DMML 9 Apr 2019

Informational Retrieval & Web Search

Page Rank

Given a query / topic

Retrieve a collection of relevant documents

Classify these as authorative sources of information vs sources of links to information

Jon Kleinberg (same year as Page Rank)

Hubs

Authorities

Stars & Directors

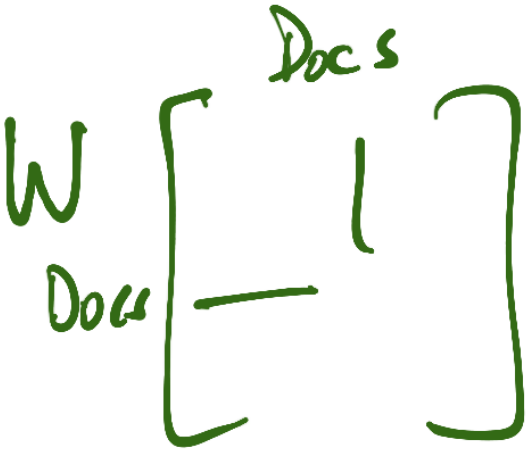
Likewise hubs get value from the quality of authorities they point to

& vice versa

document i

$$h(i) = \sum_j a(j) W[i,j]$$

hub score



Symmetrically

$$a(i) = \sum_j h(j) W[j,i]$$

$$H = WA$$

$$A = W^T H$$

Recall PageRank

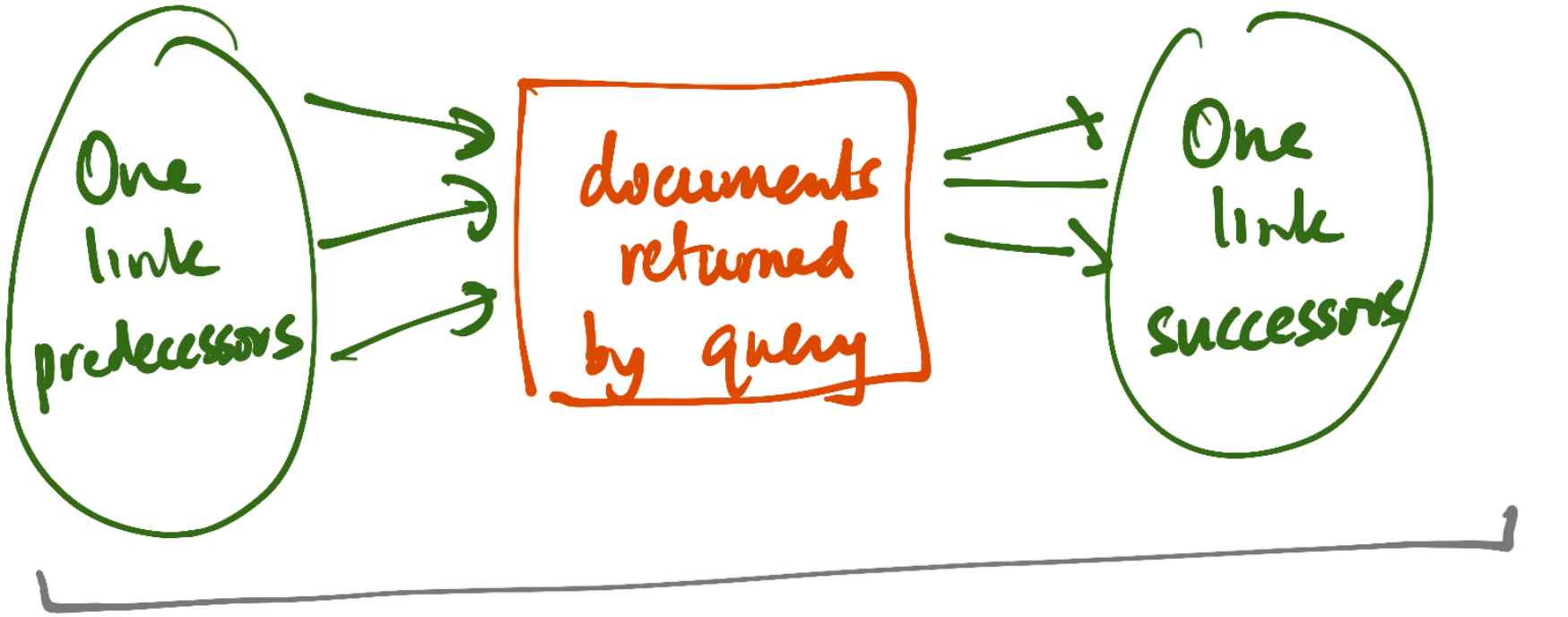
$$\pi = M\pi$$

$$H = WW^T H$$

$$A = W^T W A$$

What is W? Which documents ?

query → relevant documents



Control size

200 documents in starting set

Limit how many new documents each one "pulls in" (50)

W is the incidence matrix of this set

L need not be irreducible, aperiodic

W may have multiple "largest" eigenvalues

↳ indicative of a partition (not irreducible)

Ambiguous words

Jaguar

Two "communities"

Kleinberg called this HITS

Hypetext Induced Topic Search

Changing gears

Two problems with search

Words have different meanings

Multiple words have same meaning

Vector Space

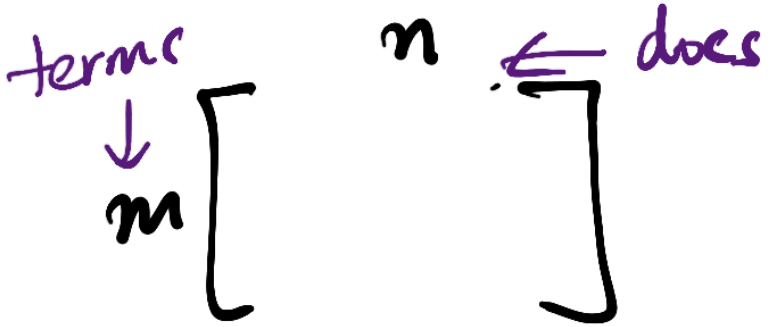
Term Doc Matrix with weights

Queries & documents as vectors and

compare using cosine similarity

Singular Value Decomposition (SVD)

D is our term-doc matrix



Unique way to write

$$D = U \Sigma V^T \quad \text{s.t.}$$

$m \times n \qquad m \times m \quad m \times n \quad n \times n$

U eigenvectors of DD^T, $U^TU = I$

V eigenvectors of D^TD, $V^TV = I$

Σ diagonal matrix

$$\begin{matrix} & & n \\ & & \downarrow \\ m \rightarrow & \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & & \ddots \\ 0 & & \sigma_r \end{bmatrix} & 0 \end{matrix}$$

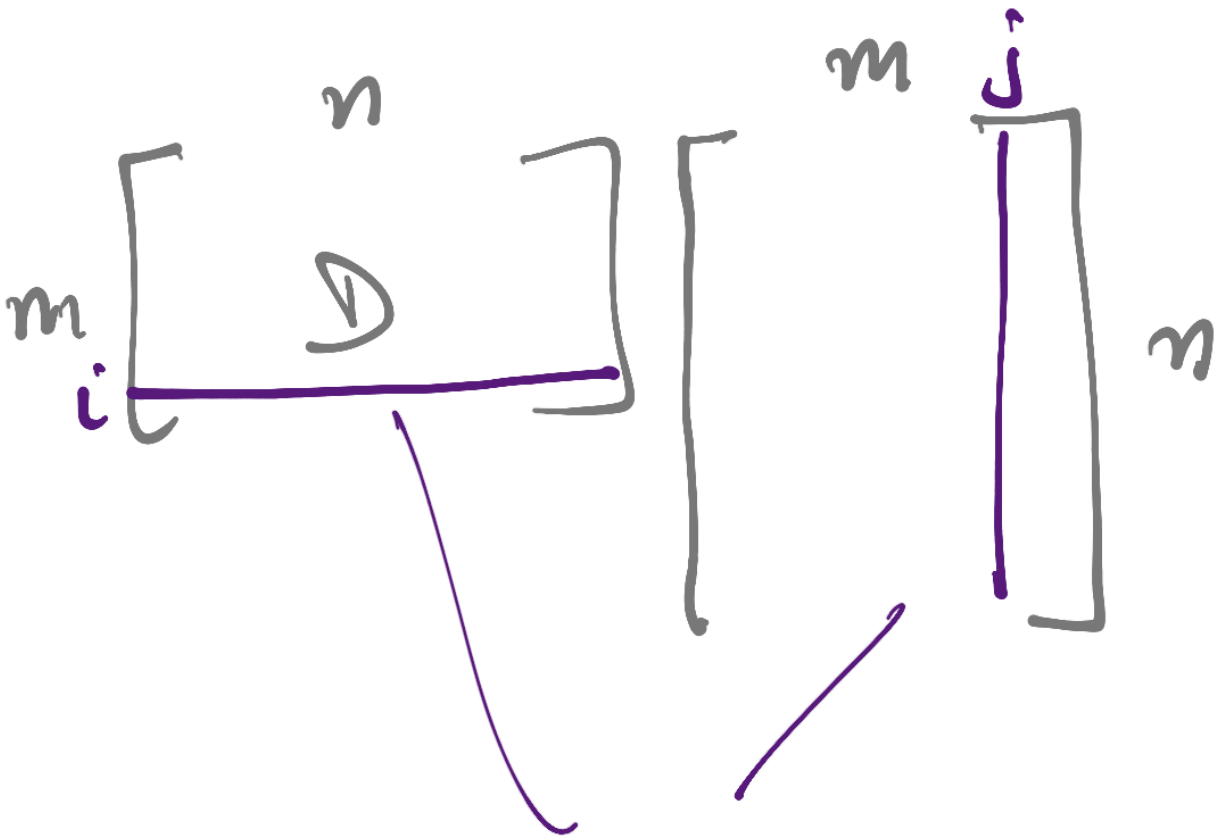
$\sigma_i = \sqrt{\lambda_i}$,

eigenvalues of D

in decreasing order

"singular values".

U eigenvectors of DD^T



Sum up all cooccurrences of

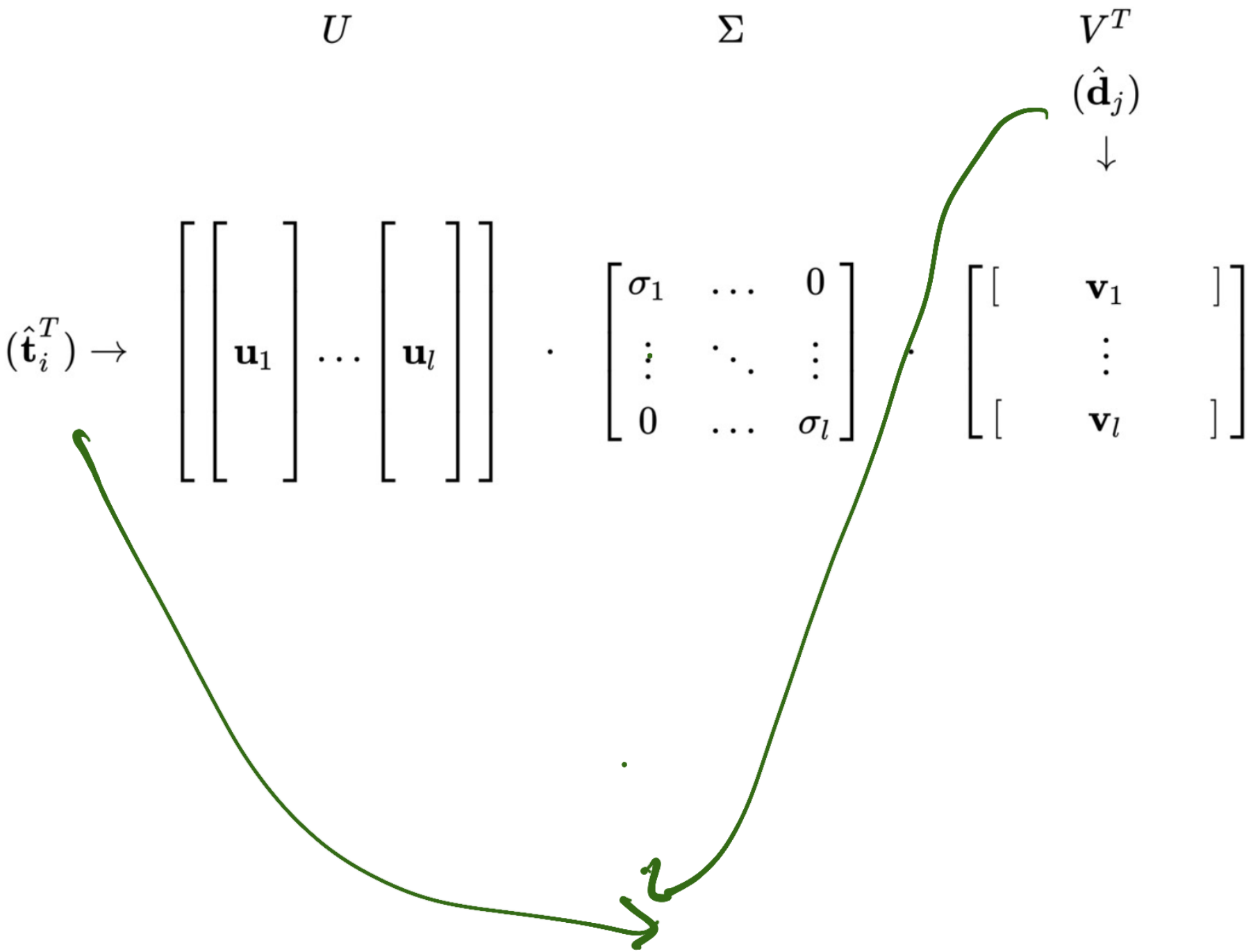
t_i, t_j

$V = D^T D$

- similarity across documents

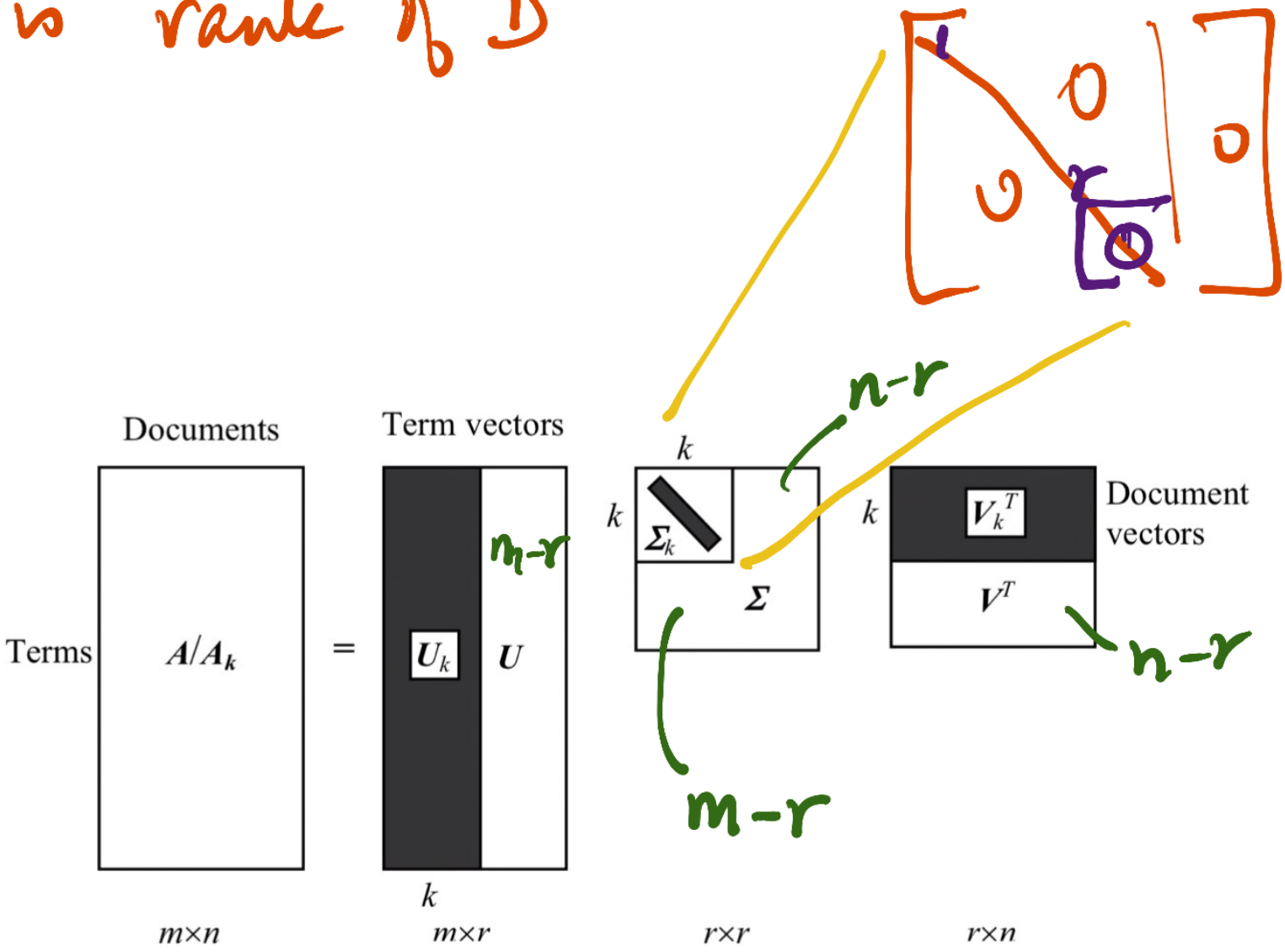
D

$$
\begin{array}{c}
 & & (\mathbf{d}_j) & & \\
 & & \downarrow & & \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,j} & \cdots & x_{m,n} \end{bmatrix}
\end{array}
$$


U
Σ
V^T
$(\hat{\mathbf{d}}_j)$
$(\hat{\mathbf{t}}_i^T) \rightarrow$
$\mathbf{u}_1$
$\mathbf{u}_l$
σ_1
σ_l
$\mathbf{v}_1$
$\mathbf{v}_l$

Σ has r singular values – eigenvalues of D

r is rank of D



Simplify

U to $m \times r$

Σ to $r \times r$

V to $r \times n$

$$
\begin{array}{ccccc}
 & U & & \Sigma & V^T \\
 & & & & (\hat{\mathbf{d}}_j) \\
 & & & & \downarrow \\
(\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot & \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix} & \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}
\end{array}
$$

$\hat{t}_i$ is transformed version of t_i

$\hat{d}_j$ is transformed version of d_j

Symmetric matrix M

Eigenvectors are orthogonal

λ_1 λ_2 λ_3 M is 3x3

v_1 v_2 v_3

Any vector $u = (u_1, u_2, u_3)$

can be expressed as

$$n_1 v_1 + n_2 v_2 + n_3 v_3$$

$$M \cdot u$$

$$= M\left(n_1 v_1 + n_2 v_2 + n_3 v_3\right)$$

Eigenvector $\quad M v_i = \lambda_i v_i$

$$= (\lambda_1 v_1) n_1 + (\lambda_2 v_2) n_2 + (\lambda_3 v_3) n_3$$

Suppose $\lambda_1 > \lambda_2 > \lambda_3$ (small)

Ignore λ_3, v_3, approximation of $M \cdot u$

Approximation of M

D - restrict to k most significant terms (directions)

document

rank k

keep only top k eigenvalues

Best approx is to truncate Σ to Σ_k

$$D = U \Sigma V^T$$

$$m \times r \quad r \times r \quad r \times n$$

$$\downarrow$$

$$D' = m \times k \quad k \times k \quad k \times n$$

$$\sum (D_{ij} - D'_{ij})^2$$

k-approx ∩ SVD

How ?

$$
\begin{array}{ccccc}
 & U & & \Sigma & V^T \\
 & & & & (\hat{\mathbf{d}}_j) \\
 & & & & \downarrow \\
(\hat{\mathbf{t}}_i^T) \rightarrow & \left[\left[\begin{array}{c} \\ \mathbf{u}_1 \\ \\ \end{array}\right] \cdots \left[\begin{array}{c} \\ \mathbf{u}_l \\ \\ \end{array}\right]\right] & \cdot & \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix} & \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}
\end{array}
$$

What is $\hat{d}_j$?

$$D = U_k \Sigma_k V_k^T$$ ← often k approx.

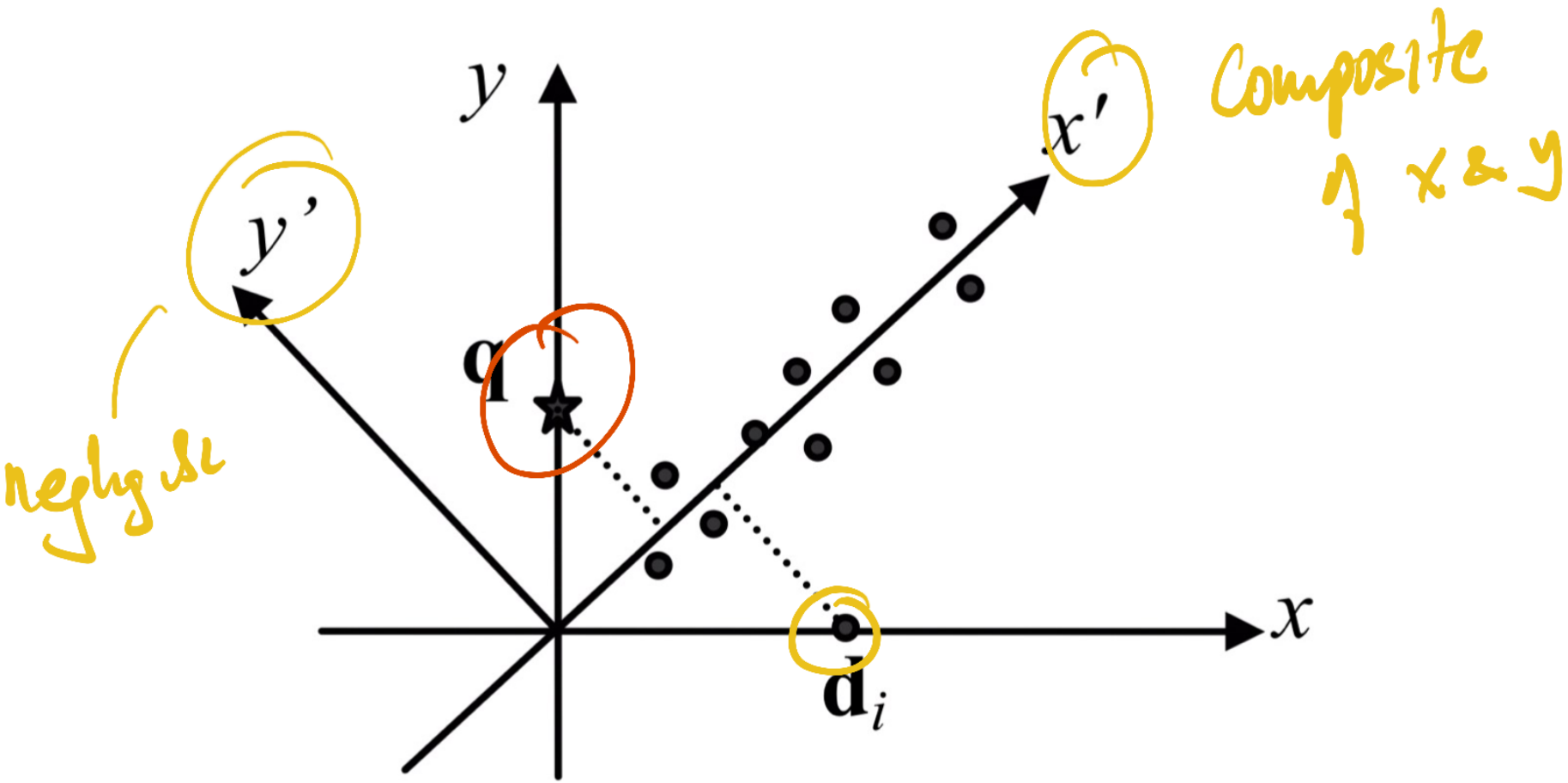
Want V_k^T – representation of documents

$$U^T D = U^T U \Sigma V^T$$

$U^T U = I$

$$\Sigma^{-1} U^T D = V^T$$

Σ^{-1}: $1/\sigma_i$



Rotate a query similarly

$$\hat{q} = \Sigma^{-1} U^T q$$

Compare $\hat{q}$ & $\hat{d}_j$ as vectors

Latent Semantic Indexing