ENRON & NETFLIX MINING

A COURSE PROJECT OF DATA INTENSIVE PROGRAMMING

PRESENTED BY DA ZHAO AND ZIYU JI
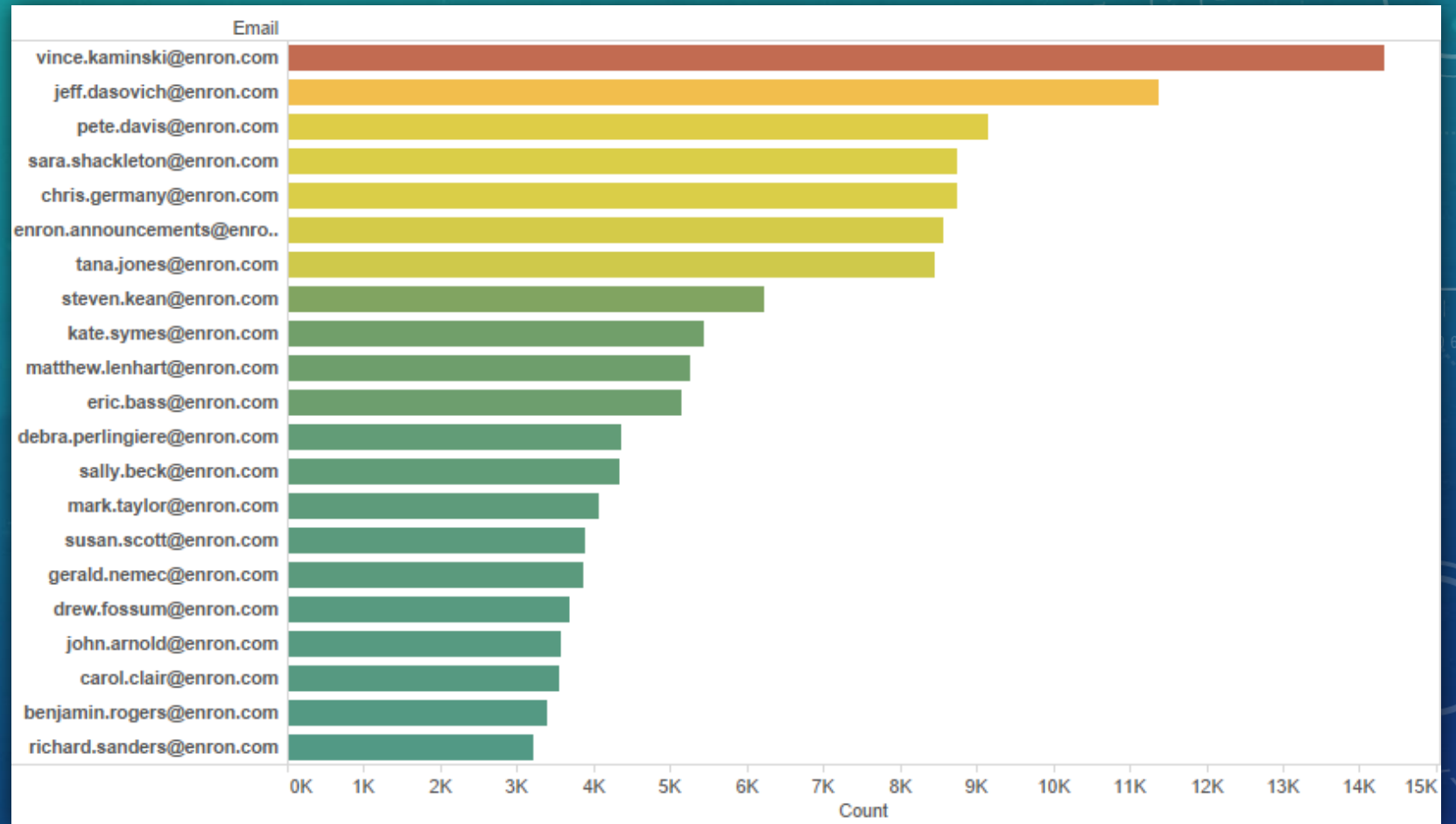
ENRON DATASET

Schema

Emails(eid:varchar, timestamp:varchar, from:varchar, to:varchar, cc:varchar, subject:varchar, context:varchar)

WHO'S WORKAHOLIC

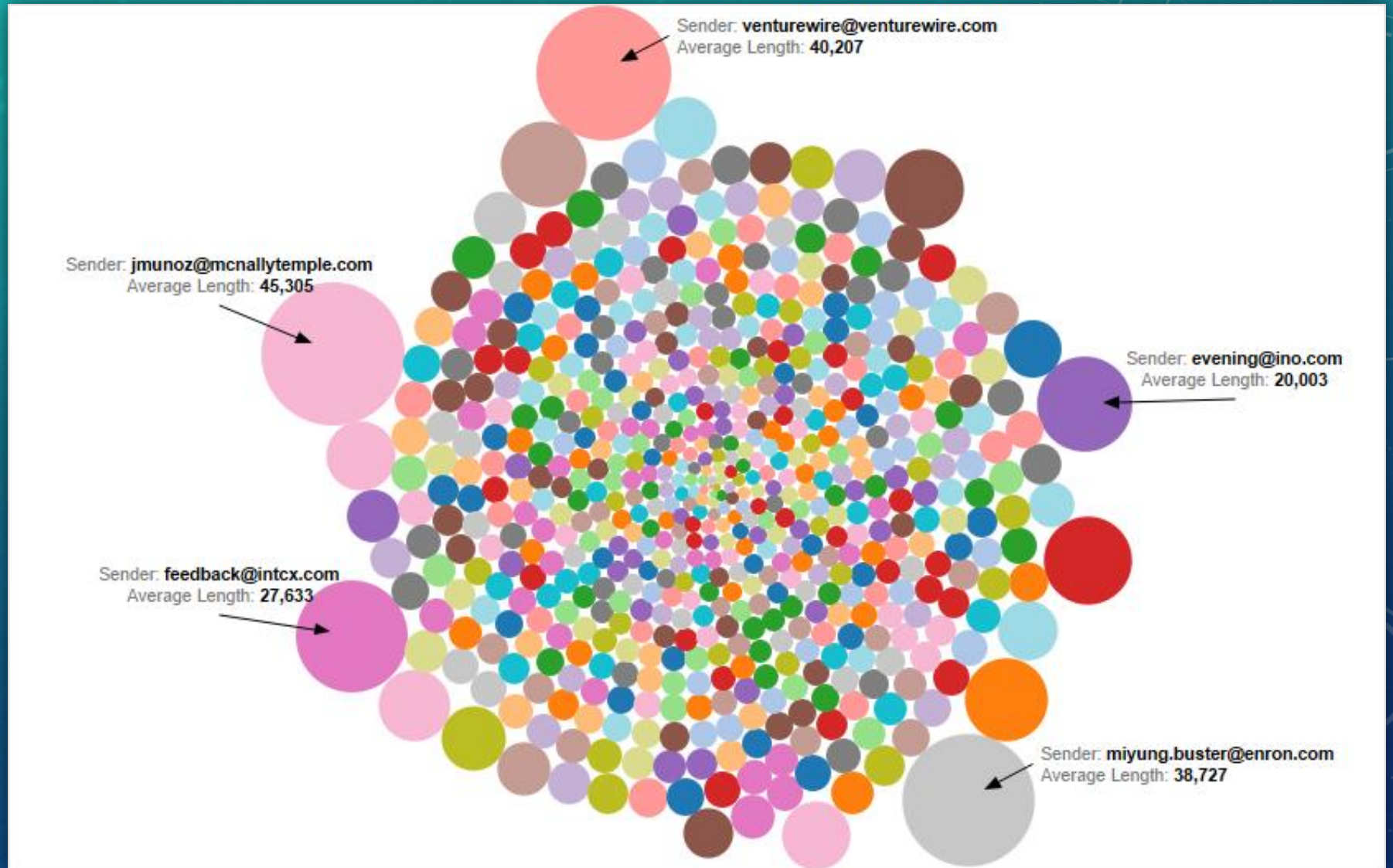- Send emails late in the evening

- Send a lot of emails

WHAT THEY ARE TALKING

High monogram frequency

Removing stop words

WHO TALKS MOST

Average email context length

NETFLIX DATASET

movie titles(mid:integer, yearOfRelease:integer, title:varchar)

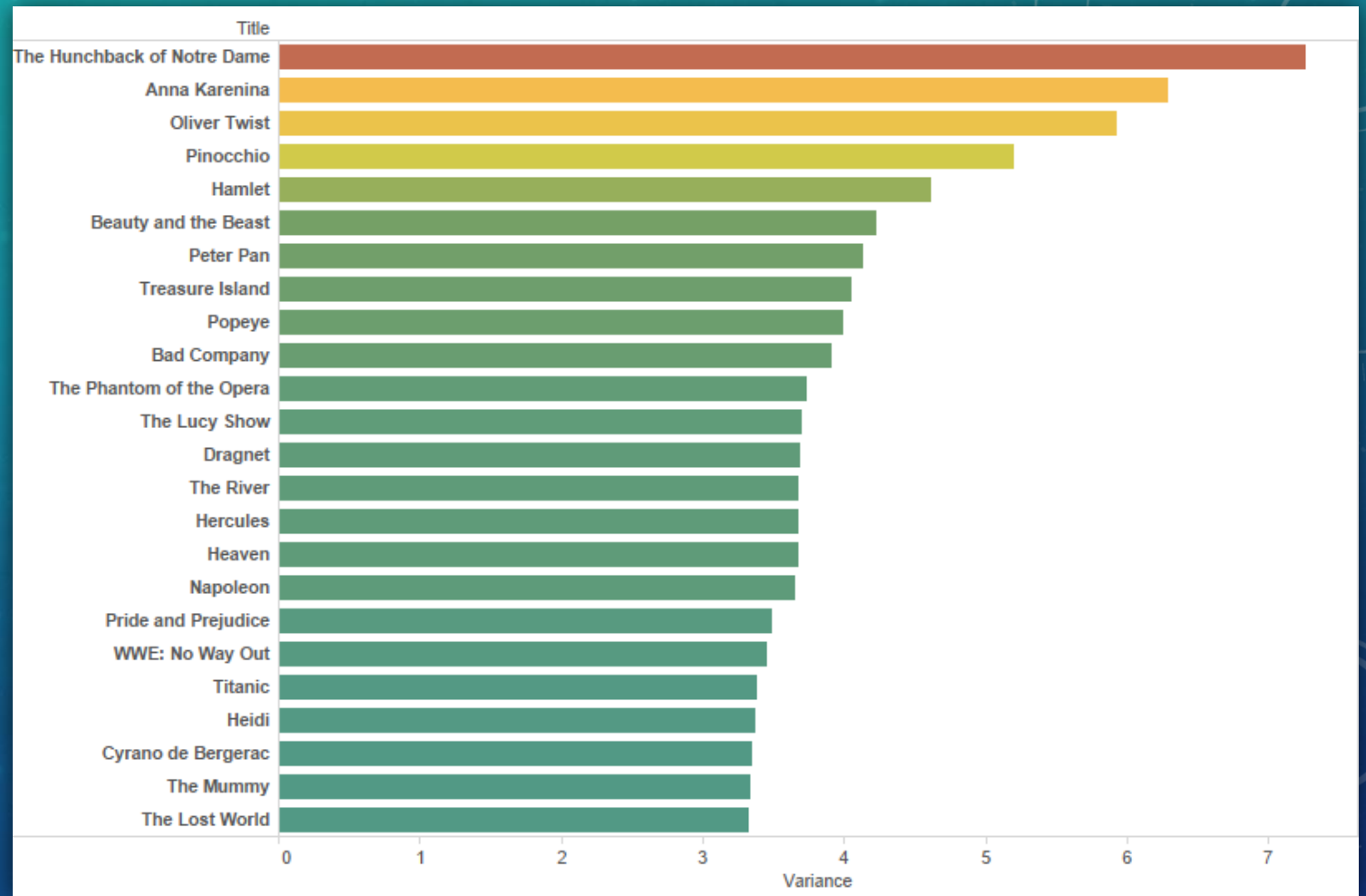movie ratings(mid:integer, custormer id:integer, date:varchar, rating: integer)

TOP 24 WIDELY WATCHED CONTROVERSIAL MOVIES

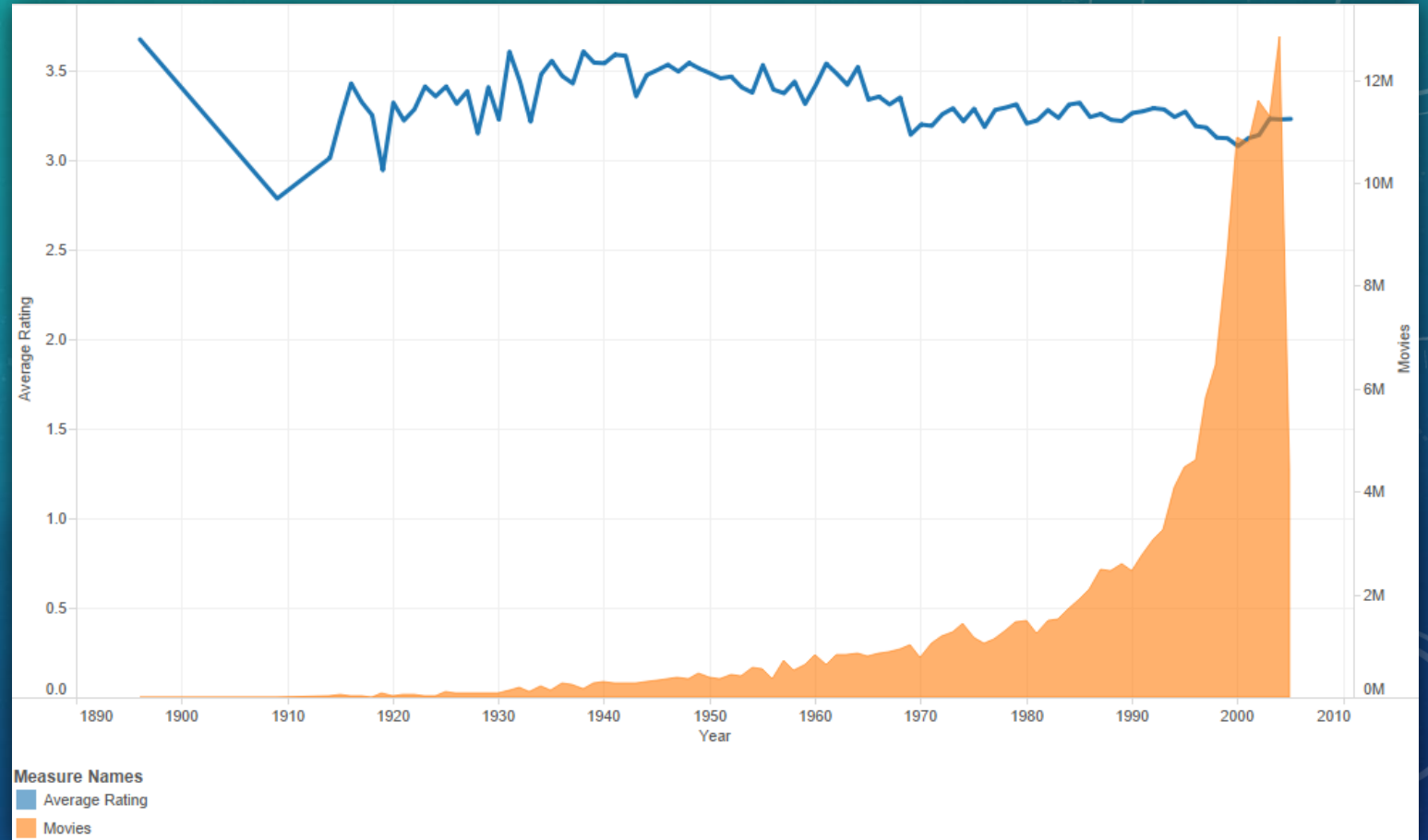Widely watched: over 1000 review

Controversial: big variance on rating

PRODUCTION AND RATINGS TREND OF MOVIES

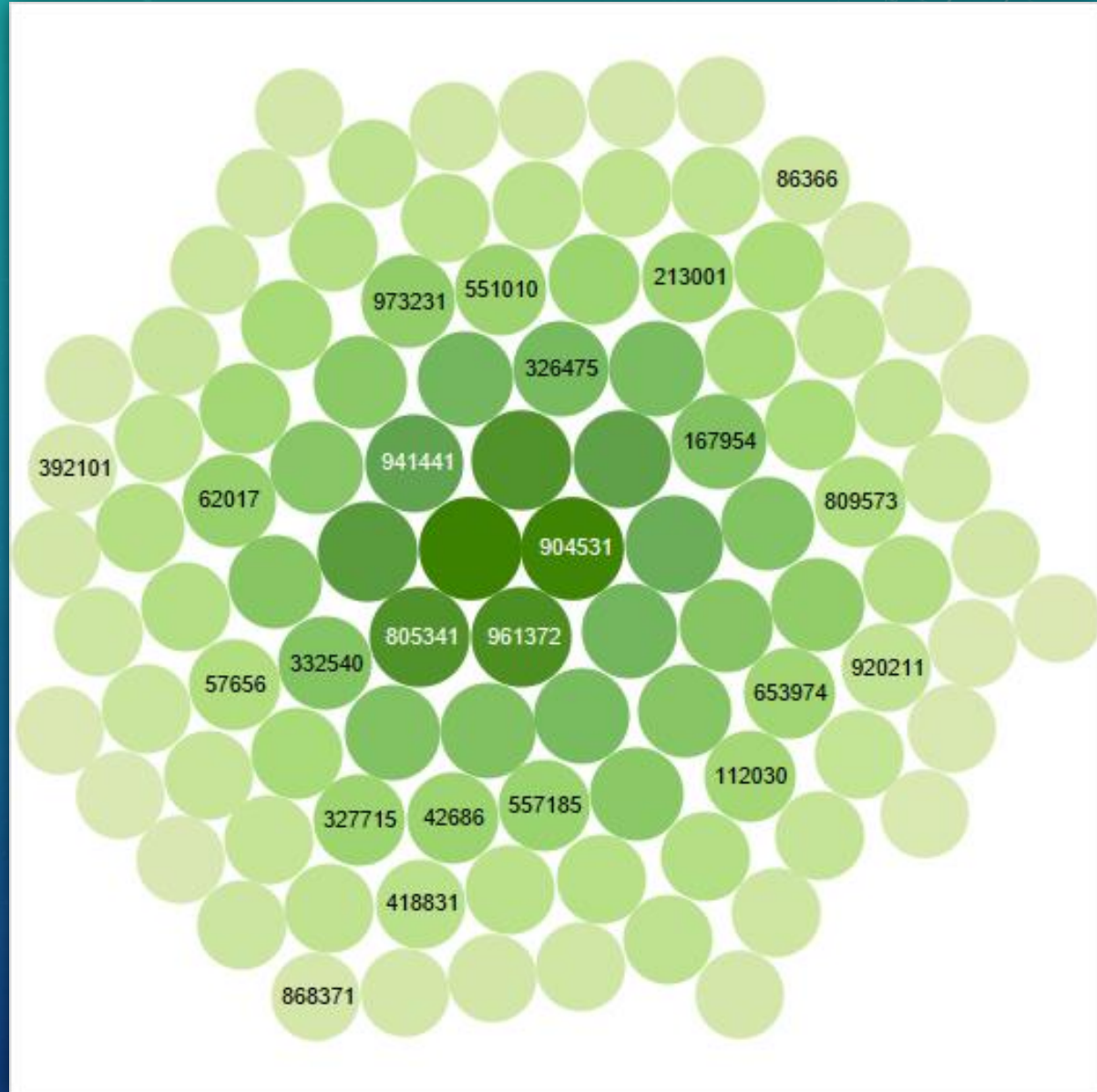Production: sum of movies released each year

Ratings: average ratings of movies each year

REVIEWERS WITH MOST DISTINCTIVE ESTHETIC STANDARDS

Disparity: Difference between personal movie rating with average rating for a movie.

High disparity over all movies.

Thank you!