

The background of the slide is a complex, abstract composition. It features a dark, muted purple or brownish background. Overlaid on this are several geometric and data-like elements. In the upper and lower portions, there are intricate networks of thin, light-colored lines forming a mesh or web-like structure. Scattered throughout these networks are numerous small, colored dots in shades of green, blue, and yellow. On the left side, there is a vertical strip containing a grid of small, light-colored plus signs. In the center, a large, white, angular shape, resembling a stylized letter 'A' or a folded piece of paper, serves as a backdrop for the title. The overall aesthetic is technical and modern, suggesting themes of data science, mathematics, or computer graphics.

# **Session 9: Dimensionality Reduction Methods**

# In-video Quiz

---

- ☐ Suppose we have word-document co-occurrence matrix  $A$  with number of words  $n = 4M$  and number of documents  $d = 50M$ . We want to approximate matrix  $A$  via Nonnegative Matrix Factorization. Which of the following choices of  $k$  would be more suitable in this application?
  - ☐  $k = 10$
  - ☐  $k = 512$
  - ☐  $k = 4M$
  - ☐  $k = 50M$ .
- ☐ Answer:  $k = 512$
- ☐ Explanation: The goal is to approximately factorize the co-occurrence matrix  $A$  with two lower-rank matrices  $U$  and  $V$ . If  $k = 4M$  or  $50M$ , the ranks of  $U$  and  $V$  could be large. In particular, for the extreme cases,  $U$  and  $V$  are full rank. If  $k = 10$ , the ranks of  $U$  and  $V$  are too small, leading to a bad approximation of matrix  $A$ .  $k = 512$  could be a suitable choice for the rank, which could give good approximation of  $A$  while enjoying the low ranker property.