Introduction

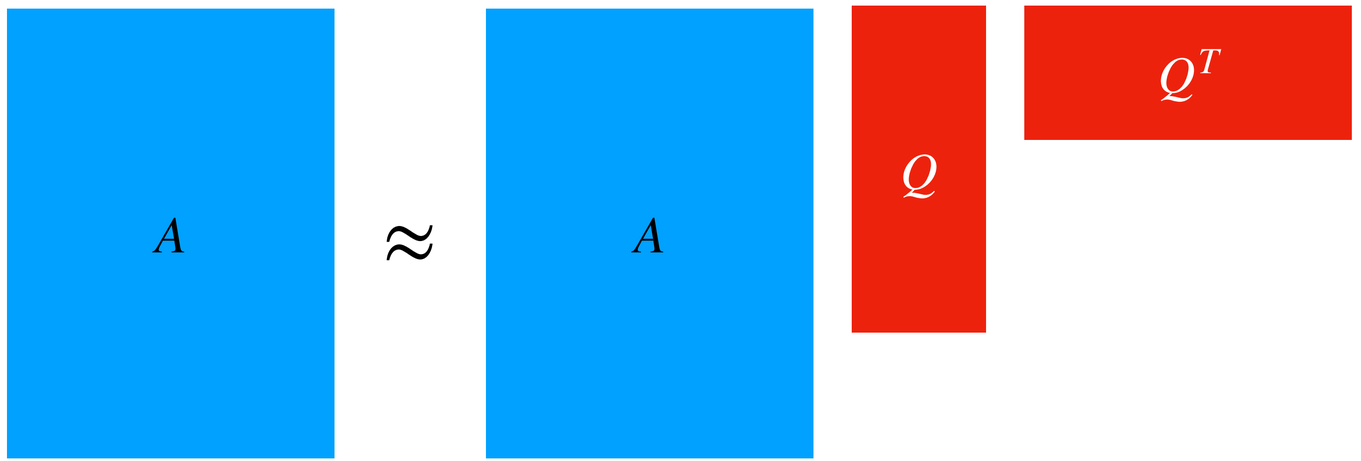
- Datasets these days are very large and presented to us in different ways
- Many existing algorithms assume arbitrary access to input data and run in super-linear time
- Can these algorithms reasonably handle large inputs which maybe presented in unconventional ways?

An Example Problem

Principal Component Analysis

- Given an $n \times d$ matrix A, compute a **good** rank k subspace Q

$$A \approx A \cdot Q \cdot Q^T$$



- Columns of Q are the **Principal Components**