Presentation Outline & Slides

Part 1: Background Introduction & Theory & Overview of Algorithms(Speaker A)

Slide 1: Title Slide

- **Title:** Efficient Low-Rank Matrix Approximation: Normal vs. Randomized Algorithms
- **Subtitle:** Analysis on Real-World Data (MovieLens & Volcano)
- **Team Members:** [Name 1], [Name 2], [Name 3]
- **Date:** December 8, 2024

Slide 2: Overview & The "Trilemma"

- **Visual:** A triangle diagram labeled "Accuracy", "Efficiency", "Interpretability".
- **Bullet Points:**
  - **Goal:** Decompose matrix A to uncover latent structures.
  - **Algorithms Implemented:**
    1. **SVD:** The Gold Standard (Best Accuracy).
    2. **NMF:** Parts-based representation (Non-negative, Interpretability).
    3. **CUR:** Selecting actual columns/rows (Data-centric Interpretability).
  - **The Challenge:** Big data requires faster, more stable methods than classical deterministic approaches.

Slide 3: Our Approach: Deterministic vs. Randomized

- **Visual:** A split screen graphic. Left: "Normal (Deterministic)", Right: "Randomized".
- **Content:**
  - **SVD:** LAPACK (Normal) vs. Random Projections (Randomized).
  - **NMF:** Random Initialization vs. **Optimized (rSVD) Initialization** (Warm Start).
  - **CUR:** Top Leverage Scores vs. Weighted Sampling.
  - **Objective:** Can randomization improve speed or stability without losing accuracy?

Part 2: Implementation & Evaluation (Speaker B)

Slide 4: Implementation from Scratch

- **Visual:** Code snippets or flowcharts (e.g., the Halko et al. rSVD flow).
- **Key Highlights:**
  - **Pure R Implementation:** No `sklearn` or wrapper packages.
  - **Randomized SVD Logic:** Random Projection $\rightarrow$ Power Iteration $\rightarrow$ QR $\rightarrow$ Low-rank SVD.

- **The "Secret Weapon" for NMF:** Using the output of Randomized SVD to initialize NMF ($W_{init} = |U_{rSVD}|$). This is our "Warm Start" strategy.

Slide 5: Real-World Datasets

- **Visual:** Two images side-by-side.
  - Left: A sparse matrix heatmap (mostly white/empty) -> **MovieLens 100k**.
  - Right: A detailed topographic map -> **Volcano Topography**.
- **Text:**
  - **MovieLens:** Highly Sparse (90%+ zeros). Challenge: Missing data recovery.
  - **Volcano:** Dense, High correlation. Challenge: Image compression / Feature extraction.

Slide 6: Result 1 - The "Warm Start" Effect (MovieLens)

- **Visual:** Comparison plot of NMF Reconstruction Error.
  - Normal NMF (Random Init): High error, unstable.
  - Optimized NMF (rSVD Init): Low error, converges instantly.
- **Key Finding:** On sparse data, standard NMF gets stuck. Our SVD-initialized NMF solves the non-convex problem much better.

Slide 7: Result 2 - Natural Low-Rank (Volcano)

- **Visual:** Original Volcano image vs. Randomized SVD Reconstruction (Rank 10).
- **Key Finding:**
  - Visuals are indistinguishable.
  - Real physical data has rapid singular value decay.
  - Randomized SVD captures global structure efficiently even with rank $k = 5$.

Part 3: Demo & Discussion (Speaker C)

Slide 8: Shiny App Showcase

- **Visual:** Screenshot of the "Comparison Lab" tab in the Dashboard.
- **Features:**
  - **Real-time Benchmarking:** "Run & Compare" button.
  - **Interactive Controls:** Switch between MovieLens/Volcano, adjust Rank k.
  - **Deep Dive:** Inspecting Eigenfaces and Leverage Scores.

Slide 9: Key Findings & Discussion

- **Visual:** Summary Table.
- **Points:**
  - **Sparsity Matters:** Randomized initialization is crucial for NMF on sparse data (MovieLens).

- **Dimensionality:** Randomized SVD shines on "Tall-and-Skinny" matrices (like the flattened Volcano data).
- **Trade-off:** Deterministic CUR is more accurate for small data, but Randomized CUR is the only path for web-scale scalability.

Slide 10: Conclusion

- **Summary:** We successfully implemented robust matrix algorithms from scratch.
- **Takeaway:** Randomization is not just an approximation; it is a computational resource that provides stability (in NMF) and scalability (in SVD).
- **Q&A:** Thank you!

Script / Speech Draft

Speaker A: Introduction

[Slide 1: Title]

"Good morning/afternoon everyone. We are [Team Names]. Today, we present our final project on 'Efficient Low-Rank Matrix Approximation'. We focused on comparing classical Deterministic algorithms against modern Randomized approaches, using real-world data."

[Slide 2: Overview]

"Matrix factorization is the engine behind many data science tasks. For this project, we implemented three core algorithms:

First, SVD, the mathematical gold standard for accuracy.

Second, NMF, which enforces non-negativity to create interpretable, parts-based representations.

And third, CUR, which selects actual columns and rows from the dataset.

Our goal was to address the 'Trilemma' of Accuracy, Efficiency, and Interpretability."

[Slide 3: Approach]

"However, classical algorithms often struggle with big data. They can be slow or unstable.

So, we implemented two versions for each algorithm:

A 'Normal' version, using standard deterministic methods like LAPACK.

And a 'Randomized' version.

We specifically wanted to answer: Can injecting randomness actually make our algorithms more robust and stable? Now, [Speaker B] will explain how we built this."

Speaker B: Implementation & Results

[Slide 4: Implementation]

"Thanks. We implemented everything from scratch in R.

For SVD, we used the Halko probabilistic method: projecting data into a smaller subspace to speed up computation.

But our most significant design choice was in NMF. Standard NMF is non-convex and sensitive to initialization. We implemented an 'Optimized NMF' that uses our Randomized SVD results as a 'Warm Start'. This theoretically places the optimization closer to the global minimum."

[Slide 5: Datasets]

"To test this, we moved away from synthetic data and used two real-world datasets with opposite characteristics:

1. **MovieLens 100k:** This is user-rating data. It is **highly sparse**—mostly zeros.
2. **Volcano Topography:** This is dense physical data. It simulates image compression."

[Slide 6: Result 1 (MovieLens)]

"The results were striking. On the sparse MovieLens data, look at this comparison [Point to NMF plot].

The Normal NMF with random initialization struggles—it gets stuck in local minima.

However, our Optimized NMF (initialized with rSVD) converges almost immediately to a much lower error. This proves that 'Warm Start' is essential for sparse data."

[Slide 7: Result 2 (Volcano)]

"On the Volcano data, we tested Randomized SVD. Even with a very low rank of 10, the reconstruction [Point to Image] is nearly perfect. This confirms that real-world physical data has a 'natural low-rank' structure, which randomized algorithms capture extremely efficiently."

Speaker C: Demo & Conclusion

[Slide 8: Shiny App]

"To visualize these findings, we built this Shiny Dashboard.

[Show Demo or Screenshot]

Here in the 'Comparison Lab', we can run both algorithms head-to-head. You can see the 'Time' and 'Error' metrics side-by-side.

We also have a 'Deep Dive' tab, where we can inspect the actual basis vectors—for example, seeing the 'Eigenfaces' of the topography or the 'User Profiles' in MovieLens."

[Slide 9: Key Findings]

"So, what did we learn?

1. **Sparsity dictates strategy:** For sparse data like MovieLens, simple random initialization fails. Using SVD to initialize NMF is a game-changer.
2. **Randomization adds Stability:** It sounds counter-intuitive, but randomized algorithms often produced *more stable* results than deterministic ones on complex landscapes.
3. **Trade-offs:** While Deterministic CUR was slightly more accurate on our small subset, Randomized CUR is the only viable option for massive scaling."

[Slide 10: Conclusion]

"In conclusion, our project demonstrates that Randomized Linear Algebra is a powerful tool. It allows us to trade a tiny amount of accuracy for significant gains in stability and speed.

Thank you for listening, and we are happy to take any questions."