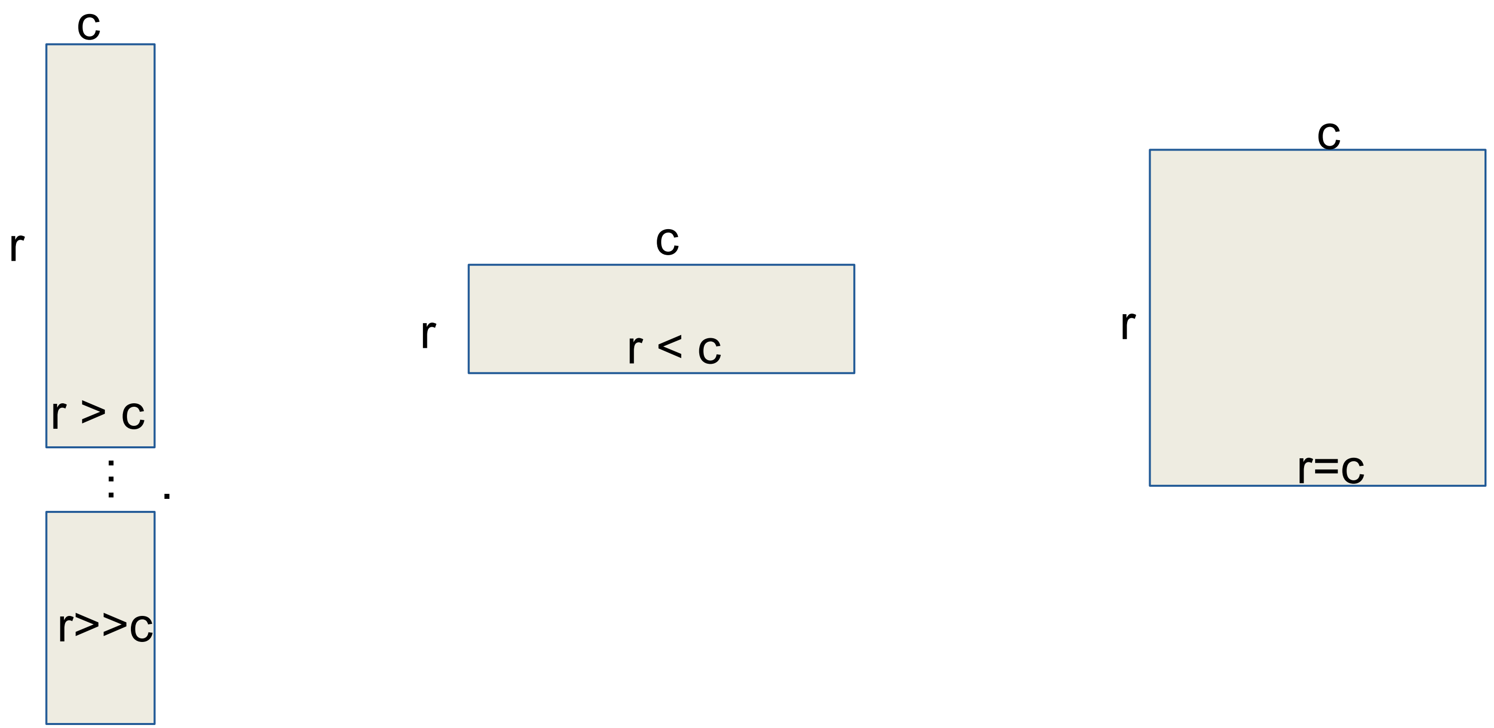
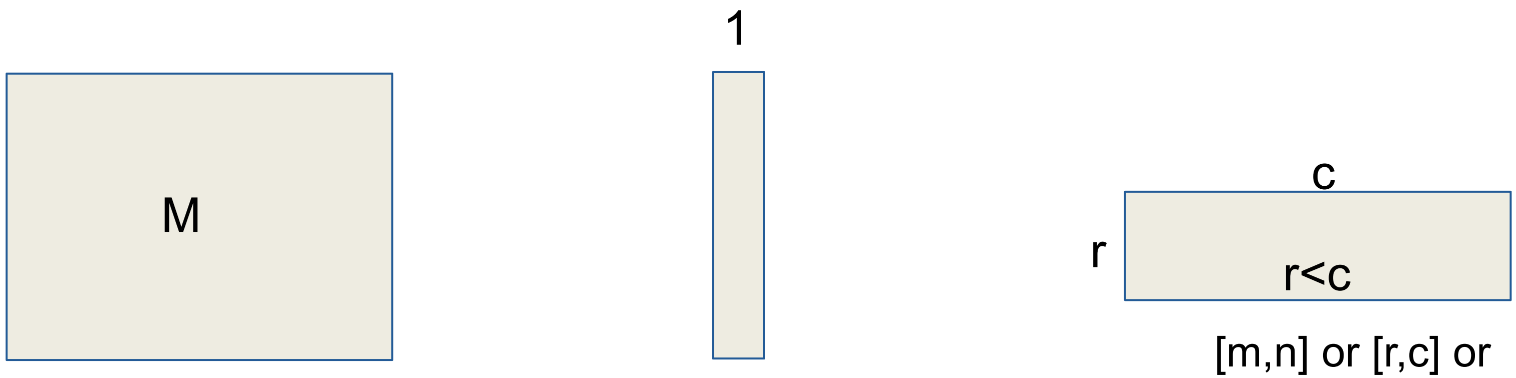
Introduction to Three Matrix Structures

Introduction to Three* Matrix Structures

*Interpreted four ways



Framing



M or X or df or data or observations

L or Y or target

[m,n] or [r,c] or [x,y] or a_ij

```
M = [[1,2,3,4],
     [4,5,6,7],
     [7,8,9,10]]
```

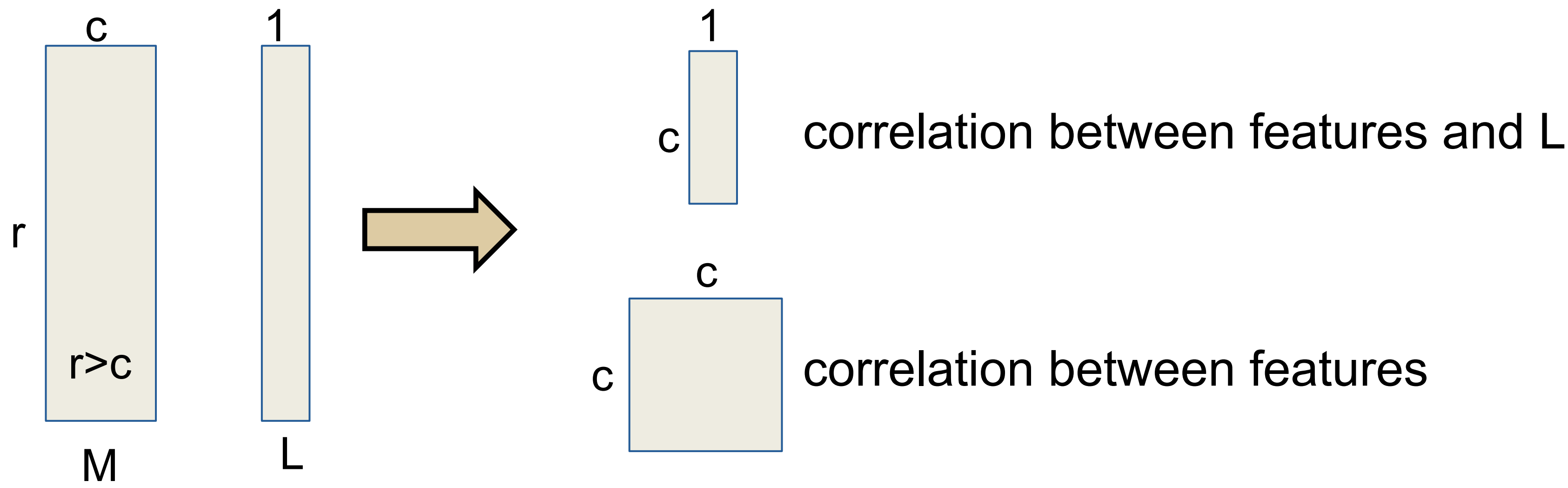
```
L = [1,2,3,4]
```

DataScience@SMU

Tall and Skinny Matrices

Tall and Skinny Matrices

Explicitly calculate correlation between columns



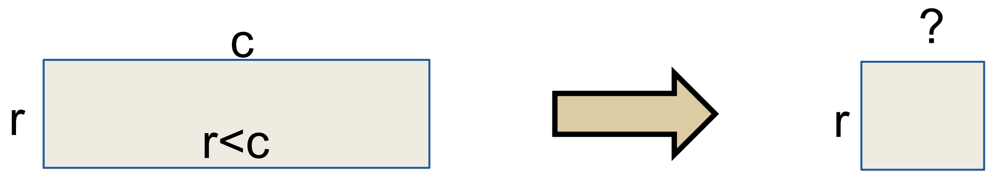
Classic question: Do we have enough data to estimate correlation between features and labels?

DataScience@SMU

Short and Fat Matrices

Short and Fat Matrices

Approximate the correlation between columns



Simplifying assumption: a good model doesn't require all the features.

PCA: preforms lossless compression on M independent of L

Lasso, Ridge Regression: utilizes L to identify features that are significant

Random Forest: utilizes L to identify features that are significant

DataScience@SMU

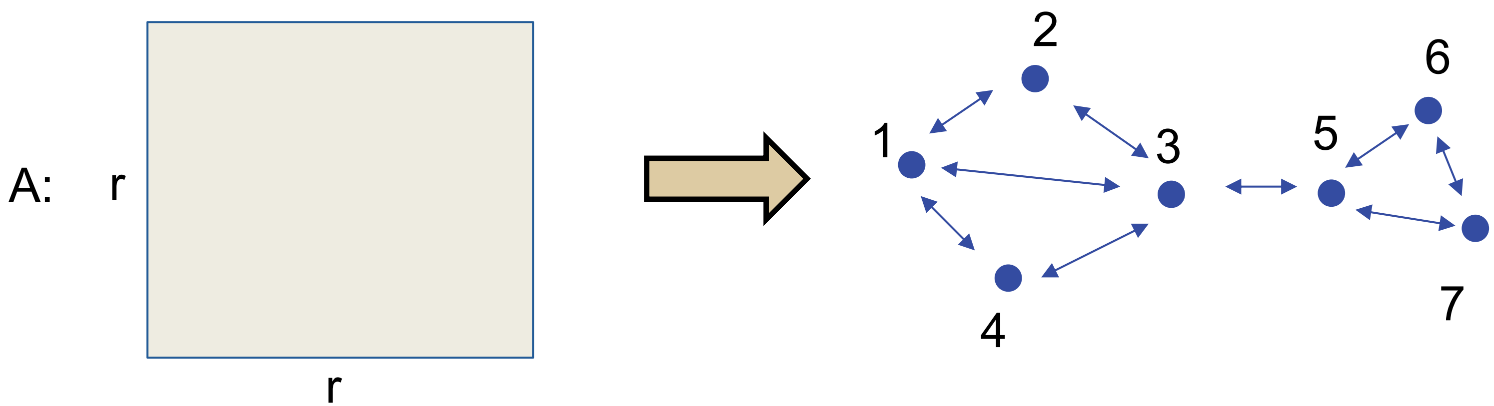
Square Matrices

Square Matrices

In network analysis, undirected graphs are represented as square matrices:

- Adjacency matrix (undirected graph example)
- Graph Laplacian (the diagonal is nonzero, contains the vertex degrees)

More generally, if we use a (dis)similarity function for each edge → dense square matric



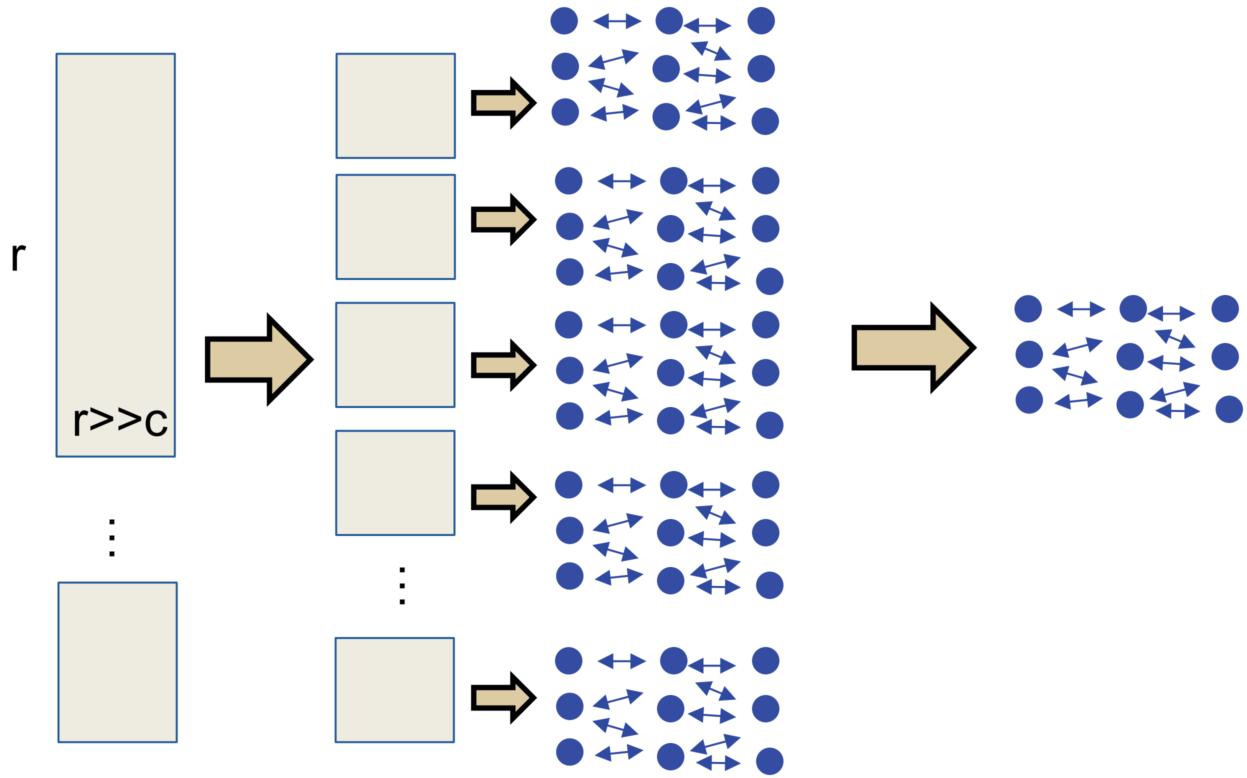
Every a_ij entry in A represents the relationship between vertex/item i and vertex/item j. Such representations are used in clustering, node centrality, and propagation models.

DataScience@SMU

Very, Very Tall and Skinny

Really Tall and Skinny Matrices

Explicitly calculate higher order correlations between columns



DataScience@SMU