

# Evaluating Recommender Systems

## Topic 8

# Agenda



- General research considerations
- Evaluations on historical datasets
- Analysis of results
- Alternative research designs

# Introduction



- Initially most recommenders have been evaluated and ranked on their prediction power – their ability to accurately predict the user's choices.
- In addition to prediction accuracy, recommender systems can be evaluated to understand how well the recommender achieves its overall goals.
- Recommender systems require that users interact with computer systems as well as with other users. Therefore, a user study can be conducted to understand
  - ▣ Do users find interactions with a recommender system useful?
  - ▣ Are they satisfied with the quality of the recommendations they receive?
  - ▣ What drives people to contribute knowledge such as ratings and comments?
  - ▣ What is it exactly that users like about receiving recommendation?

# General Research Considerations

- Subject
  - ▣ Online customers, students, historical user sessions, simulated users
- Research Method
  - ▣ Experimental, quasi-experimental
- Settings
  - ▣ Lab, Field
  - ▣ User study, Off-line

# General Research Considerations

- Types of Errors
  - ▣ Type I error vs. Type II error
  - ▣ Precision, Recall, F-measure
  - ▣ Accuracy, Error rate

		Proposed by recommender: 	
		Yes	No
 Liked by user:	Yes	Correct predictions	False negatives
	No	False positives	Correct omissions

# General Research Considerations

## □ Hypothesis

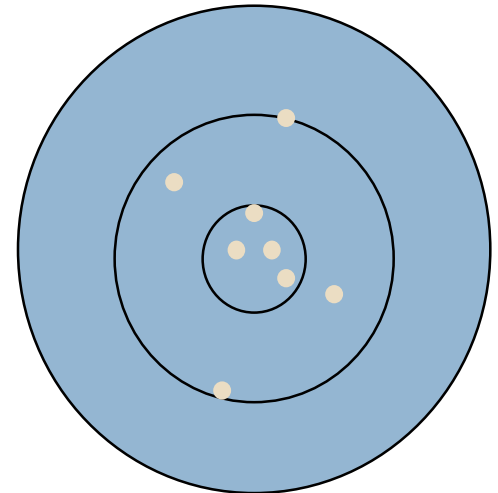
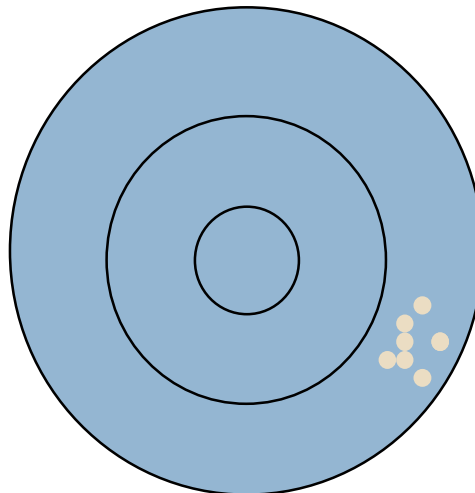
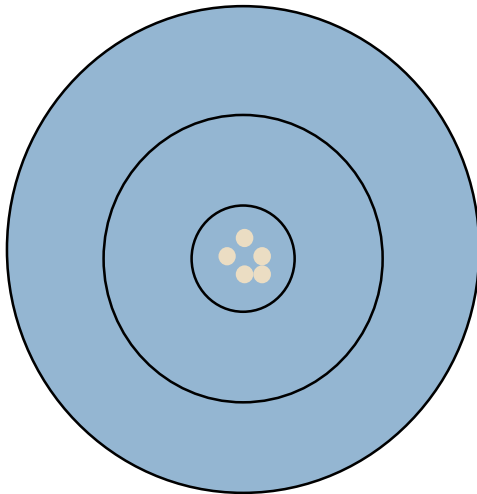
- ▣ Before running the experiment, we must form a hypothesis, a proposed explanation for a phenomenon. A study is conducted to test the hypothesis.
- ▣ Null hypothesis vs. Alternative hypothesis
- ▣ Reject/Fail to reject a hypothesis

	<b><math>H_0</math> is valid: Innocent</b>	<b><math>H_0</math> is invalid: Guilty</b>
<b>Reject <math>H_0</math> I think he is guilty!</b>	Type I error False positive Convicted!	Correct outcome Negative Convicted!
<b>Don't reject <math>H_0</math> I think he is innocent!</b>	Correct outcome Positive Freed!	Type II error False negative Freed!

# General Research Considerations

## □ Reliability and Validity

- Reliability: Absence of inconsistencies and errors in the data and measurement
- Validity
  - Measurement validity: Measuring what the items supposed to measure
  - Internal validity: Effects observed are due to the controlled test conditions (treatments)
  - External validity: Generalizability of the findings



# Subjects

- People are typically the subjects – online customers, web users, system users, students, or general population
- User profiles containing preference information such as ratings, purchase transactions, or click-through data can be split into training and testing partitions for the evaluation of the recommendation algorithm
- Synthetic datasets should be used only to test recommendation methods for obvious flaws or to measure technical performance criteria such as average computation times
- Natural datasets include historical interaction records of real users
  - ▣ Explicit user ratings
  - ▣ Implicit user feedback such as purchases or add-to-basket actions
  - ▣ Sparsity

$$sparsity = 1 - \frac{|R|}{|I| \cdot |U|}$$



# Popular Data Sets

Table 7.2. *Popular data sets.*

Name	Domain	Users	Items	Ratings	Sparsity
BX	Books	278,858	271,379	1,149,780	0.9999
EachMovie	Movies	72,916	1,628	2,811,983	0.9763
Entree	Restaurants	50,672	4,160	N/A	N/A
Jester	Jokes	73,421	101	4.1M	0.4471
MovieLens 100K	Movies	967	4,700	100K	0.978
MovieLens 1M	Movies	6,040	3,900	1M	0.9575
MovieLens 10M	Movies	71,567	10,681	10M	0.9869
Netflix	Movies	480K	18K	100M	0.9999
Ta-Feng	Retail	32,266	N/A	800K	N/A

# Research Methods

- Variables
  - ▣ Dependent vs. independent
  - ▣ Control variables
- Measures
  - ▣ Pretest and posttest
  - ▣ Dependent variables are measured before and after the treatment
- Settings
  - ▣ Random assign is important
  - ▣ Quasi-experimental design lacks random assignment
  - ▣ Control group should be included
  - ▣ Sample size
  - ▣ Quantitative research vs. Qualitative research
  - ▣ Cross sectional vs. Longitudinal
  - ▣ Lab studies vs. field studies

# Example of Experimental Design

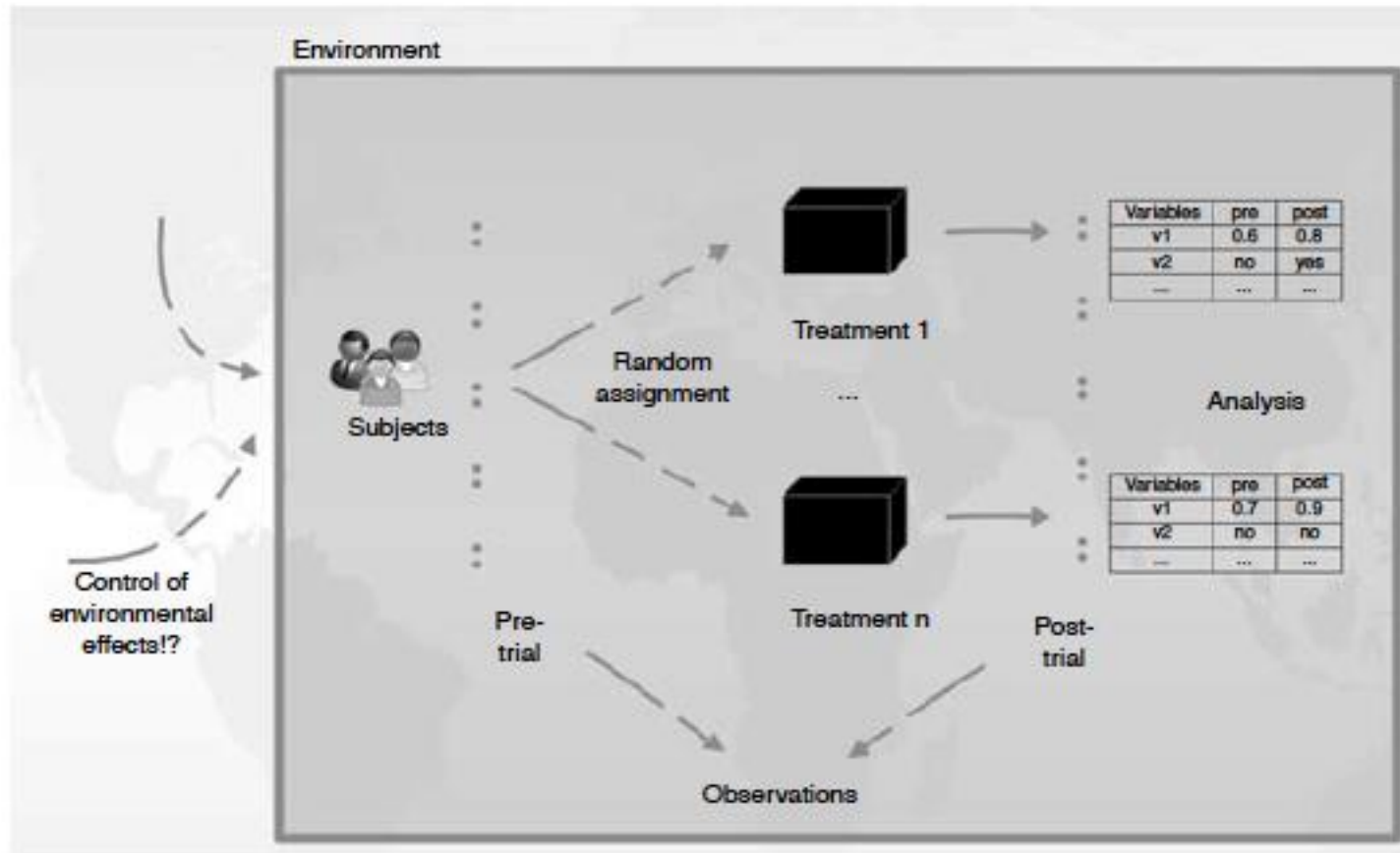


Figure 7.2. Example of experiment design.

# Evaluation on Historical Cases

## □ Methodology

### ▣ A group of user profiles

- For training vs. testing/validation purposes
- Random split, model building, and evaluation
- N-fold cross-validation
  - Original sample is partitioned into N subsamples. Then, N-1 subsamples are used for training; one subsample for evaluation
- Leave-one-out cross-validation
  - Using a single observation in each subsample as the validation data, where N is the total number of user profiles
- All but N – assigns a fixed number N to the testing set of each evaluated user vs. Given N – sets the size of the training partition to N elements
- Prediction task – computes a missing rating vs. classification task – selects a ranked list of n items (i.e., the recommendation set)

# Evaluation on Historical Cases

## □ Metrics

### ▣ Mean absolute error (MAE)

$$MAE = \frac{\sum_{u \in U} \sum_{i \in testset_u} |rec(u, i) - r_{u,i}|}{\sum_{u \in U} |testset_u|} \quad NMAE = \frac{MAE}{r_{max} - r_{min}}$$

### ▣ Accuracy of classifications

$$P_u = \frac{|hits_u|}{|recset_u|} \quad R_u = \frac{|hits_u|}{|testset_u|} \quad F1 = \frac{2 \cdot P \cdot R}{P + R}$$

### ▣ Accuracy of ranks

$$rankscore_u = \sum_{i \in hits_u} \frac{1}{2^{\frac{rank(i)-1}{\alpha}}} \quad rankscore'_u = \frac{rankscore_u}{rankscore_u^{max}}$$
$$rankscore_u^{max} = \sum_{i \in testset_u} \frac{1}{2^{\frac{idx(i)-1}{\alpha}}} \quad liftindex_u = \begin{cases} \frac{1 \cdot S_{1,u} + 0.9 \cdot S_{2,u} + \dots + 0.1 \cdot S_{10,u}}{\sum_{i=1}^{10} S_{i,u}} & : \text{if } hits_u > 0 \\ 0 & : \text{else} \end{cases}$$

# Evaluation on Historical Cases

## □ Additional Metrics

### ▣ User coverage (Ucov)

$$Ucov = \frac{\sum_{u \in U} \rho_u}{|U|}$$

$$\rho_u = \begin{cases} 1 & : \text{if } |reset_u| > 0 \\ 0 & : \text{else} \end{cases}$$

### ▣ Catalog coverage (Ccov)

$$Ccov = \frac{|\bigcup_{u \in U} reset_u|}{|I|}$$

### ▣ Intra-list similarity (ILS)

$$ILS_u = \frac{\sum_{i \in reset_u} \sum_{j \in reset_u, i \neq j} sim(i, j)}{2}$$

# Analysis of Results

- T-test
  - ▣ Used to test whether the observed differences between two sample means are due to chance or represents a true difference between populations
- One-way Analysis of Variance (ANOVA)
  - ▣ Used to compare the means of two or more groups
- Factorial ANOVA
  - ▣ Used to compare the means from four or more groups in a factorial design in order to decide whether the differences between means may be due to chance or one of the factors or a combination of the factors

# Alternative Types of Research Designs

- Preexperimental design
  - ▣ No control group, no random assignment
  - ▣ Used to collect preliminary or pilot data
- Experimental design
  - ▣ Rule out threats to internal validity through the use of control groups and random assignment
- Quasiexperimental design
  - ▣ Includes one or more control groups, but do not employ random assignment
- Nonexperimental design