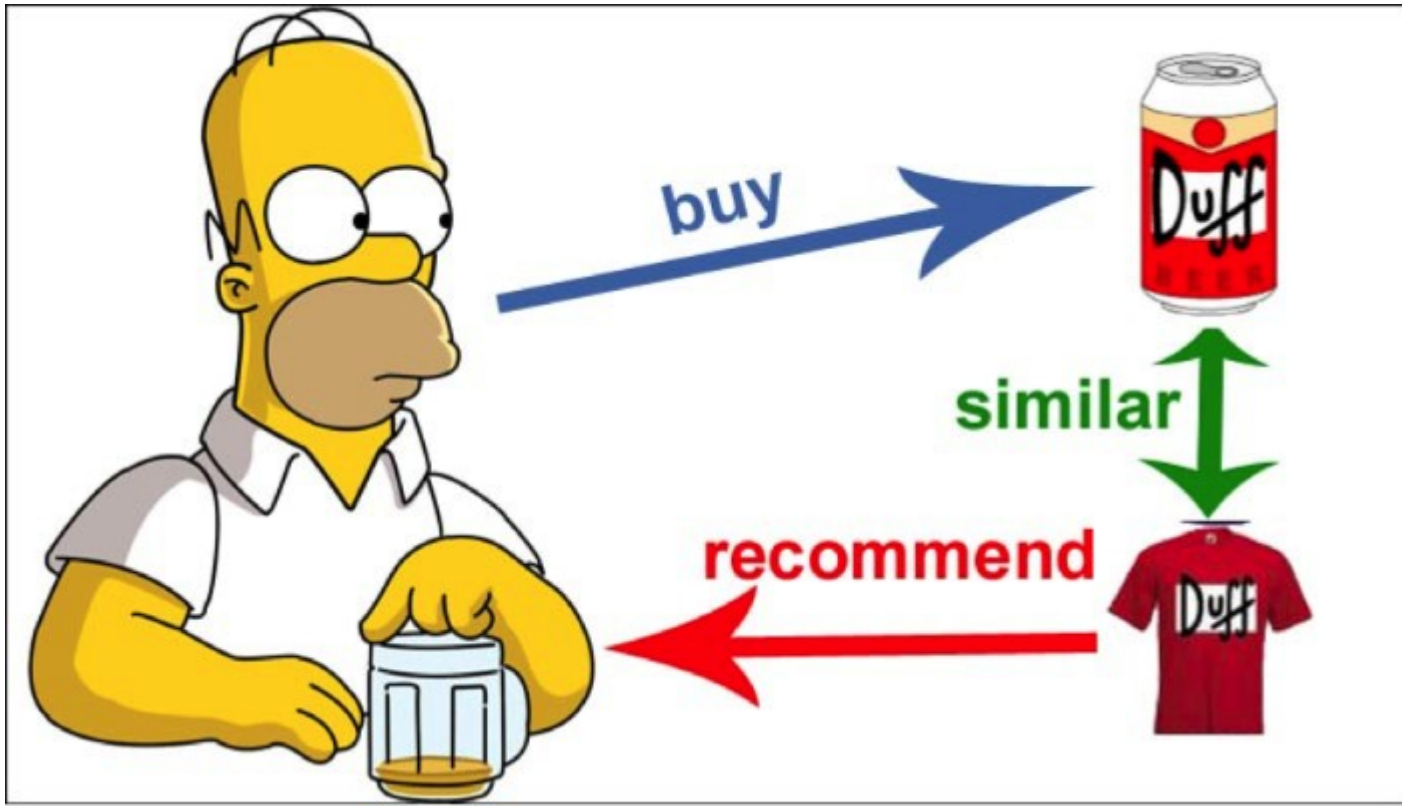
Collaborative filtering

A motivating application

In this lecture

- Recommender systems only as problem drivers
  - More about advanced collaborative filtering further in this course
- This lecture
  - Notions and algorithmic challenges posed by recommender systems
  - Representative of a vast range of applications

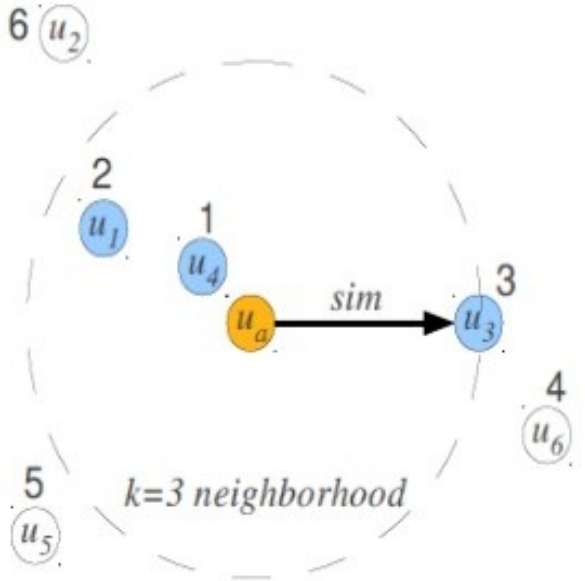
Collaborative filtering in a picture (by LinkedIn) ...



… and a nutshell (by Navisro Analytics)

User-based CF

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	2.0	1.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	3.0	?	?	3.0	2.0	2.0	?	3.0
u_4	4.0	?	?	2.0	1.0	1.0	2.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	1.0	1.0	?	1.0
u_a	?	?	4.0	3.0	?	1.0	?	5.0
r_a	3.5	4.0			1.3		2.0	



$$\hat{r}_{aj} = \frac{1}{\sum_{i \in \mathcal{N}(a)} s_{ai}} \sum_{i \in \mathcal{N}(a)} s_{ai} r_{ij}$$

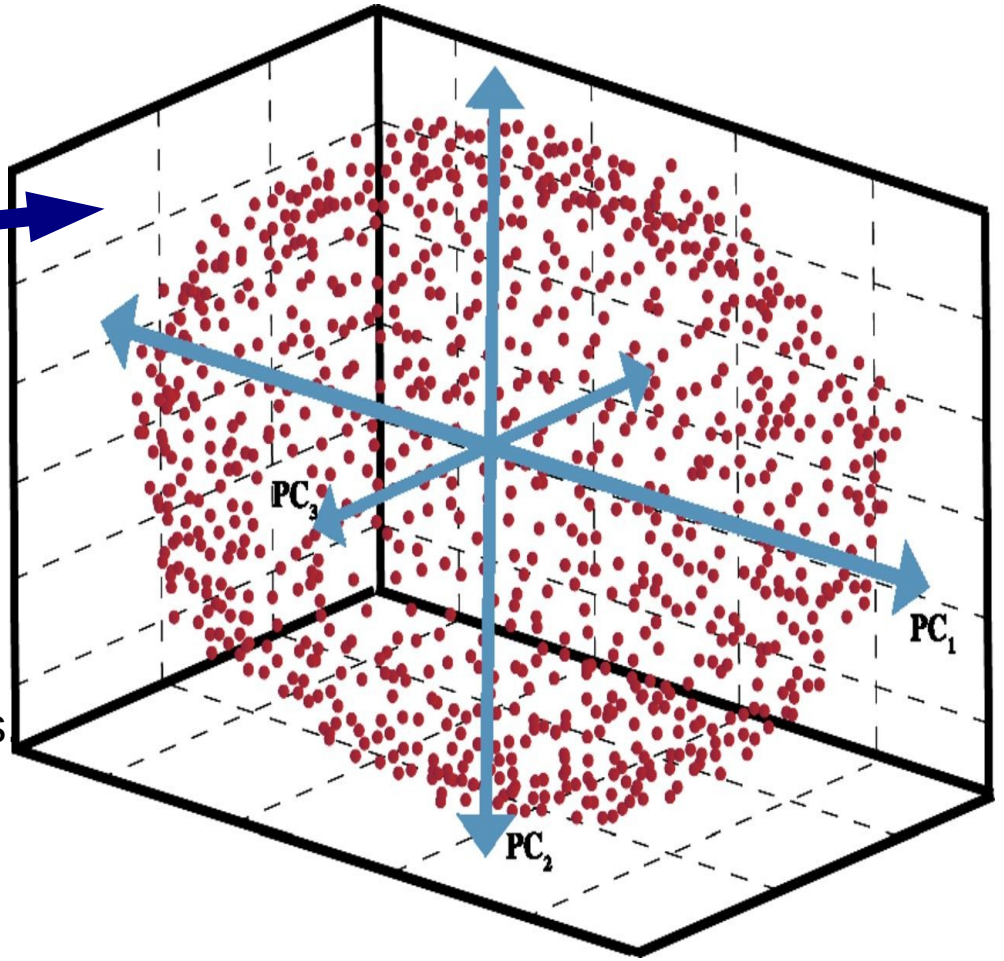
$$\text{sim}_{\text{Cosine}}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Reference: Recommenderlab vignette, http://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf

Everything is a point

Feature vector is a point in a high-dimensional space

user \ item	A	B	C	D	. . .
Joe	2	-2	2	-2	
Sue	1	-1	1	-1	
John	0	0	0	0	



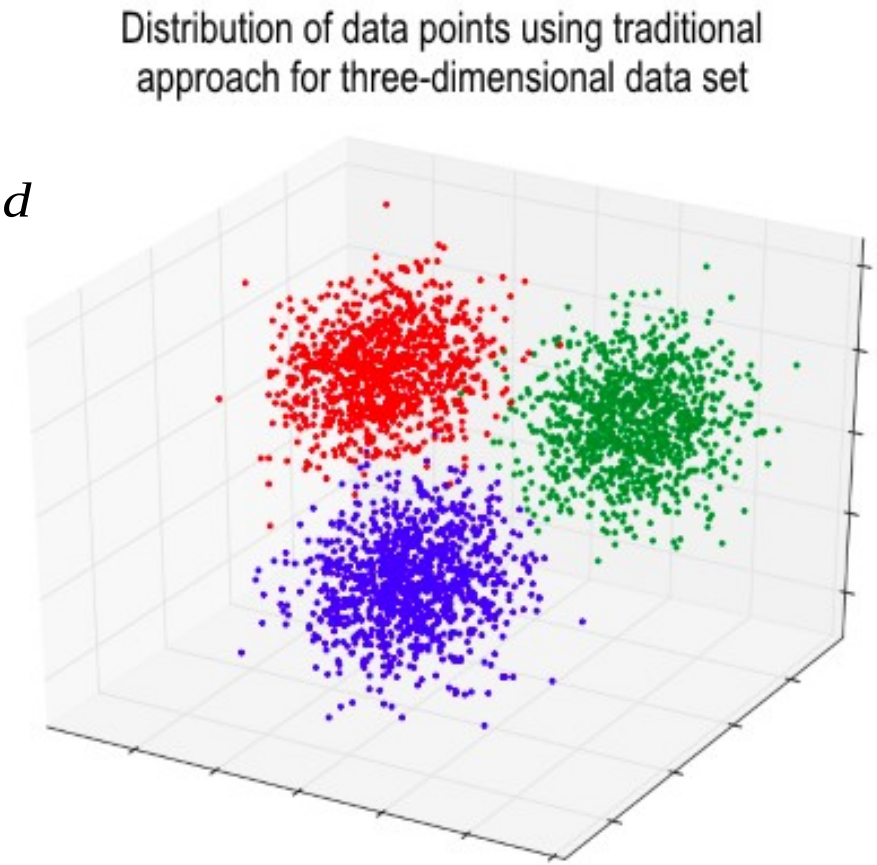
Number of dimensions = number of features
In the right picture: dimensions are the (3) users

Two general problems

- Classification
  - Aggregate or (cluster) users/items according to **similarity**
- Prediction: given new user:
  - Assign user to closest cluster/group wrt training set
  - Predict user's behaviour/preferences based on average behaviour/preferences of assigned group

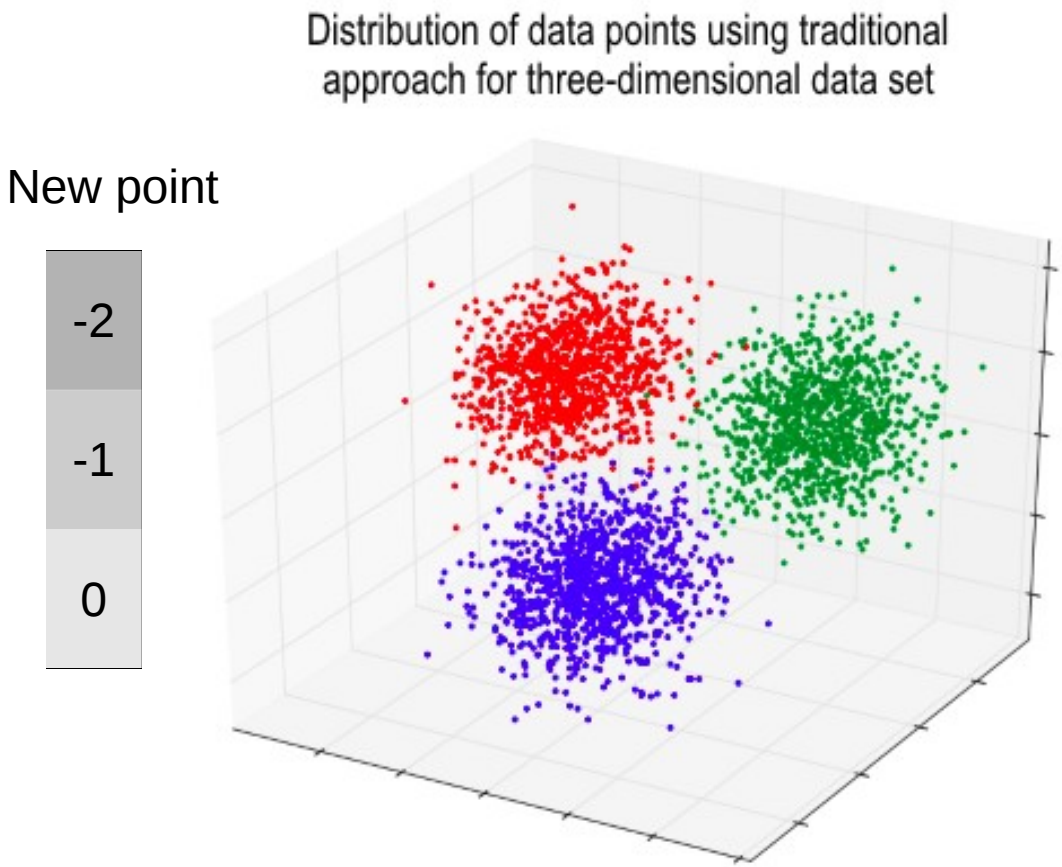
Classification

- Aggregate points into k groups based on similarity
- Challenges
  - Similarity → distance in $\mathbb{R}^d$
  - Similarity estimation
  - Clustering / classification
  - Computation
    - Potentially expensive
    - Distribute when possible
- Real time/off-line
  - Often performed off-line



Prediction

- New user/item u
  - A point in $\mathbb{R}^d$
- Find “closest” group to u
- Predict behaviour based on assigned group's preferences
- Challenges
  - Requires distance
  - Often performed on line



Immediately related topics

- Distributed/cluster computing
  - MapReduce, Spark, Giraffe …
- Similarity estimation
  - Locality Sensitive Hashing, MinHashing …
- Advanced topics in recommender systems
  - Beyond collaborative filtering: Netflix challenge, latent factors and decompositions
- Dimensionality reduction
- Clustering