

INTRODUCTION



To DATA SCIENCE

Introduction to Data Science

Case Study: the Netflix Dataset - Class 4

Giora Simchoni

gsimchoni@gmail.com and add #intro2ds in subject

Stat. and OR Department, TAU

INTRODUCTION



To DATA SCIENCE

The Netflix Prize

- In 2006 Netflix was still a DVD-rental business
- Users would rate the movies from 1 to 5
- Netflix training data: 100,480,507 ratings that 480,189 users gave to 17,770 movies
- Metric on predicted scores \hat{y} vs. true scores y : $RMSE = \sqrt{\frac{1}{n} \sum_i (\hat{y}_i - y_i)^2}$
- Netflix algorithm achieved an RMSE of 0.95
- “Beat our algorithm by over 10% and win \$1 million dollars!”
- Over 20,000 teams, over 150 countries, over 13,000 submissions
- Prize awarded in 2009 for the BellKor’s Pragmatic Chaos team, with an RMSE of 0.855

The Netflix (mini) Dataset



- 10,000 users, 99 movies
- Can you predict the 100th movie's scores “Miss Congeniality”?

Intro to Data Science

INTRODUCTION



To DATA SCIENCE