

# INTRODUCTION TO DATA SCIENCE

This lecture is  
based on course by E. Fox and C. Guestrin, Univ of Washington

22/12/2020

WFAiS UJ, Informatyka Stosowana  
I stopień studiów

# Recommending system: films

2

Machine learning:  
recommending system

## □ Personalizacja

**You Tube**

100 Hours a Minute  
*What do I care about?*

Information overload



Browsing is "history"  
– Need new ways  
to discover content

Personalization: Connects *users & items*

viewers

videos

# Recomending system:

3



Connect users with movies they may want to watch

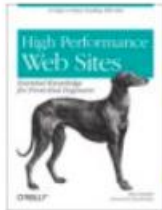
# Recomending system:

4

amazon.com

[Help](#) | [Close window](#)

## Recommended for You



**High Performance Web Sites:  
Essential Knowledge for  
Front-End Engineers**

by Steve Souders (Author)

**Our Price: \$19.79**

**Used & new** from \$16.24

[Add to Cart](#)

[Add to Wish List](#)

## Because you purchased...

**Programming Collective Intelligence: Building  
Smart Web 2.0 Applications** (Paperback)

by Toby Segaran (Author)

### Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)



Book Title	Author	Format	Price	Rating
Even Faster Web Sites: Performance... (Paperback)	Steve Souders	Paperback	\$23.10	★★★★★ (7)
Simply JavaScript (Paperback)	Kevin Yank	Paperback	\$26.37	★★★★★ (19)
The Art & Science of Java (Paperback)		Paperback		★★★★★ (5)

Categories: [Any Category](#) | [Algorithms](#) | [Boxed Sets](#) | [Business & Culture](#) | [Java](#) | [Networking](#) | [Networks, Protocols & APIs](#) | [New](#) | [SQL](#)

Recommendations combine  
global & session interests

# Recommending system: popularity?

5

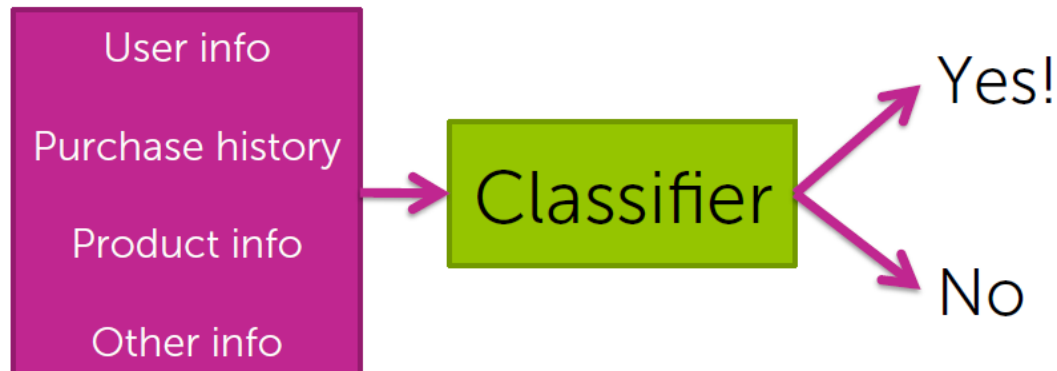
- **Popularity?**
  - ▣ **Ranking vs number of downloading?**
  - ▣ **No personalisation in this case**

# Recommending system: classification

6

## □ Classification?

- ▣ What is probability that I will buy this product?
- ▣ Personalisation: purchase history, monthly and yearly trends, etc.



# Recommending system: correlations

7

- **Analyse correlations. Customers who bought product A also bought product B**
  - ▣ **Correlation matrix**

User  purchased *diapers*

1. Look at *diapers* row of matrix
2. Recommend other items with largest counts
  - *baby wipes, milk, baby food,...*

# Recommending system: correlations

8
















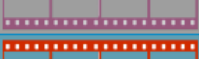


- **Analyse correlations. Customers who bought product A also bought product B**
  - ▣ **Should we normalise correlation matrix?**
  - ▣ **How to quantify that products are „products“?**
- **Limitation of correlations:**
  - ▣ **It is not looking at the purchasing history (trends in time)**
  - ▣ **How to add a new customer (no info on correlations)?**



# Recommending system: films

9

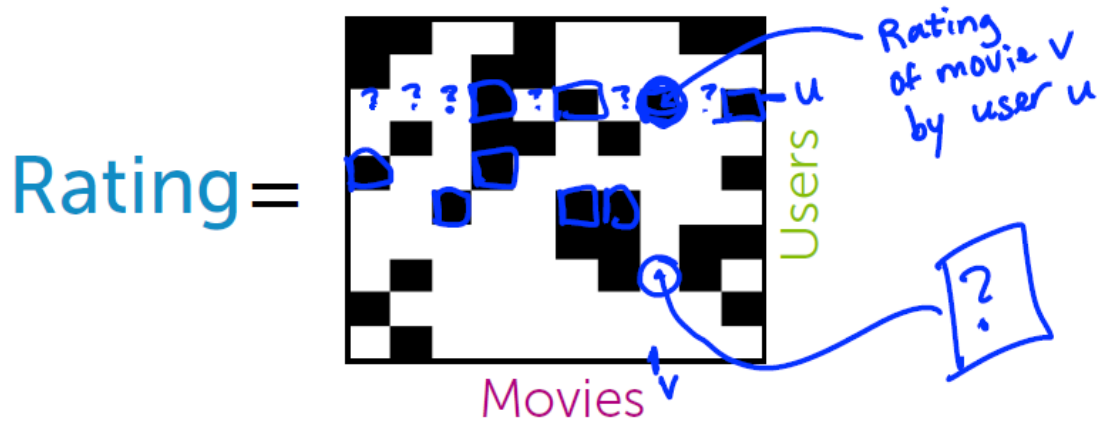
- Users watch movies and rate them

User	Movie	Rating
		★★★★☆
		★★★★★
		★★★☆☆
		★★★☆☆
		★★★★☆
		★★★☆☆
		★★★★☆
		★★★★★
		★★★★☆

Each user only watches a few of the available movies

# Recommending system: films

10



- **Data:** Users score some movies

$Rating(u,v)$  known for black cells  
 $Rating(u,v)$  unknown for white cells

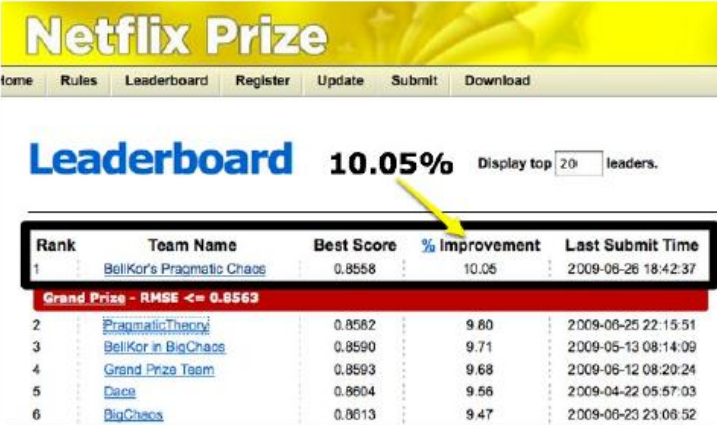
- **Goal:** Filling missing data?



# Recommending system: optimisation

11

- Squeezing last bit of accuracy by blending models
- Netflix Prize 2006-2009
  - 100M ratings
  - 17,770 movies
  - 480,189 users
  - Predict 3 million ratings to highest accuracy
  - **Winning team blended over 100 models**



**Netflix Prize**

Home Rules Leaderboard Register Update Submit Download

**Leaderboard** 10.05% Display top 20 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	<a href="#">BellKor's Pragmatic Chaos</a>	0.8558	10.05	2009-06-26 18:42:37
<b>Grand Prize - RMSE &lt;= 0.8563</b>				
2	<a href="#">PragmaticTheory</a>	0.8562	9.80	2009-06-25 22:15:51
3	<a href="#">BellKor in BigChaos</a>	0.8590	9.71	2009-05-13 08:14:09
4	<a href="#">Grand Prize Team</a>	0.8593	9.68	2009-06-12 08:20:24
5	<a href="#">Dace</a>	0.8604	9.56	2009-04-22 05:57:03
6	<a href="#">BigChaos</a>	0.8613	9.47	2009-06-23 23:06:52

# Recommending system: how effective?

12

## The world of all baby products



# Recommending system: how effective?

13

## User likes subset of items





# Recommending system: how effective?

14

## How many liked items were recommended?

The image displays a variety of baby products. Items circled in blue include a wooden rocking chair, a baby monitor, a car seat, a hanging mobile, and a pair of baby shoes. Items crossed out with blue X's include a crib, a pair of baby shoes, a set of baby bottles, and a baby bottle. Items enclosed in pink boxes include a baby stroller, a baby monitor, a box of Kirkland Baby Wipes, a baby stroller, a baby bottle, and two rubber ducks. A purple stick figure stands in the center of the collection.

**Recall**  
$$\frac{\# \text{ liked \& shown}}{\# \text{ liked}} = \frac{3}{5}$$

# Recommending system: how effective?

15

## How many recommended items were liked?

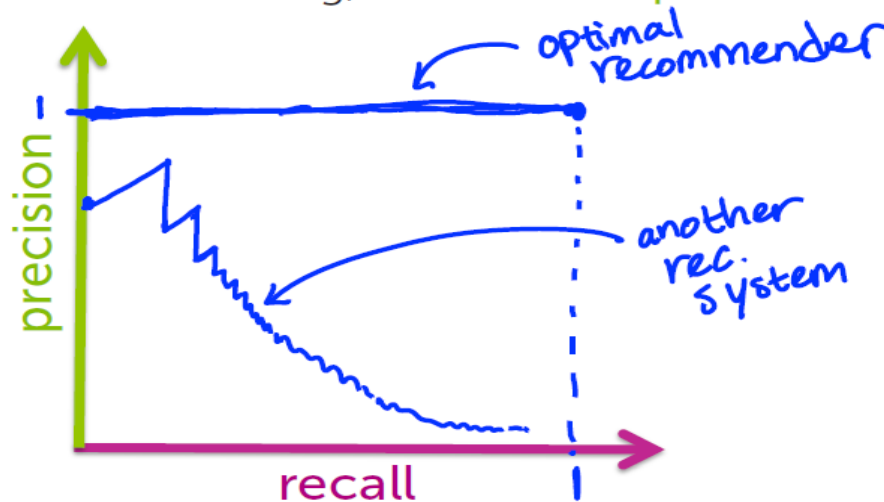
**Precision**  
$$\frac{\# \text{ liked \& shown}}{\# \text{ shown}}$$
  
$$= \frac{3}{11}$$

# Recommending system: how effective?

16

## Precision-recall curve

- **Input:** A specific recommender system
- **Output:** Algorithm-specific precision-recall curve
- To draw curve, vary threshold on # items recommended
  - For each setting, calculate the precision and recall



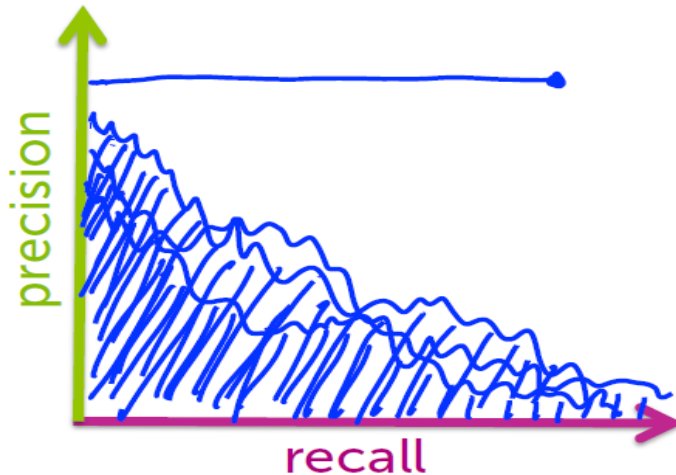


# Recommending system: how effective?

17

## Which Algorithm is Best?

- For a given **precision**, want **recall** as large as possible (or vice versa)
- One metric: largest **area under the curve (AUC)** ★
- Another: set desired recall and maximize precision (precision at k)



# Recommending system

18

## Models

- Collaborative filtering
- Matrix factorization
- PCA

## Algorithms

- Coordinate descent
- Eigen decomposition
- SVD

## Concepts

- Matrix completion, eigenvalues, random projections, cold-start problem, diversity, scaling up