

Evaluation Challenge

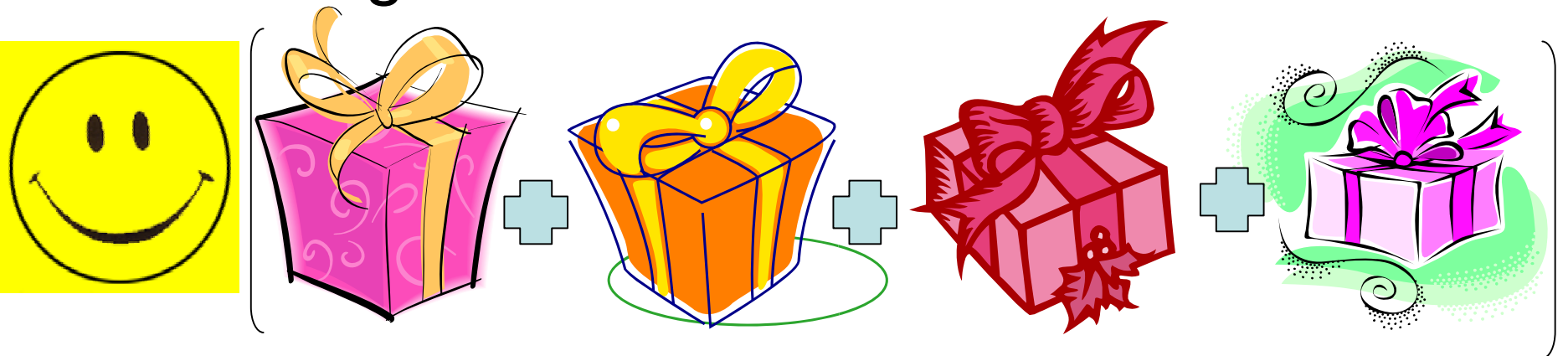
Introduction

Motivation

- Provide a case study to show difficulties in evaluation
- Get more interest in evaluation
- Provide a low threshold way of contributing
- Elicit innovative evaluation designs
- Encourage controversial discussion

System to be evaluated

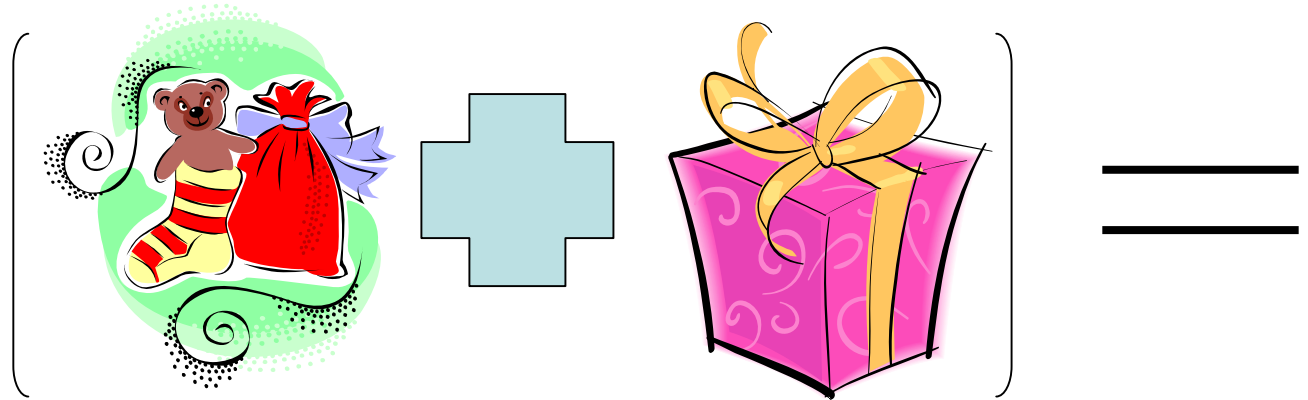
- Recommender system
- Recommends music clips to groups of users
- Ratings 1 to 10 for each music clip
- Models happiness of individuals and uses this to guide recommendations



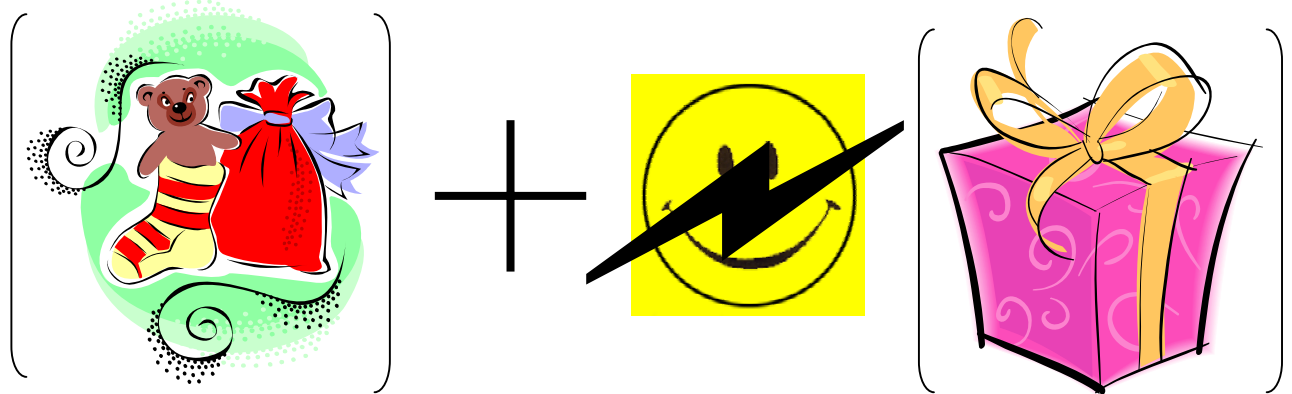
Model 1:



Satisfaction decreases over time



δ_x



$0 \leq \delta \leq 1$

$\delta=0$: No memory

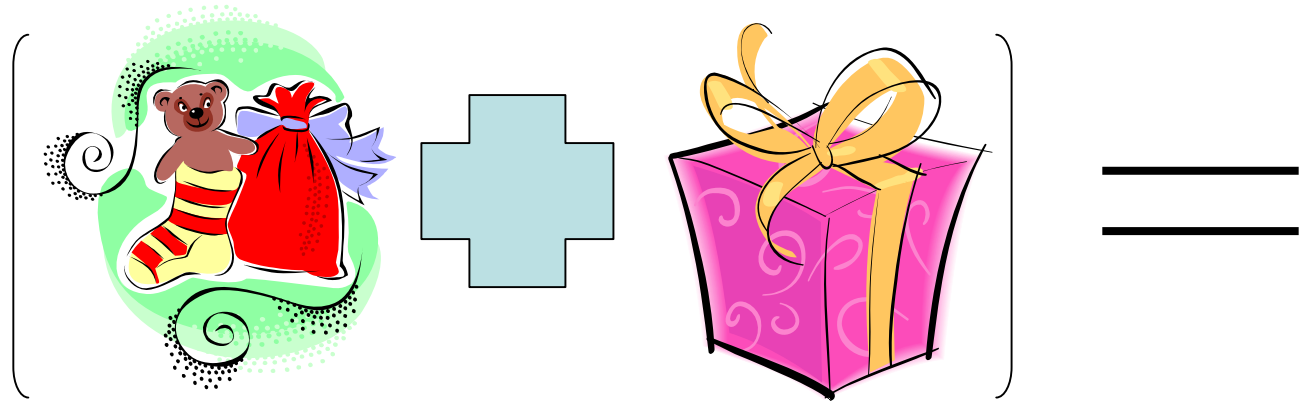
$\delta=1$: Perfect memory

Impact



Quadratic(
Rebalanced(
Rating()))

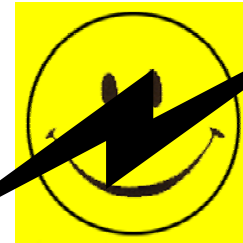
Model 2: Happiness is bounded



δ_x

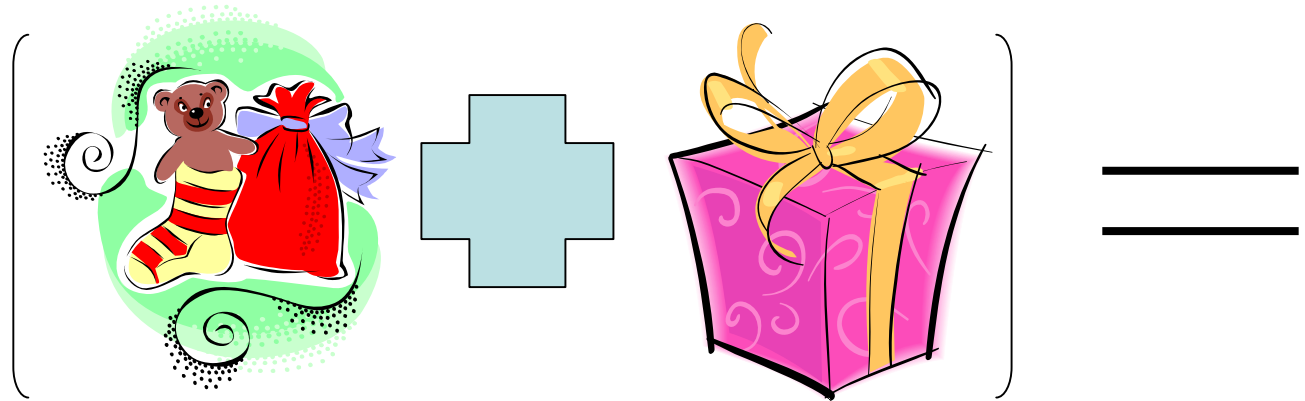


+



$(1+\delta)$

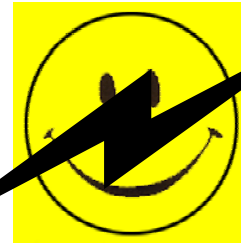
Model 3: Impact depends on mood



δ_x



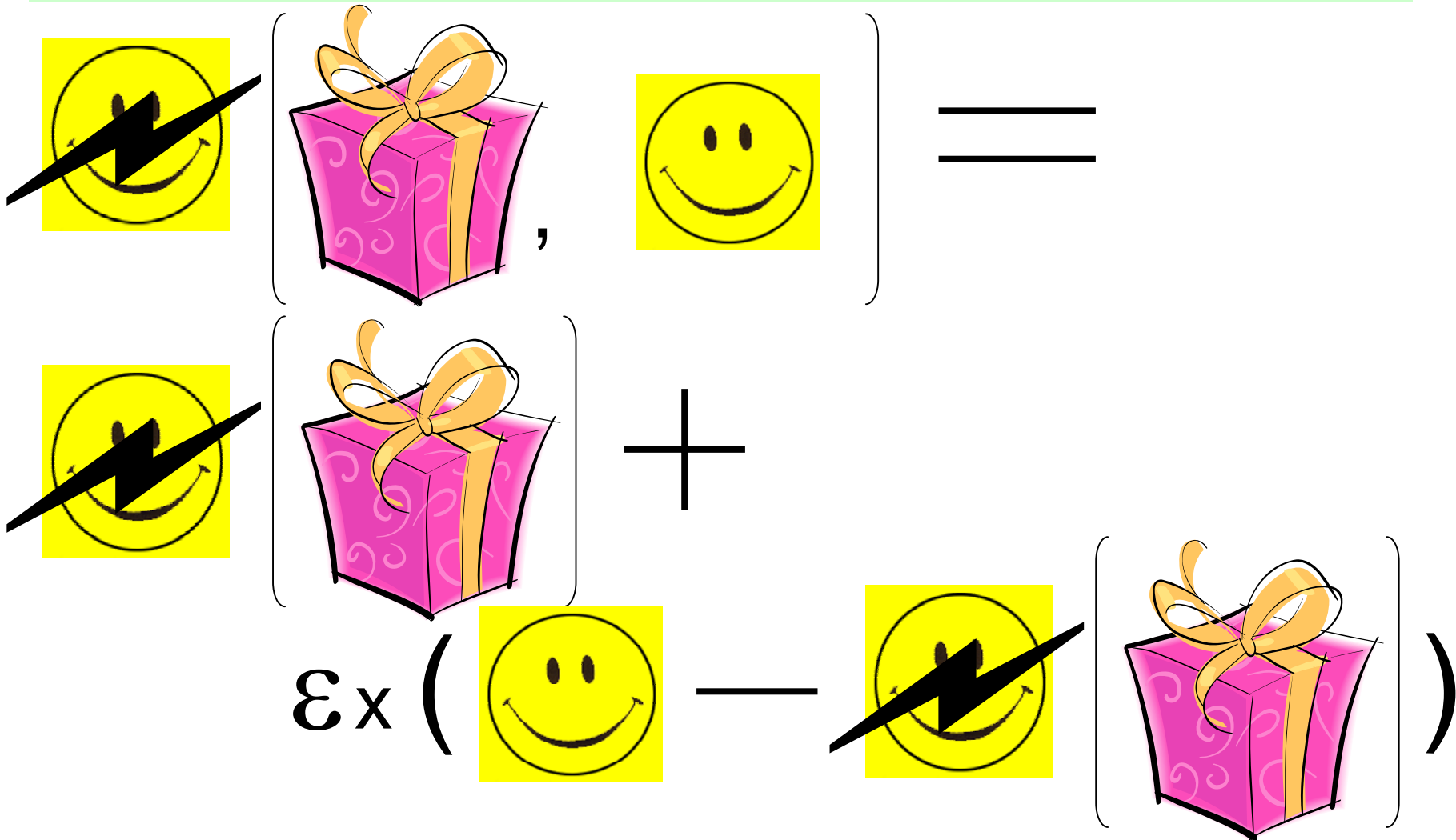
+



δ_x



Impact depends on mood



$0 \leq \varepsilon \leq 1$

$\varepsilon=0$: No impact mood

$\varepsilon=1$: Mood determines all

Evaluation goals

Which of the three models:

- Accurately predicts when an individual is happy versus unhappy?
- Is best at predicting inter-individual differences in happiness?
- Achieves the highest modelling precision: accurately predicts how happy a user will be?

Structure of session

Intention:

- Presentations of the best contributions
- Discuss their merits and limitations
- Vote on winner

However, we only received two contributions....

Plan:

- Presentations of both contributions
- Decide and discuss the main issues for evaluation

Inherent complexity of evaluation

- Need accurate ratings for individuals which is hard due to order effect
- Need to know satisfaction of subjects after each item
- Items should be topically unrelated and preferably not have emotional content