SVD For matrix Completion

The problem:

We have an incomplete Matrix (R)

$$R = \begin{bmatrix} 2 & 1 & \boxed{?} & 5 & 3 \\ 1 & 4 & 3 & 2 & \boxed{?} \\ 2 & \boxed{?} & 4 & 3 & 3 \end{bmatrix}$$

Can we predict the incomplete entries?

⇒ We cannot develop any system that can provide an exact answer

⇒ But we can develop a protocol that can give ~~a~~ very good estimates for the missing values

⇒ But How is that possible?

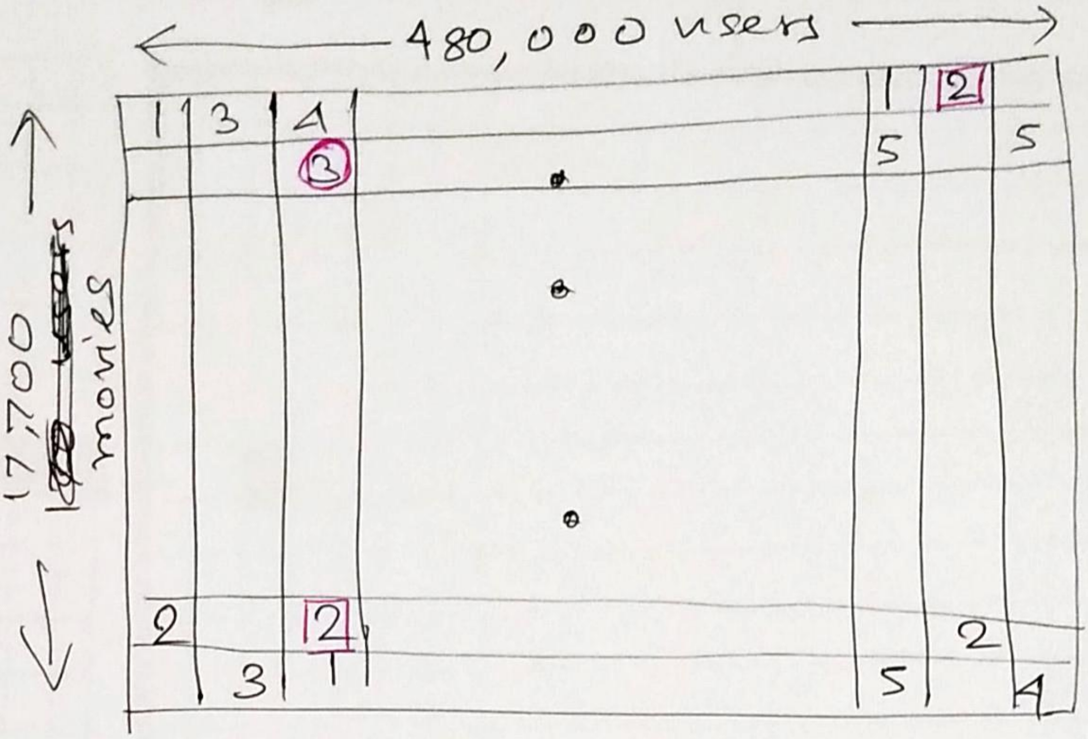
⇒ ~~Be~~, We presume that the entries of the matrix is an outcome of a physical process & not ~~becau~~ random numbers.

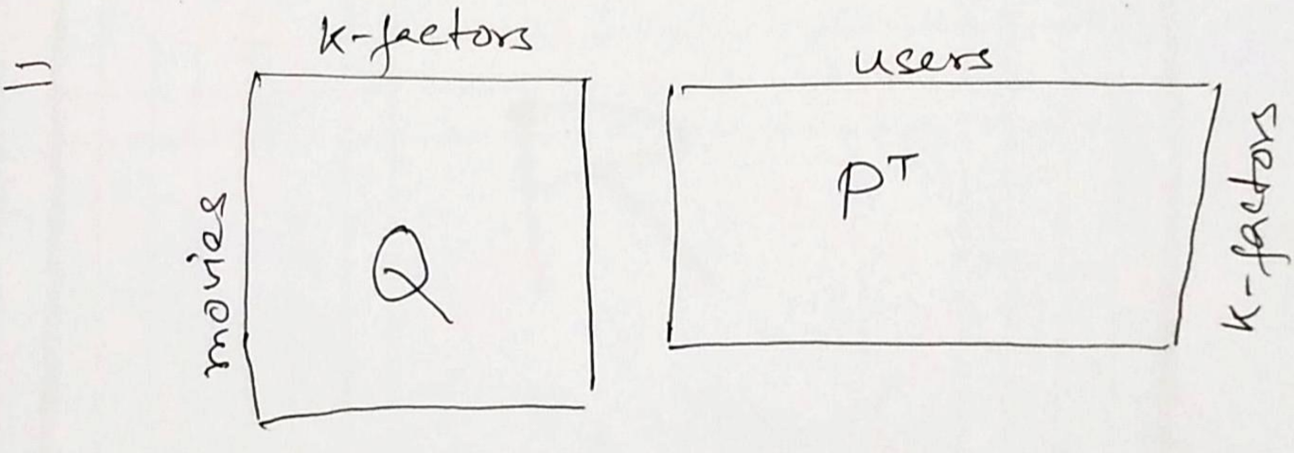
⇒ Therefore the idea is to find ~~the~~ the latent space of the physical process by the avialable entries. Then use ~~thes~~ these latent variables to estimate the missing entries.

SVD & Matrix Completion : Pg 2

This is a billion \$ business:

Netflix movie rating predictions: R



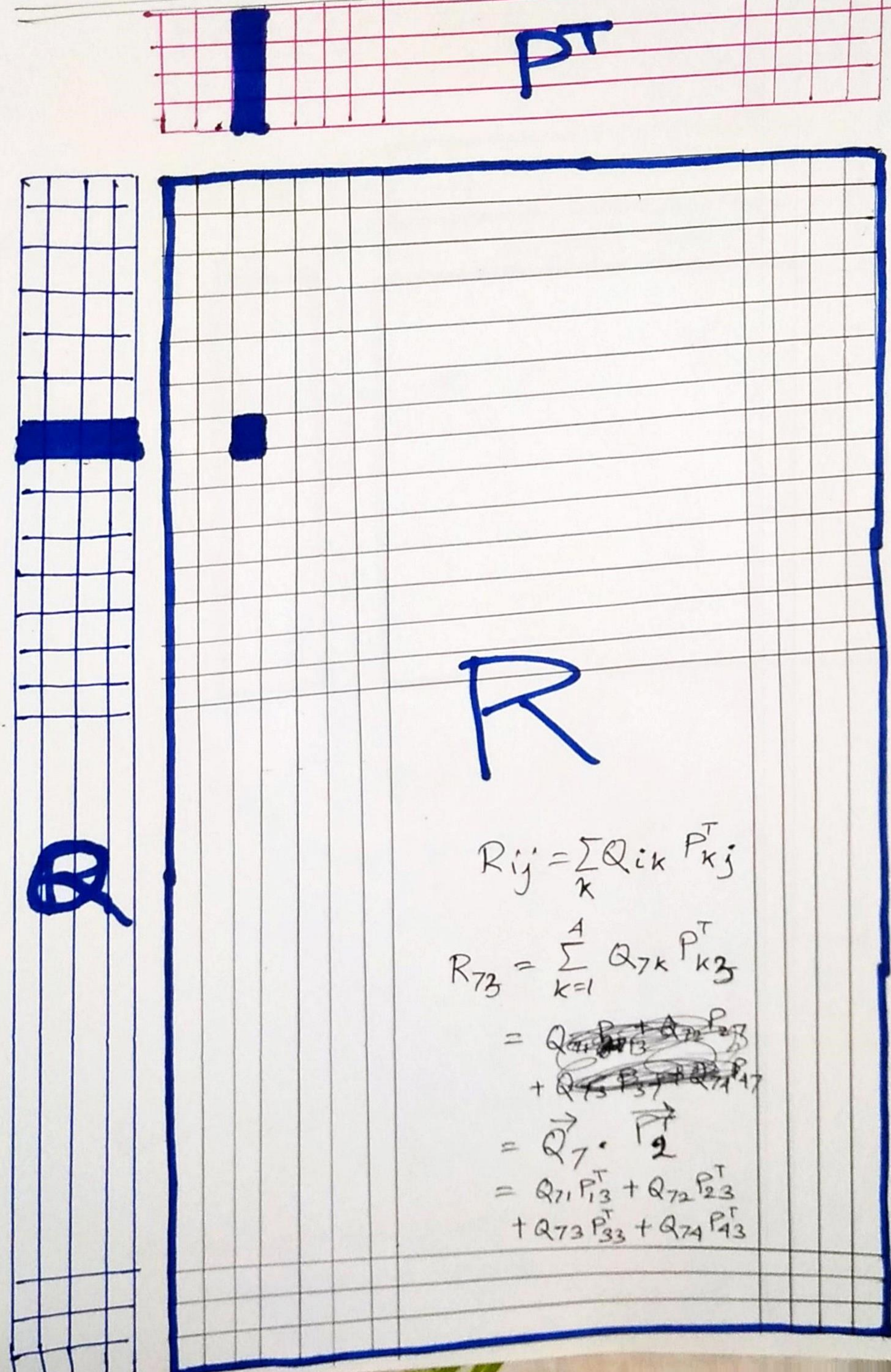


Thin k-factors

SVD: $A = U \Sigma V^*$ or $U \Sigma V^T$

Problem: we have missing values in R

SVD & Matrix completion: Pg 3 , P & Q are thin matrices



$$R_{ij} = \sum_k Q_{ik} P^T_{kj}$$

$$R_{73} = \sum_{k=1}^{4} Q_{7k} P^T_{k3}$$

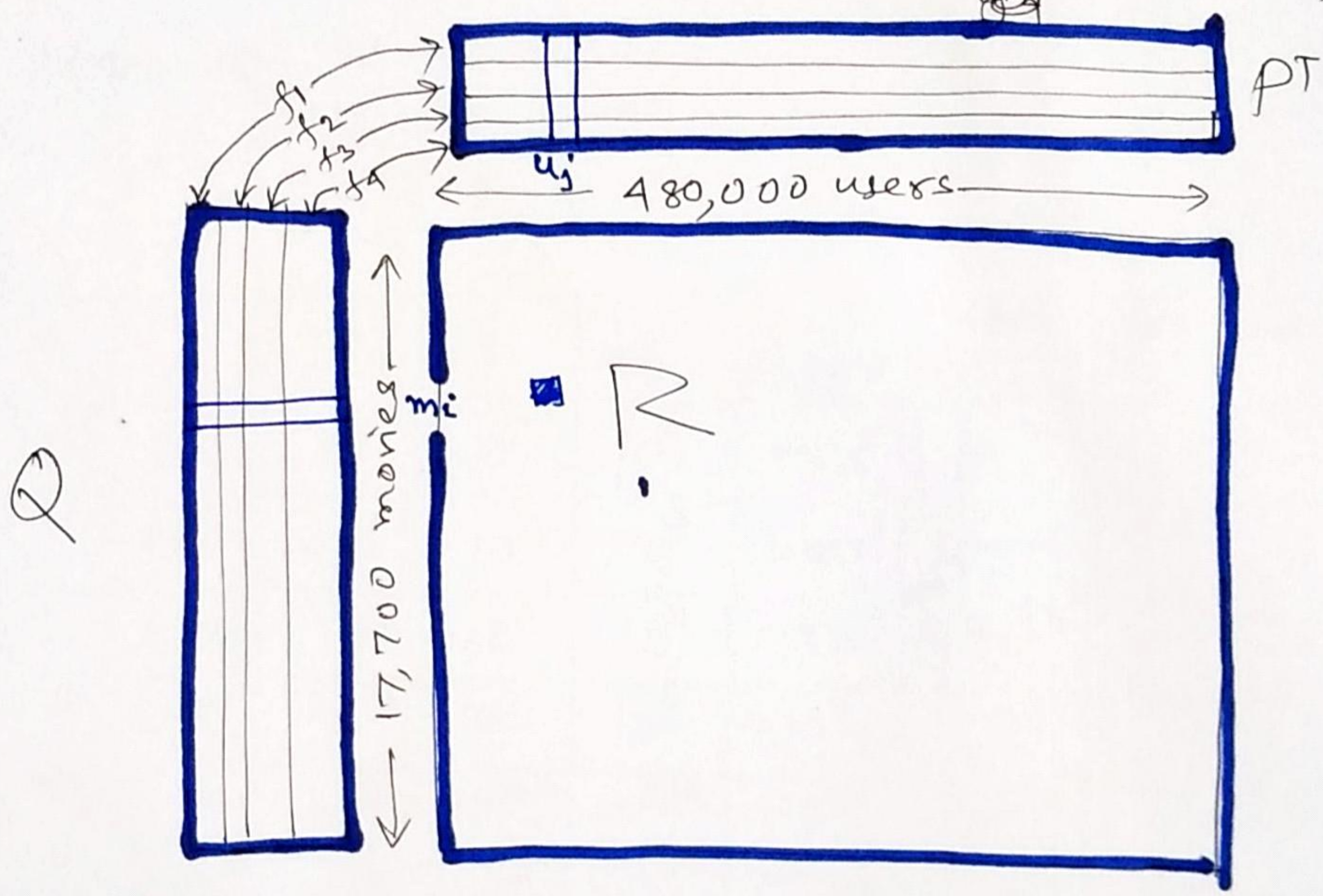
$$= \cancel{Q_{71} P_{13} + Q_{72} P_{27}} + \cancel{Q_{73} P_{37} + Q_{74} P_{47}}$$

$$= \vec{Q}_7 \cdot \vec{P^T_3}$$

$$= Q_{71} P^T_{13} + Q_{72} P^T_{23} + Q_{73} P^T_{33} + Q_{74} P^T_{43}$$

SVD & Matrix Completion 2 Pg 4

Factors k : sci-fi (f_1), romance (f_2), comedy (f_3), crime-~~thir~~ thriller (f_4) etc.



$$R_{ij} = \vec{Q}_i \cdot \vec{P}_j^{\,T}$$

$$= \sum_k Q_{ik} P^T_{kj}$$

$$= \sum_k Q_{ik} P_{jk}$$

⇐ Note: transpose removed 'coz index exchanged

But what to do with missing values in R ?

⇒ We can not perform the SVD in the traditional sense : Numerical methods will fall apart !!!

SVD & Matrix Completion: Pg 5

Let us perform a back-of-the-envelop calculation to understand whether it is a ~~solvaba~~ solvable problem or not!

Size of $R \to m \times n$

" " $Q \to m \times k$

" " $P^T \to k \times n$

m	n	k	R	P	Q	Total Storage (P+Q)	Compression ratio R/(P+Q)	Comment
10^3	10^3	100	10^6	10^5	10^5	2×10^5	5	
10^6	10^6	100	10^{12}	10^8	10^8	2×10^8	5000	
10^9	10^9	100	10^{18}	10^{11}	10^{11}	2×10^{11}	5×10^6	
10^9	10^5	100	10^{14}	10^7	10^{11}	$\sim10^{11}$	10^3	
10^{23}	10^4	100	10^{27}	10^6	10^{25}	$\sim10^{25}$	10^2	
10^{23}	10^{23}	100	10^{46}	10^{25}	10^{25}	$\sim10^{25}$	10^{21}	

→ Number of unknowns $\sim 10^{11}$

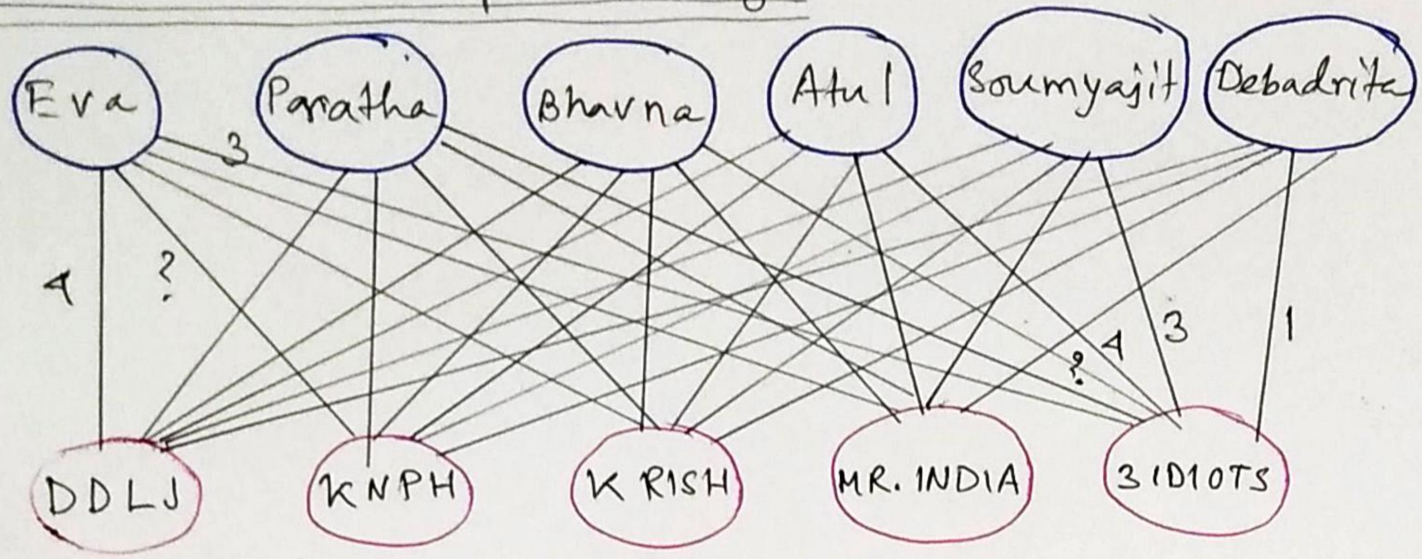
Even if 1% Entries in R is available we have $10^{14} \times 10^{-2} = 10^{12}$ equations available

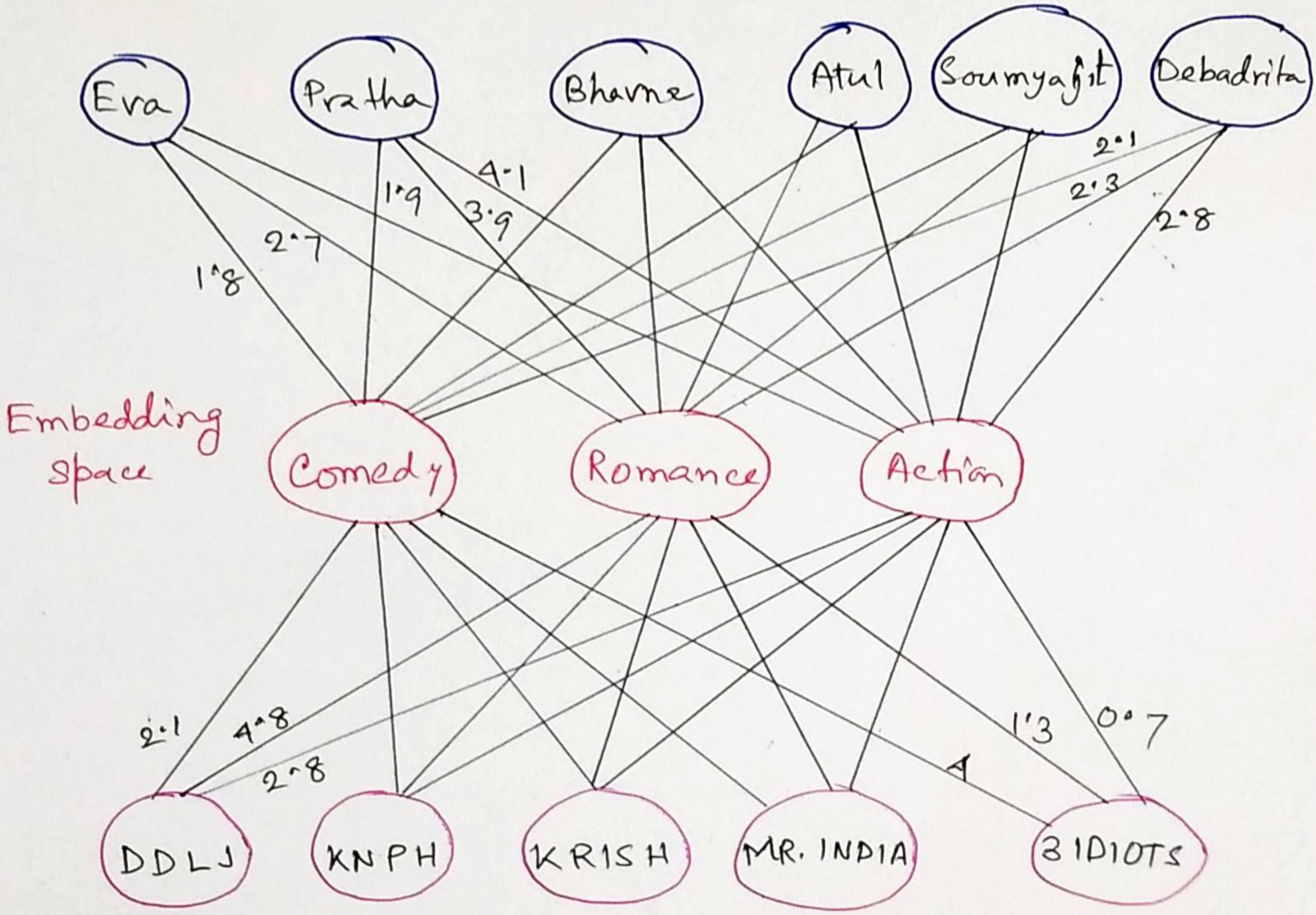
⇒ So ~~these~~ we still have nearly 10 times more equations avaiable than needed

⇒ These systems of equations must be solvable.

⇒ Solution method: → one possible candidate: Gradient descent

SVD & Matrix completion : Pg 6





Discuss the connection with Auto-encoders.