NYPD Misconduct Database Analysis

COS 424/524, SML 302 Assignment 3

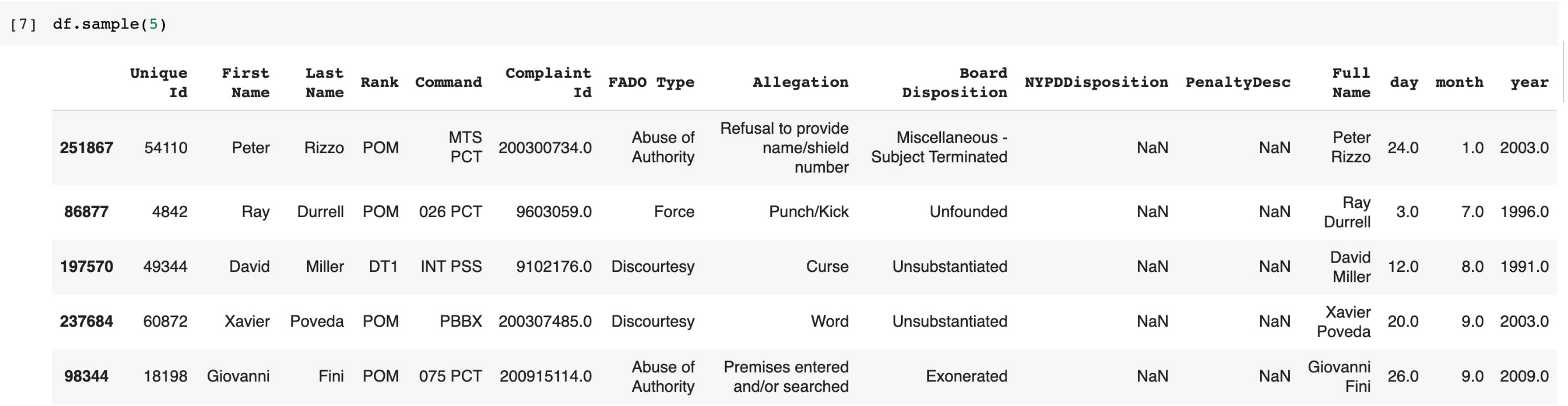
Unsupervised Learning

Hi! I’m Deniz

NYPD Misconduct Complaint Database

- Repository of complaints made by the public on record at the Civilian Complaint Review Board (CCRB).
- 323,911 records spanning 81,550 current or former NYPD officers!
- Obtained by NYCLU and made into search tool:
  - https://www.nyclu.org/en/campaigns/nypd-misconduct-database
- Further information on CCRB website:
  - https://www1.nyc.gov/site/ccrb/policy/data-transparency-initiative-allegations.page

NYPD Misconduct Complaint Database



```
[7] df.sample(5)
```

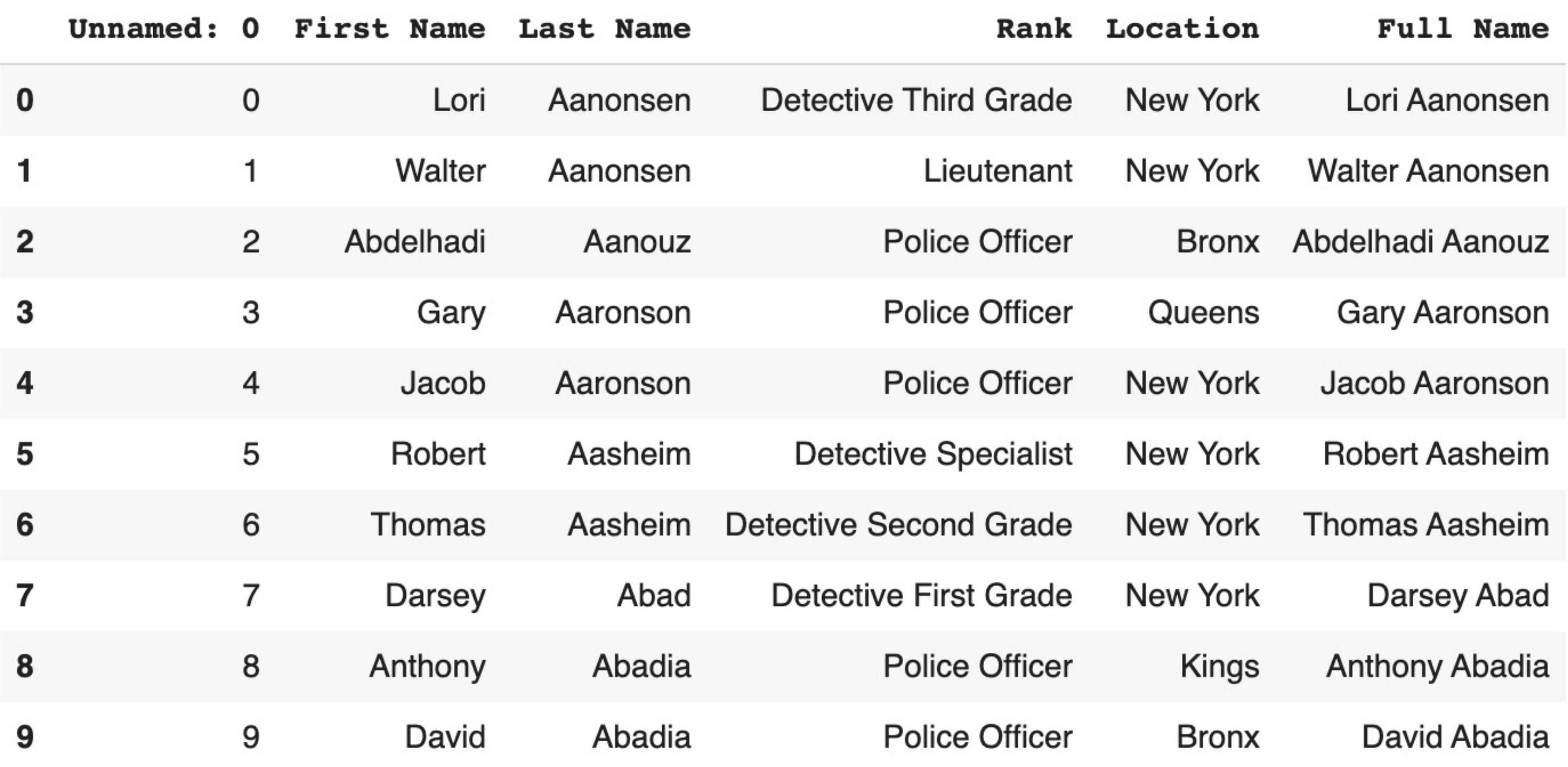
	Unique Id	First Name	Last Name	Rank	Command	Complaint Id	FADO Type	Allegation	Board Disposition	NYPDDisposition	PenaltyDesc	Full Name	day	month	year
251867	54110	Peter	Rizzo	POM	MTS PCT	200300734.0	Abuse of Authority	Refusal to provide name/shield number	Miscellaneous - Subject Terminated	NaN	NaN	Peter Rizzo	24.0	1.0	2003.0
86877	4842	Ray	Durrell	POM	026 PCT	9603059.0	Force	Punch/Kick	Unfounded	NaN	NaN	Ray Durrell	3.0	7.0	1996.0
197570	49344	David	Miller	DT1	INT PSS	9102176.0	Discourtesy	Curse	Unsubstantiated	NaN	NaN	David Miller	12.0	8.0	1991.0
237684	60872	Xavier	Poveda	POM	PBBX	200307485.0	Discourtesy	Word	Unsubstantiated	NaN	NaN	Xavier Poveda	20.0	9.0	2003.0
98344	18198	Giovanni	Fini	POM	075 PCT	200915114.0	Abuse of Authority	Premises entered and/or searched	Exonerated	NaN	NaN	Giovanni Fini	26.0	9.0	2009.0

Command column could contain very useful info, especially about precinct.
Documented in CCRB_filespecs.xlsx

CapStat.NYC Police Officer database

- 12,450 current police officers.
- Information about their rank and district
- Not a perfect match with the misconduct database!
    - 52061 of the 323911 complaints matched to a police officer in the CapStat.NYC database.
- Compiled by fellow Princeton student Wendy Ho

CapStat.NYC Police Officer database



	Unnamed: 0	First Name	Last Name	Rank	Location	Full Name
0	0	Lori	Aanonsen	Detective Third Grade	New York	Lori Aanonsen
1	1	Walter	Aanonsen	Lieutenant	New York	Walter Aanonsen
2	2	Abdelhadi	Aanouz	Police Officer	Bronx	Abdelhadi Aanouz
3	3	Gary	Aaronson	Police Officer	Queens	Gary Aaronson
4	4	Jacob	Aaronson	Police Officer	New York	Jacob Aaronson
5	5	Robert	Aasheim	Detective Specialist	New York	Robert Aasheim
6	6	Thomas	Aasheim	Detective Second Grade	New York	Thomas Aasheim
7	7	Darsey	Abad	Detective First Grade	New York	Darsey Abad
8	8	Anthony	Abadia	Police Officer	Kings	Anthony Abadia
9	9	David	Abadia	Police Officer	Bronx	David Abadia

Lots of data! What to predict?

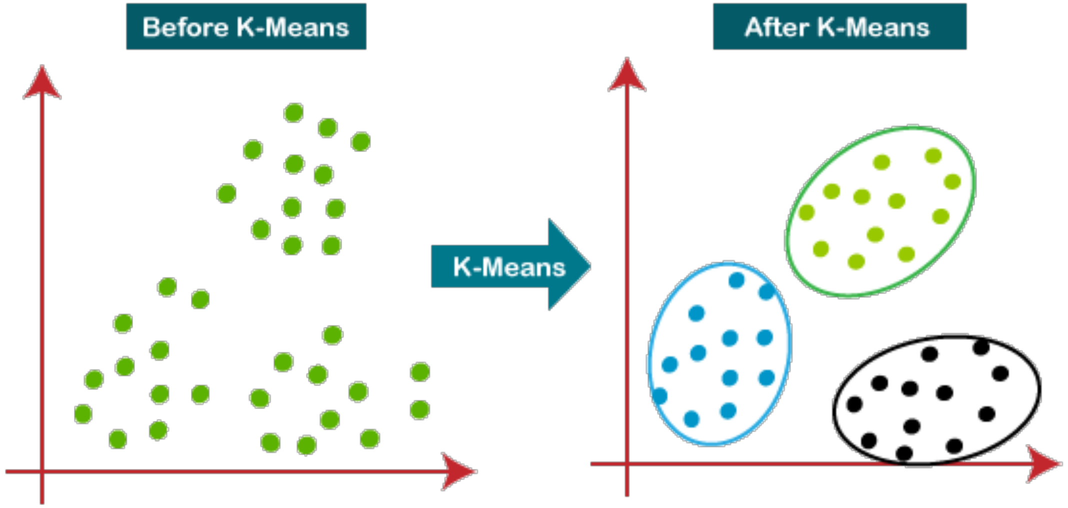
- That's up to you!
- Previous assignments had a specific prediction task.
- This assignment much more open ended.
  - We would like you to use **unsupervised learning** to find patterns in the data.
- Some ideas in assignment description

Unsupervised Learning

- Vague definition.
  - “Find hidden patterns in unlabeled data”
- Very ill-defined problem. In practice refers to a set of commonly used techniques:
- Clustering
  - Find a group of meaningful clusters and assign each data point to a cluster.
- Dimensionality Reduction
  - Find a low-dimensional representation of the data that preserves certain qualities.
- Generative modeling (we won’t need this)
  - Learn the distribution of the data and/or learn to generate new data.

Clustering

- Find archetypes of different groups
- Methods:
  - K-Means
  - Gaussian Mixture Models
  - Nonparametric models

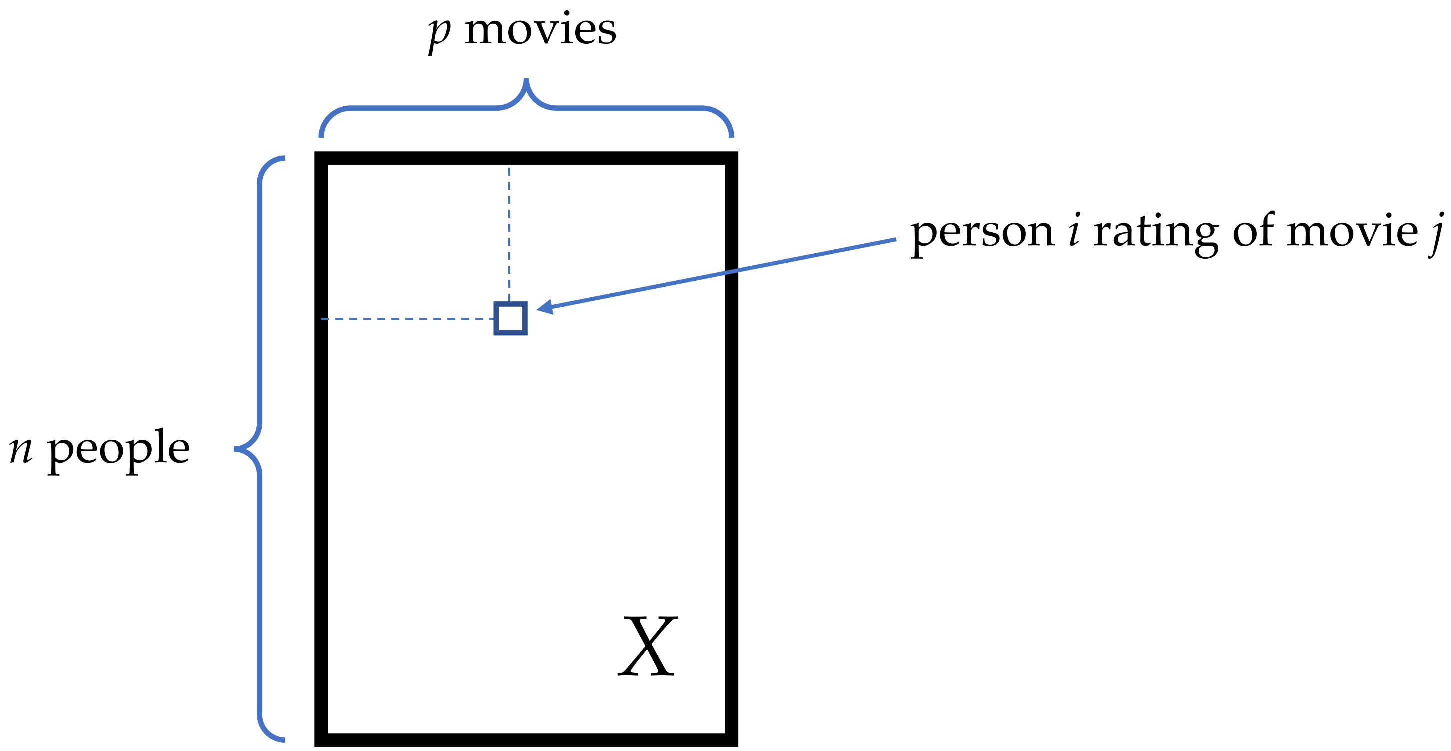


Dimensionality Reduction

- Represent high dimensional data with fewer variables.
  - Lose some information, but hopefully keep the important ones for the downstream task.
- Methods:
  - PCA
  - Factor Analysis
  - NNMF
  - LDA
  - Zero-inflated methods

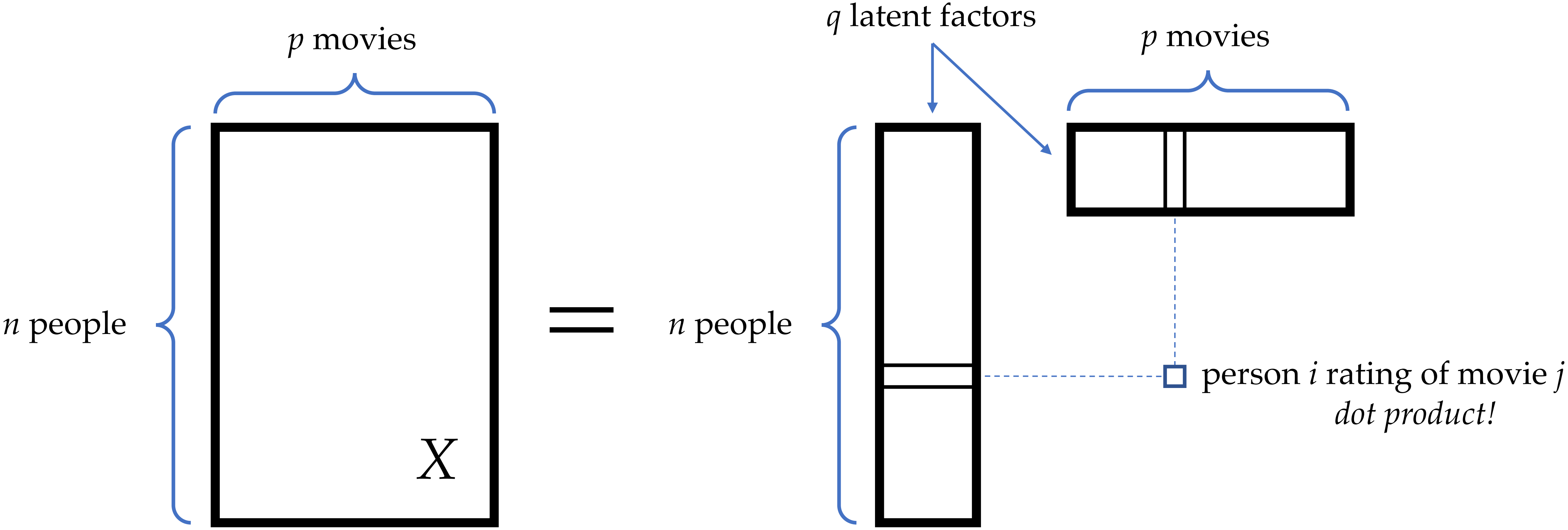
Dimensionality Reduction: Matrix Factorization

- Given n by p data matrix X:



Dimensionality Reduction: Matrix Factorization

- Factor into two components:

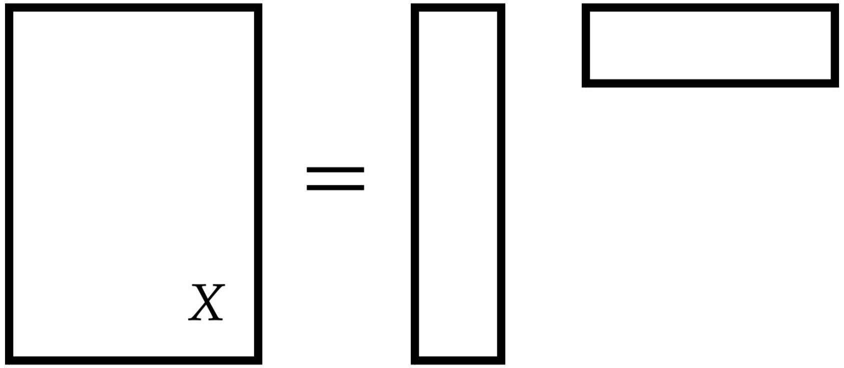


person *i* rating of movie *j*

dot product!

Dimensionality Reduction: Matrix Factorization

- How to factor?
- Note that this is intrinsically an approximation, unless X happens to be low rank (it will not be in real life)
- But oftentimes it is “approximately low rank”
- SVD gives an ”optimal” low-rank matrix factorization.
- Non-negative Matrix Factorization gives all non-negative entries (more interpretable)



Back to HW3

- Two csv files with datasets. One for complaints one for police officers.
- Jupyter notebook with data exploration + simple SVD example.