# Explicit data

- ✘ Buys a product, rates a film, or gives a thumbs up or down to a post.
- ✘ Clearly showing how customers feel about a product.
- ✘ Clear, unambiguous, and gives us a definite picture of the user.

# Explicit data Issues

- ✘ Harder to collect.
  - People seldom leave a rating
- ✘ Shallow.
  - People tend to rate in the extremes

# Application

- ✘ Have more obvious value than implicit data, therefore will naturally be favorable.
- ✘ A single click tells us that a user liked a video or rated a product positively

# Implicit Data

- ✘ Passively track different sorts of user behaviors
- ✘ No direct input from the users regarding preferences
- ✘ Purchase history, watching habits, browsing history...

# Implicit data in spark als

1.

$$p_{ui} = \begin{cases} 1 & r_{ui} > 0 \\ 0 & r_{ui} = 0 \end{cases}$$

2.

$$c_{ui} = 1 + \alpha r_{ui}$$

r_ui: implicit score
P_ui: a binary variable indicating whether user u likes item i
C_ui: a confidence score indicating the confidence of observing P_ui
Alpha: hyperparameter to be tuned


# Implicit data in spark als

Loss function:

$$\min_{x_\star, y_\star} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda \left( \sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$


# Differences in spark als

*implicitPrefs* : specifies whether to use the *explicit feedback* ALS variant or one adapted for *implicit feedback* data (defaults to false meaning *explicit feedback*)

Reason for default to explicit:
The standard approach to matrix factorization based collaborative filtering treats the entries in the user-item matrix as explicit preferences given by the user to the item.

# Deal with missing data

Difference lies in how we deal with all the missing data in our very sparse matrix. For explicit data we treat them as just unknown fields that we should assign some predicted rating to.
But for implicitly we can't just assume the same since there is information in these unknown values as well. As stated before we don't know if a missing value means the user disliked something, or if it means they love it but just don't know about it. Basically we need some way to learn from the missing data. So we'll need a different approach to get us there.