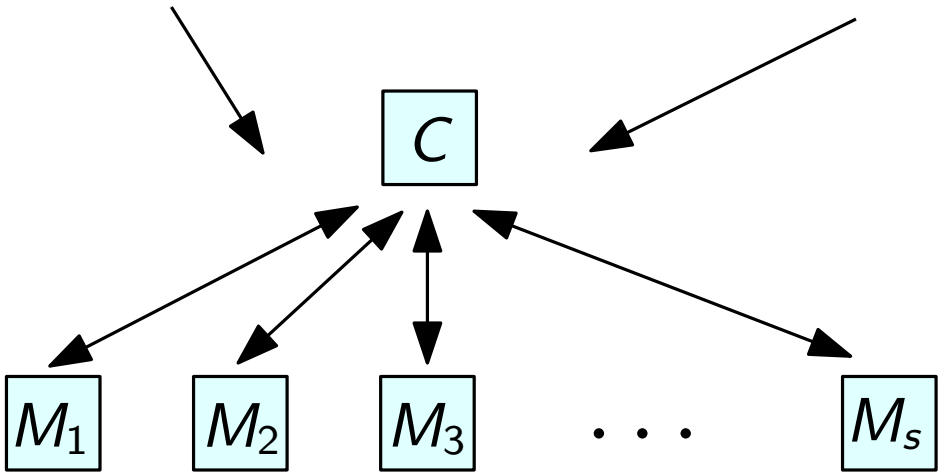
Big data is often collected/stored distributedly, while we want to compute on the global dataset

sensor networks



network routers

cloud computation



the coordinator model

communication $\rightarrow$ your data/energy/time bill

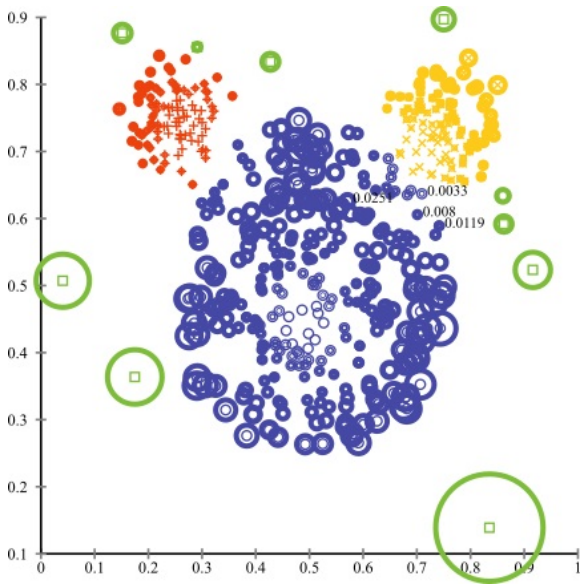
- small comm. cost
- small # comm. rounds

Fundamental problems

I study fundamental problems, both theoretically and practically, in *databases*, *data mining*, *machine learning* and *bioinformatics*.

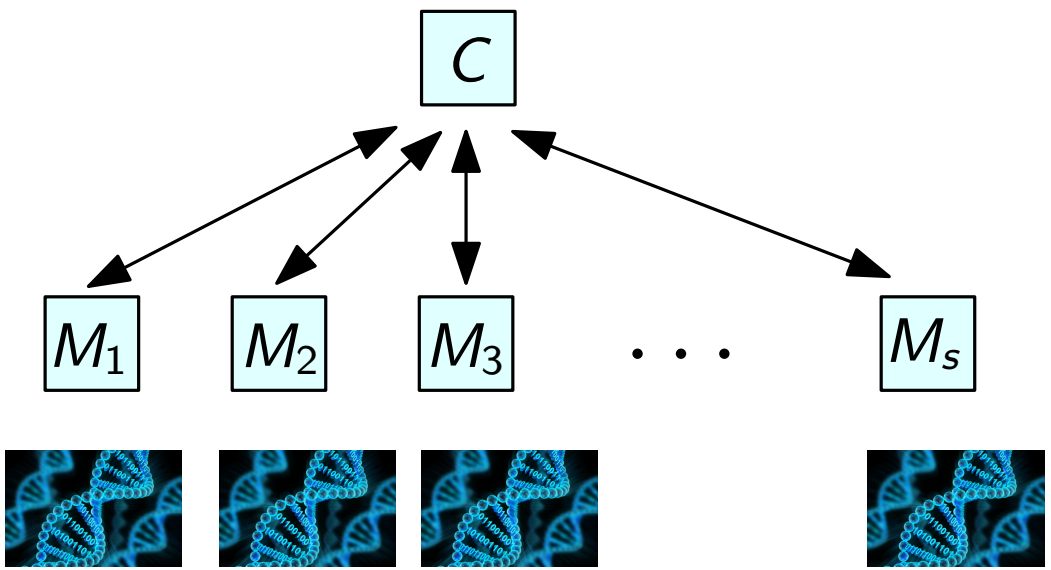
For example:

- string similarity joins
- clustering (with outliers)



String similarity joins

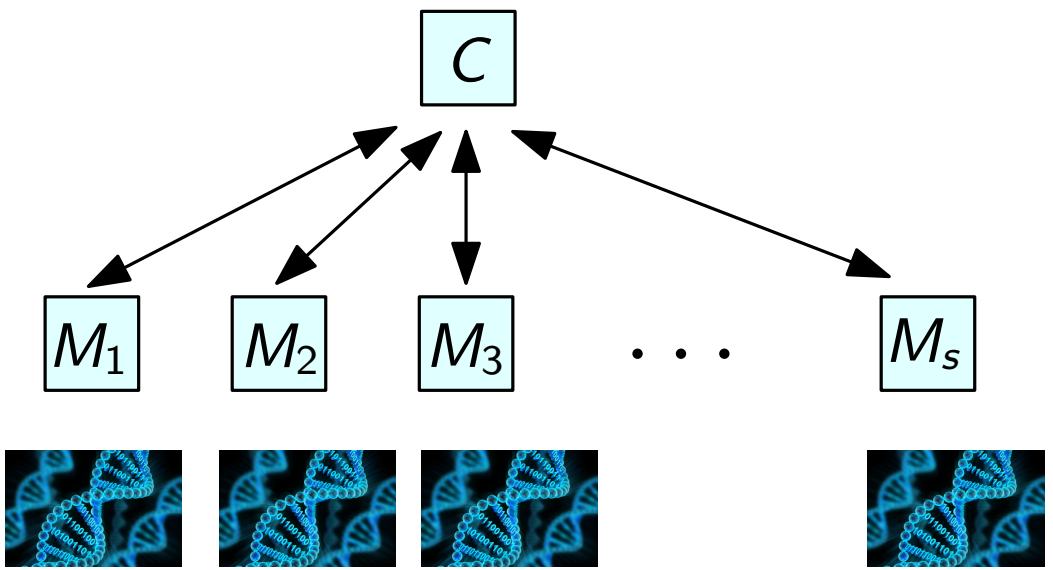
Find all pairs of strings (s_i, s_j) s.t. edit-distance$(s_i, s_j) \leq K$, where K is a threshold measuring similarity



- sketch-based approach; 1 round
- communication:
  sketch size $\times$ #strings

String similarity joins

Find all pairs of strings (s_i, s_j) s.t. edit-distance$(s_i, s_j) \leq K$, where K is a threshold measuring similarity



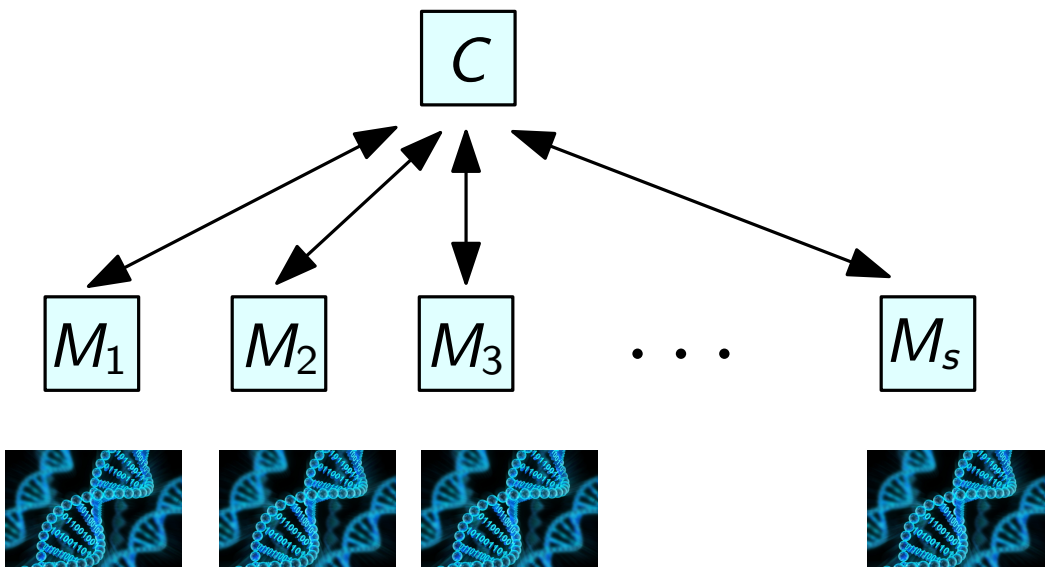
- with Belazzougui, *FOCS'16*
  - –First sketch of sublinear size: $O(K^8 \log^5 n)$ when $K \ll n$.

    Solves a major open problem in the area of sketching algos
  - – Make use of multiple random walk embeddings

- – sketch-based approach; 1 round
- – communication: sketch size $\times$ #strings

String similarity joins

Find all pairs of strings (s_i, s_j) s.t. edit-distance$(s_i, s_j) \leq K$, where K is a threshold measuring similarity

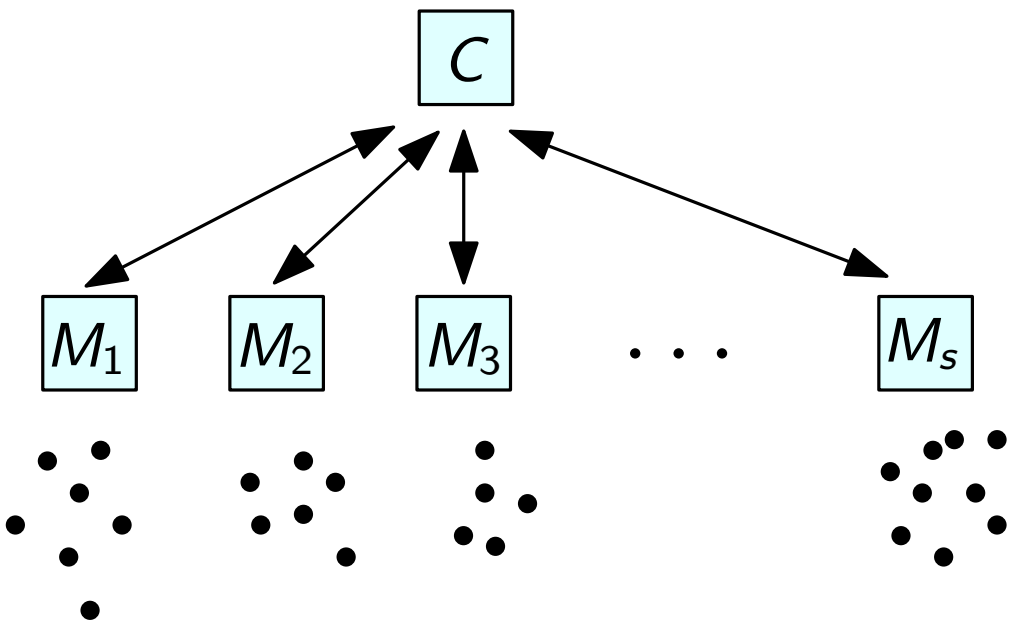


- sketch-based approach; 1 round
- communication: sketch size $\times$ #strings

- with Belazzougui, *FOCS'16*
  - First sketch of sublinear size: $O(K^8 \log^5 n)$ when $K \ll n$.
    Solves a major open problem in the area of sketching algos
  - Make use of multiple random walk embeddings
- with H. Zhang, *KDD'17*
  - A more practical algorithm
  - Run less than 10 hours on a 16-core server for 320,000 DNA strings each of size $n = 5000$, and $K = 1000$
  - Orders of magnitude faster than previous algos

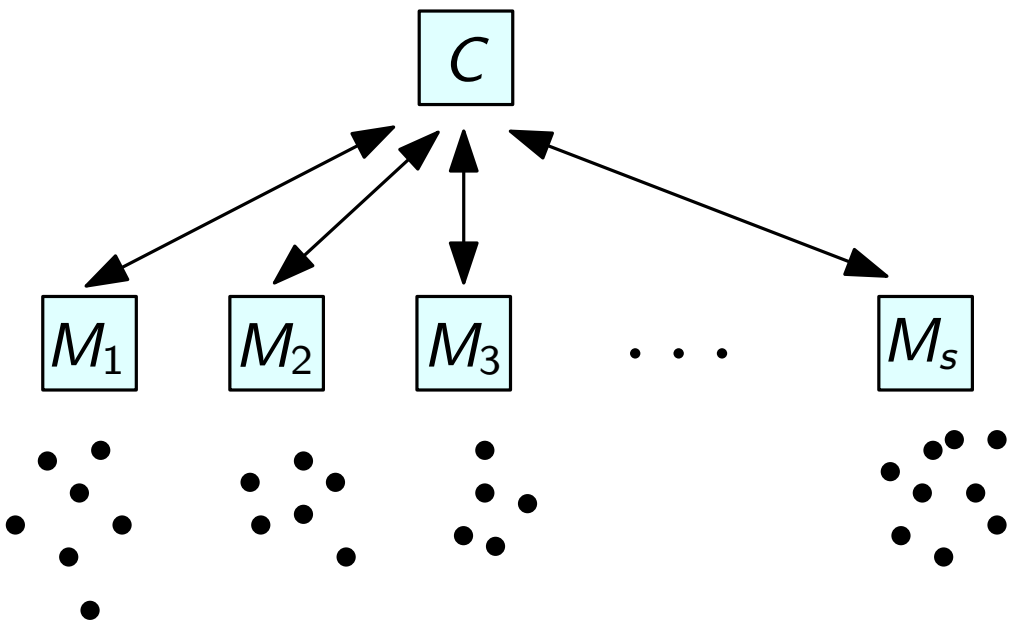
Clustering (with outliers)

Output k centers and t outliers,
under objective functions
k-means/median/center



Clustering (with outliers)

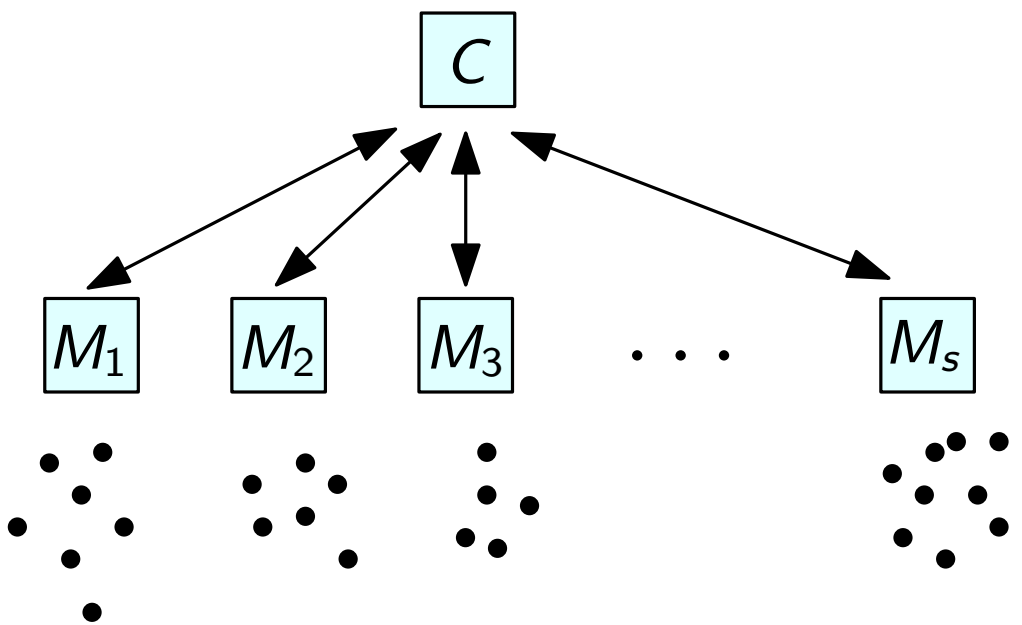
Output k centers and t outliers, under objective functions k-means/median/center



- with Guha and Li, *SPAA'17*
  - A set of algorithms for (k, t)-means/median/center with theoretically optimal commmunication $\tilde{O}(sk + t)$, and 2 rounds.
  - **Best paper award**

Clustering (with outliers)

Output k centers and t outliers, under objective functions k-means/median/center



- with Guha and Li, *SPAA'17*
  - A set of algorithms for (k, t)-means/median/center with theoretically optimal commmunication $\tilde{O}(sk + t)$, and 2 rounds.
  - **Best paper award**
- with Chen and Sadeqi Azer, under submission
  - Practical 1-round algorithm
  - Beat previous algos in all metrics (communication cost, accuracy, local running time)


THANK YOU
Qin Zhang's CONTACT INFO
EMAIL: qzhangcs@indiana.edu
http://homes.soic.indiana.edu/qzhangcs/
Projects funded by NSF
INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING