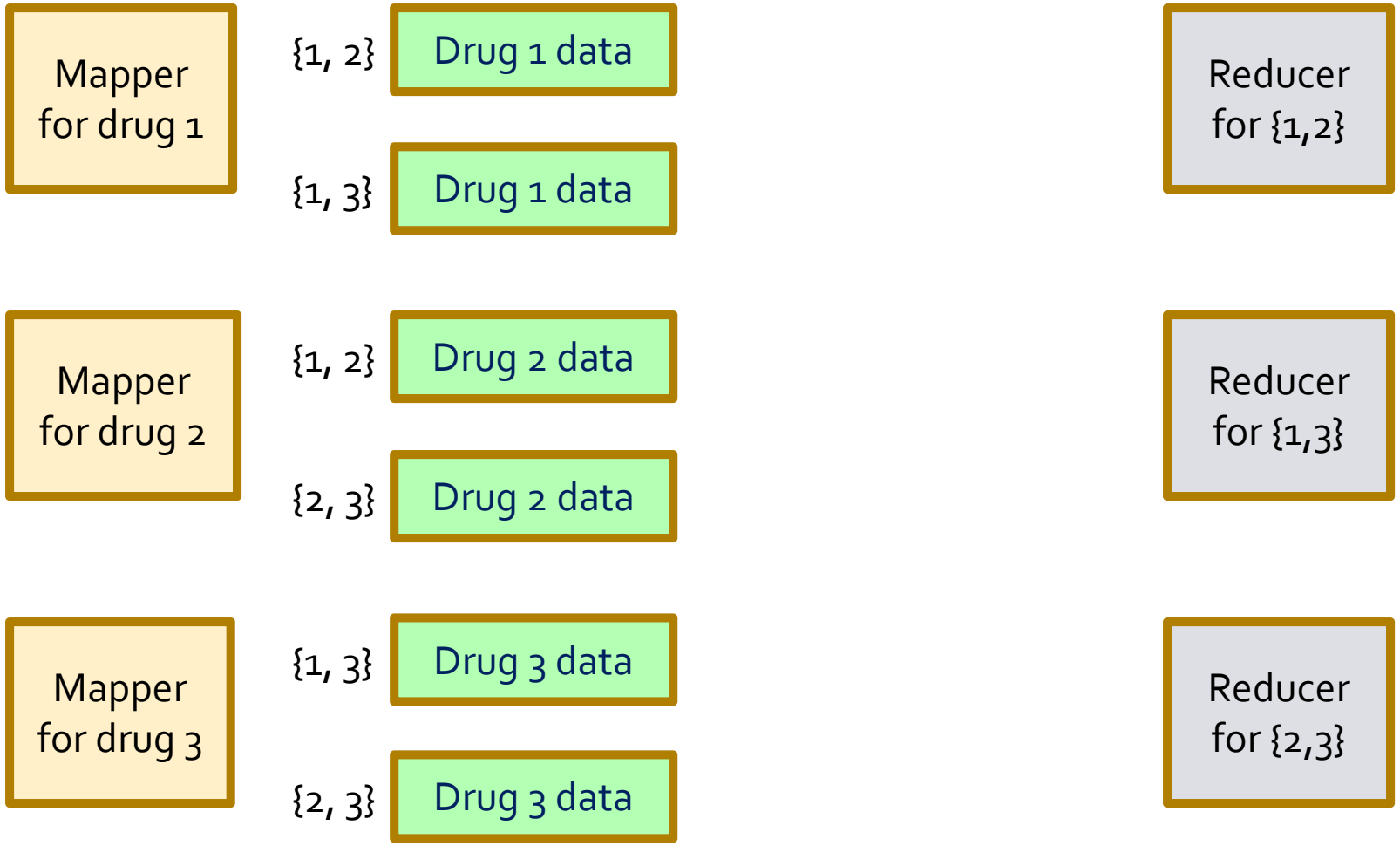
An Example Problem

- Data consists of records for 3000 drugs.
  - List of patients taking, dates, diagnoses.
  - About 1M of data per drug.
- Problem is to find drug interactions.
  - Example: two drugs that when taken together increase the risk of heart attack.
- Must examine each pair of drugs and compare their data.

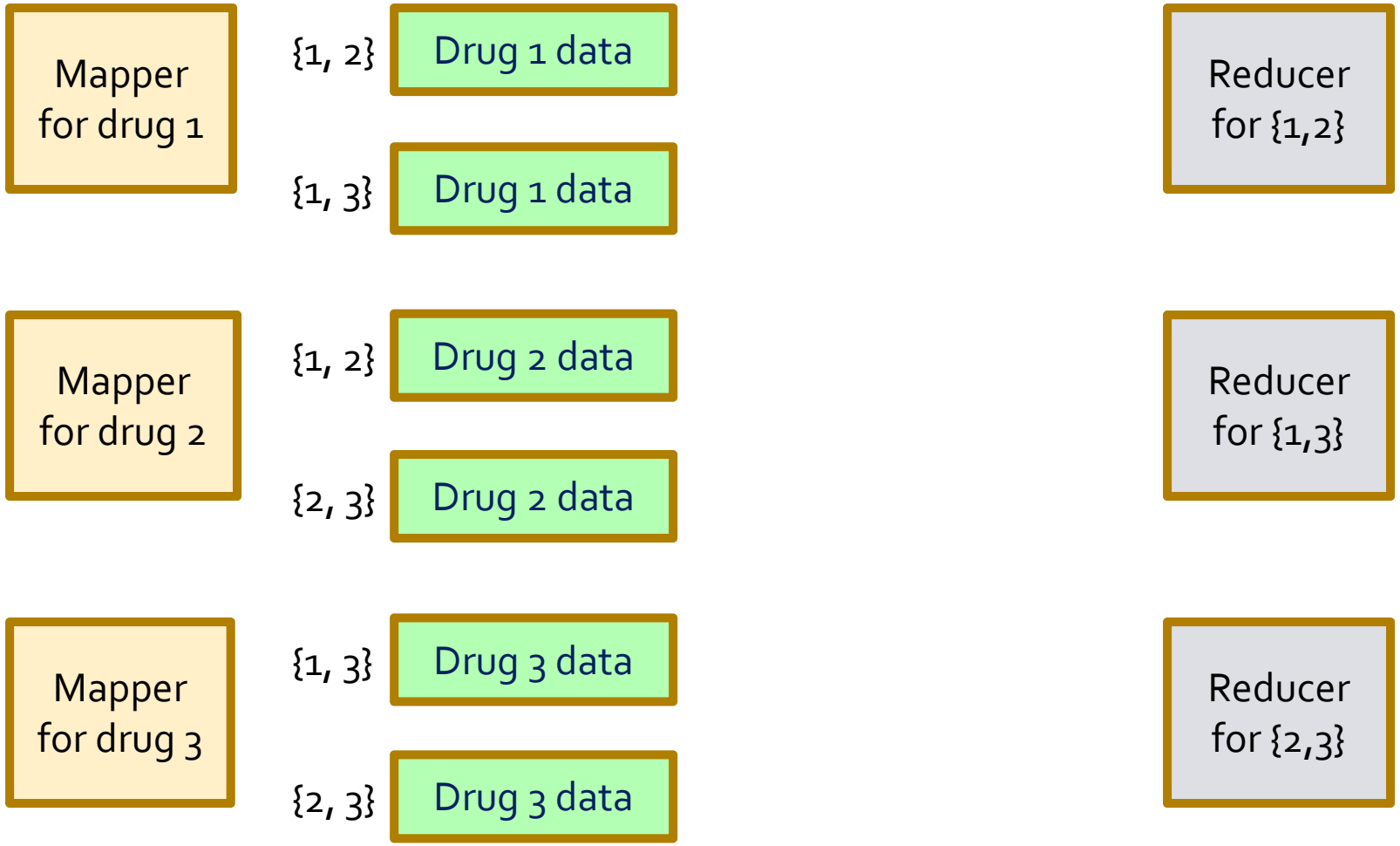
Initial Map-Reduce Algorithm

- The first attempt used the following plan:
  - Key = set of two drugs $\{i, j\}$.
  - Value = the record for one of these drugs.
- Given drug i and its record R_i, the mapper generates all key-value pairs $(\{i, j\}, R_i)$, where j is any other drug besides i.
- Each reducer receives its key and a list of the two records for that pair: $(\{i, j\}, [R_i, R_j])$.

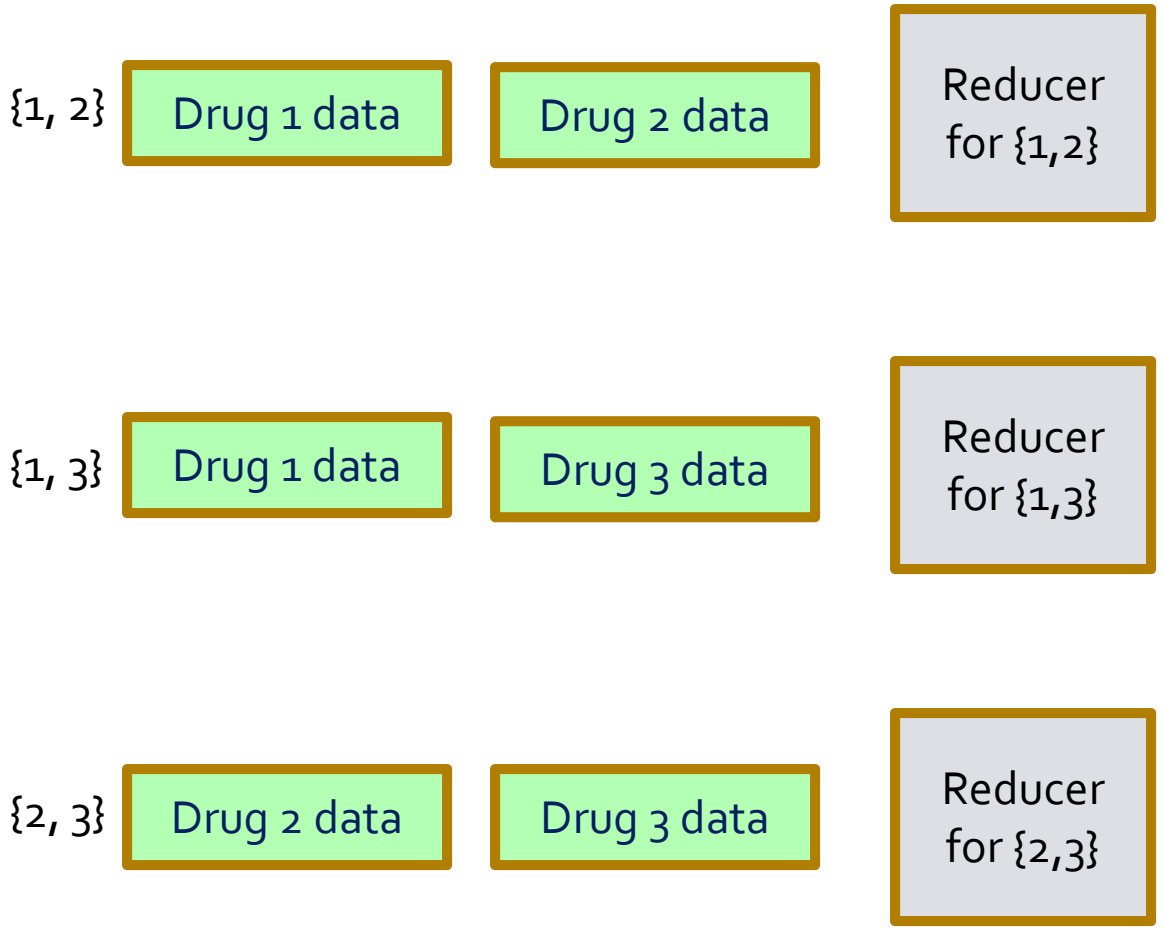
Example: Three Drugs



Example: Three Drugs



Example: Three Drugs



What Went Wrong?

- 3000 drugs
- times 2999 key-value pairs per drug
- times 1,000,000 bytes per key-value pair
- = 9 terabytes communicated over a 1Gb Ethernet
- = 90,000 seconds of network use.

A Better Approach

- The way to handle this problem is to use fewer keys with longer lists of values.
- Suppose we group the drugs into 30 groups of 100 drugs each.
  - Say G_1 = drugs 1-100, G_2 = drugs 101-200,…, G_{30} = drugs 2901-3000.
  - Let g(i) = the number of the group into which drug *i* goes.

The Map Function

- A key is a set of two group numbers.
- The mapper for drug *i* produces 29 key-value pairs.
  - Each key is the set containing g(i) and one of the other group numbers.
  - The value is a pair consisting of the drug number i and the megabyte-long record for drug *i*.

The Reduce Function

- The reducer for pair of groups {*m*, *n*} gets that key and a list of 200 drug records – the drugs belonging to groups *m* and *n*.
- Its job is to compare each record from group *m* with each record from group *n*.
  - Special case: also compare records in group *n* with each other, if *m* = *n*+1 or if *n* = 30 and *m* = 1.
- Notice each pair of records is compared at exactly one reducer, so the total computation is not increased.

The New Communication Cost

- The big difference is in the communication requirement.
- Now, each of 3000 drugs' 1MB records is replicated 29 times.
  - Communication cost = 87GB, vs. 9TB.