Alternating Least Squares in Spark

Patrick Landreman

Collaborative Filtering estimates unknown matrix values

Product
(Physical Measurements)

Users

(Patients)

M
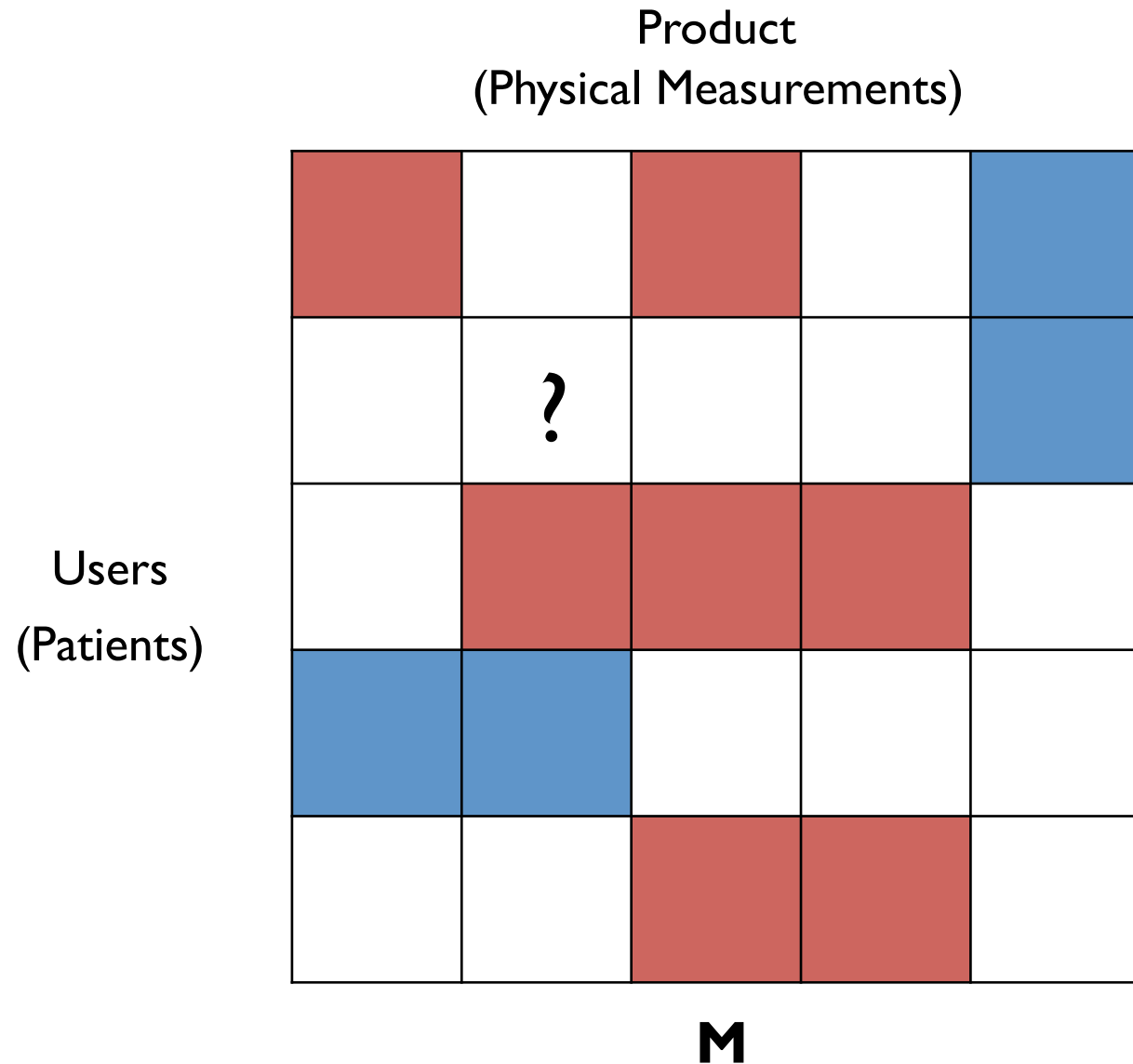
Spark estimates M using **Alternating Least Squares**
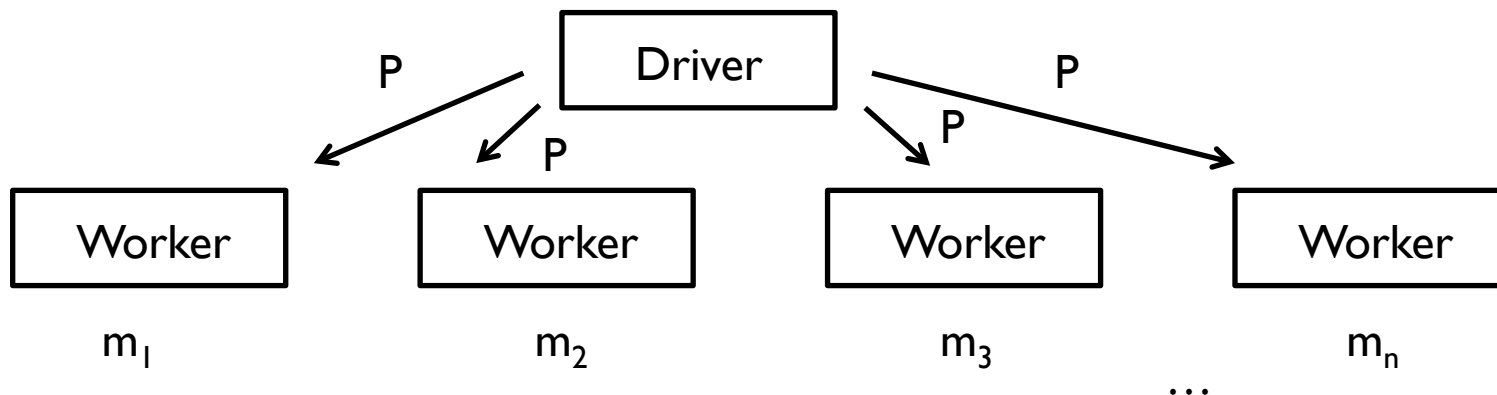
$$\hat{M} = PQ$$
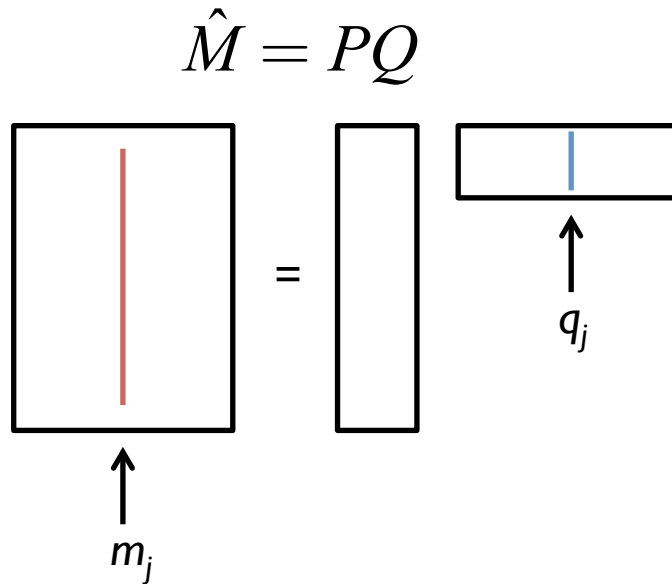
can solve for each column in parallel:

$$q_j = \left(P^T S P + \lambda I \right)^{-1} P m_j$$

$$S_{ii} = \begin{cases} 1 & \text{if } r_{ij} \neq 0 \\ 0 & \text{else} \end{cases}$$

Naïve attempt using Broadcasts

$$q_j = \left(P^T S P + \lambda I \right)^{-1} P m_j$$

Problems:

- Matrix arithmetic not implemented in Spark

- LAPACK requires flat arrays (how much do data conversions cost?)

- Setup time for sbt + LAPACK, etc.