UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Recommender Systems

Evaluation Methods

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

One problem

$$f(u,i)$$

Many solutions

Gazillions of algorithms

- Collaborative
- Content-based
- Knowledge-based
- Hybrid

Which one to choose?

Which one to choose?

“Research has shown that deep matrix factorization is the best method, we should implement it” **#NOT**

Better: ask questions

- “Why do I need a recommender system?”
- “What are the constraints of the system?” (e.g., latency, privacy, data size, data sparsity)

Evaluation

> *Evaluation is a systematic determination of a subject's merit, worth and significance, using criteria governed by a set of standards.*

- https://en.wikipedia.org/wiki/Evaluation

What to evaluate?

Three fundamental targets

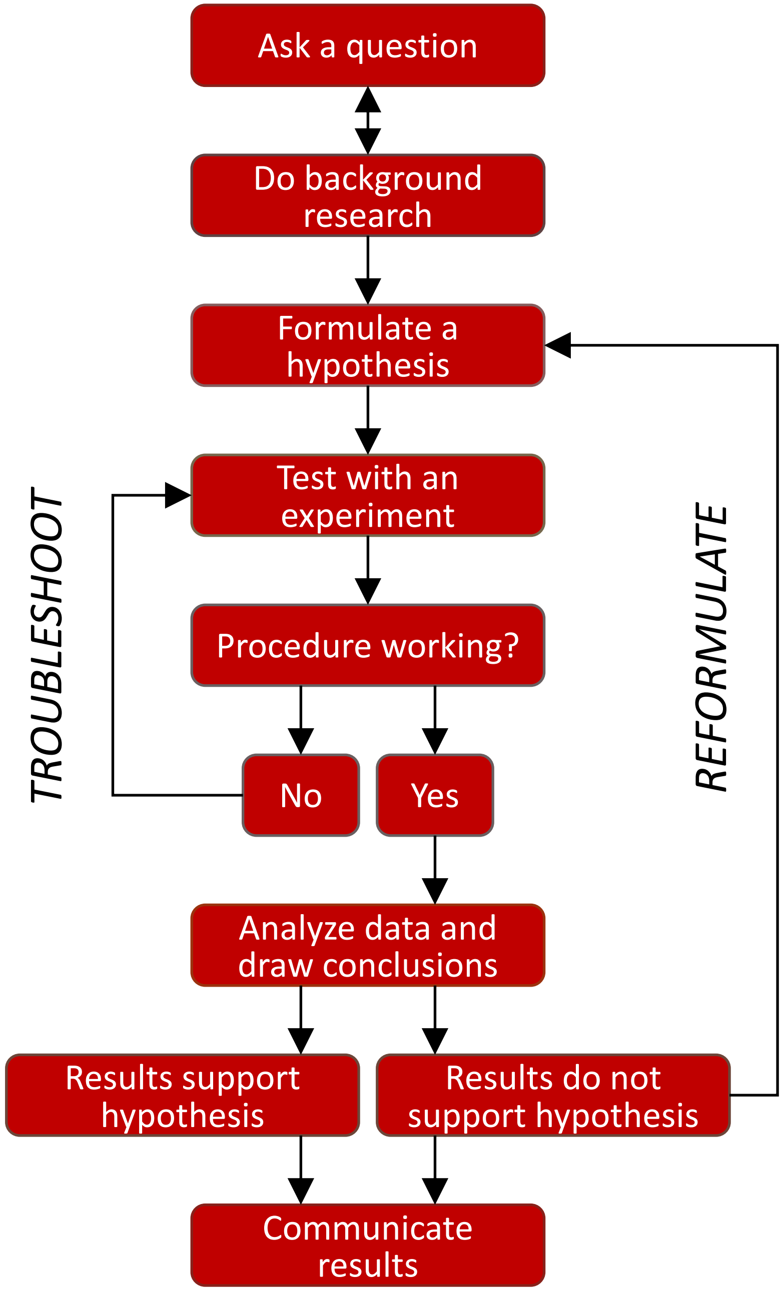
- Systems (efficiency)
- Methods (effectiveness)
- Applications (user utility)

Evaluation plays a critical role for all three

- Our primary focus is on “methods” research

How to evaluate?

Scientifically, of course!



Asking questions

What problem are you trying to solve?

- Or in recommendation parlance, what **task**?

Hard to solve an ill-defined task!

- Is it a well-known task? Review the literature!
- Is it unlike anything done before?

Asking (new) questions

Characterize the task (see class #2)

- How is the system used?
- What are the inputs? Outputs?
- How do you define success?

Defining success

Can we improve the **user satisfaction**?

- Can we recommend good movies to watch?
- Can we recommend good books to read?
- Can we recommend good follow-up stories?

Defining success

Can we improve the **system performance**?

- ~~Can we recommend good movies to watch?~~
  - Can we retain user subscriptions?
- ~~Can we recommend good books to read?~~
  - Can we increase sales?
- ~~Can we recommend good follow-up stories?~~
  - Can we display more ads?

Formulating hypotheses

A hypothesis is an explanation for a phenomenon

- *The moonlight is produced when little green men on the moon throw a party, but they will hide whenever anyone on earth looks for them, and will flee into deep space whenever a spacecraft comes near*

Is this a ***scientific*** hypothesis?

Formulating hypotheses

A hypothesis must be falsifiable

- Ideally concerning an isolated component
  e.g., *dim. reduction improves collaborative filtering*

It either holds or does not…

- … with respect to the considered data (scope)
- … perhaps under certain conditions (extent)

Formulating hypotheses

Methods are not devised arbitrarily

- We always have a hypothesis (whether implicit or explicit) for why our work should improve
- Even the best results are useless if nobody understands what you are trying to solve

So, spell out your hypotheses!

Performing experiments

Key components

- Experimental setup
- Analysis of results

Key concern: ***reproducibility***

- Must specify each and every detail needed for reproducing our method and the experiment

Experimental setup

Task definition

Research hypotheses

Reference comparisons

Evaluation methodology

Reference comparisons (aka baselines)

"My method is 90% accurate"

- Meaningless without a reference comparison
- Rephrasing: is it better or worse?

Choice of baseline depends on the hypothesis

- Key question: what are you trying to show?

Choosing baselines

Vanilla baselines

- Have the proposed effect turned off
  e.g., collaborative filtering without dim. reduction

Competing baselines

- Exploit the proposed effect in a different manner
  e.g., probabilistic matrix factorization

Evaluation methodology

Feedback

- Implicit
- Explicit

Mode

- Retrospective
- Prospective

	retrospective	*prospective*
implicit	counterfactual evaluation	online evaluation
explicit	offline evaluation	

Online evaluation

Prospective experiments

- How well can we predict future preferences?

Benchmarked using live user interactions

- Poorly reproducible
- Highly realistic

Online evaluation

Focus on implicit user feedback

- Derived from observable user activity
- Captured during natural interaction

Implicit signals with various levels of noise

- Clicks, dwell-times, purchase decisions

Allows for detecting causation

Controlled experiments

When different variants run concurrently, only two things could explain a change in metrics

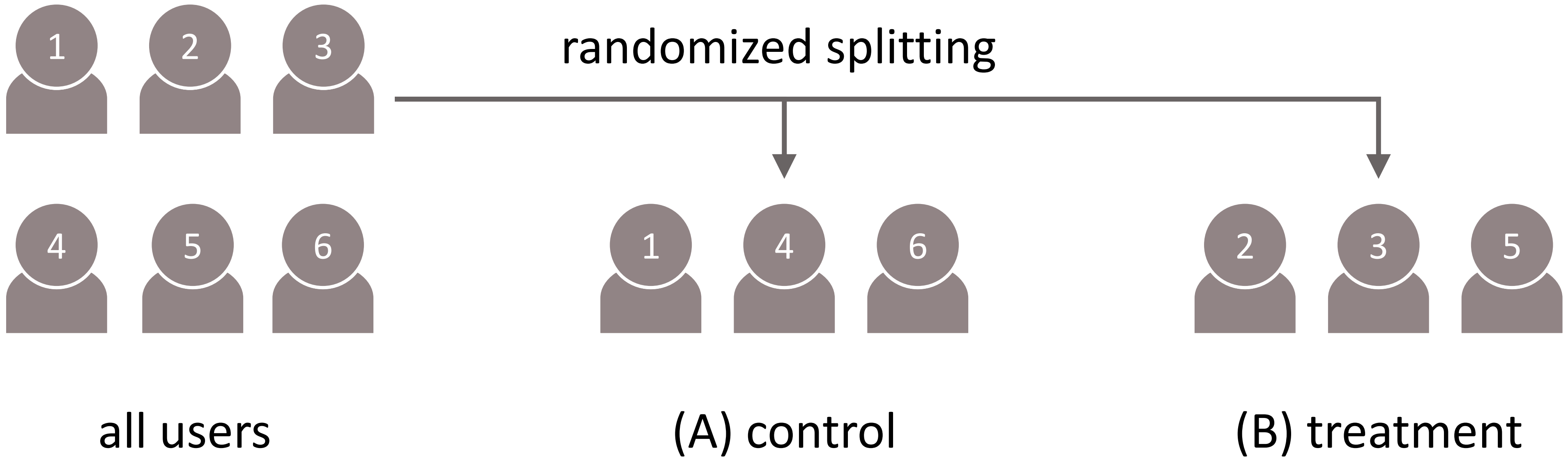
- Their “feature(s)” (A vs. B)
- Random chance

Everything else happening affects both the variants

- For #2, we conduct statistical tests for significance

A/B test

Each user is exposed to a single variant



Offline evaluation

Testing with people is always expensive

- A/B test may take too long to converge
- Exposing users to prototypes is risky

We need a cheap and rapid protocol

- *Simulate* the behavior of real users

Offline evaluation

Retrospective experiments

- How well can we predict (hidden) past preferences?

Benchmarked using static datasets

- Highly reproducible
- Poorly realistic

Offline evaluation

Goal is to *estimate* the recommender’s quality

- High-throughput evaluation
- Answer important research questions

Often can’t answer if recommender really works

- User-based evaluation needed
- Link to business metrics is weak

Public datasets

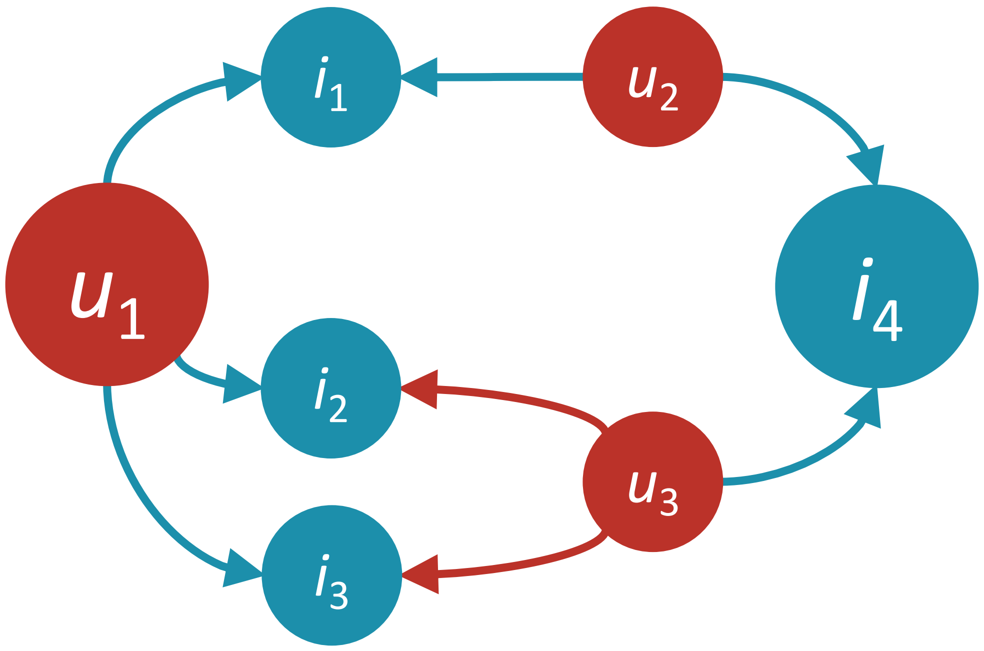
Many available datasets for movies, music, books, food, papers, jokes, tags, dates, healthcare, you name it!

- http://cseweb.ucsd.edu/~jmcauley/datasets.html
- https://gist.github.com/entaroadun/1653794
- https://toolbox.google.com/datasetsearch
- https://www.kaggle.com/datasets

You can build your own

Three core components

- A set of users
- A set of items
- A map of preferences

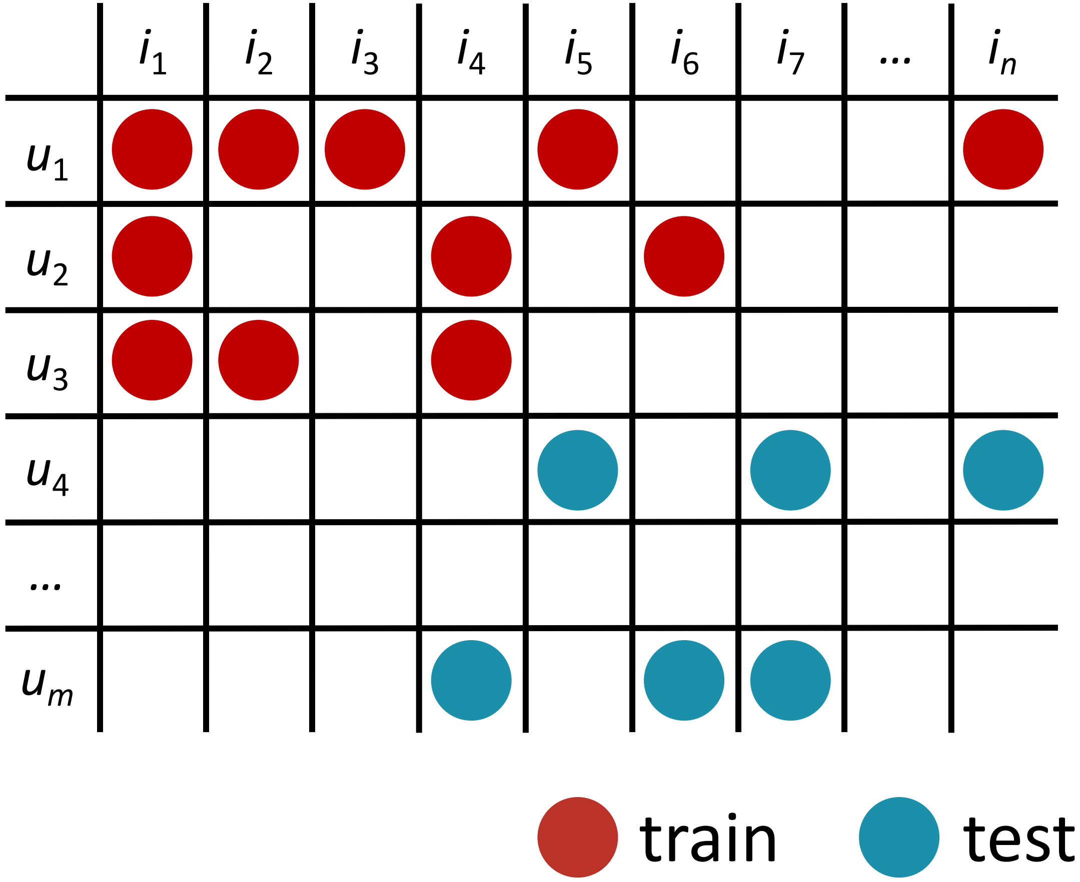


How to simulate user behavior?

	i_1	i_2	i_3	i_4
u_1	5	3	3	
u_2	5			3
u_3		1	2	1

split available data
into **training** and **test**

Splitting by user



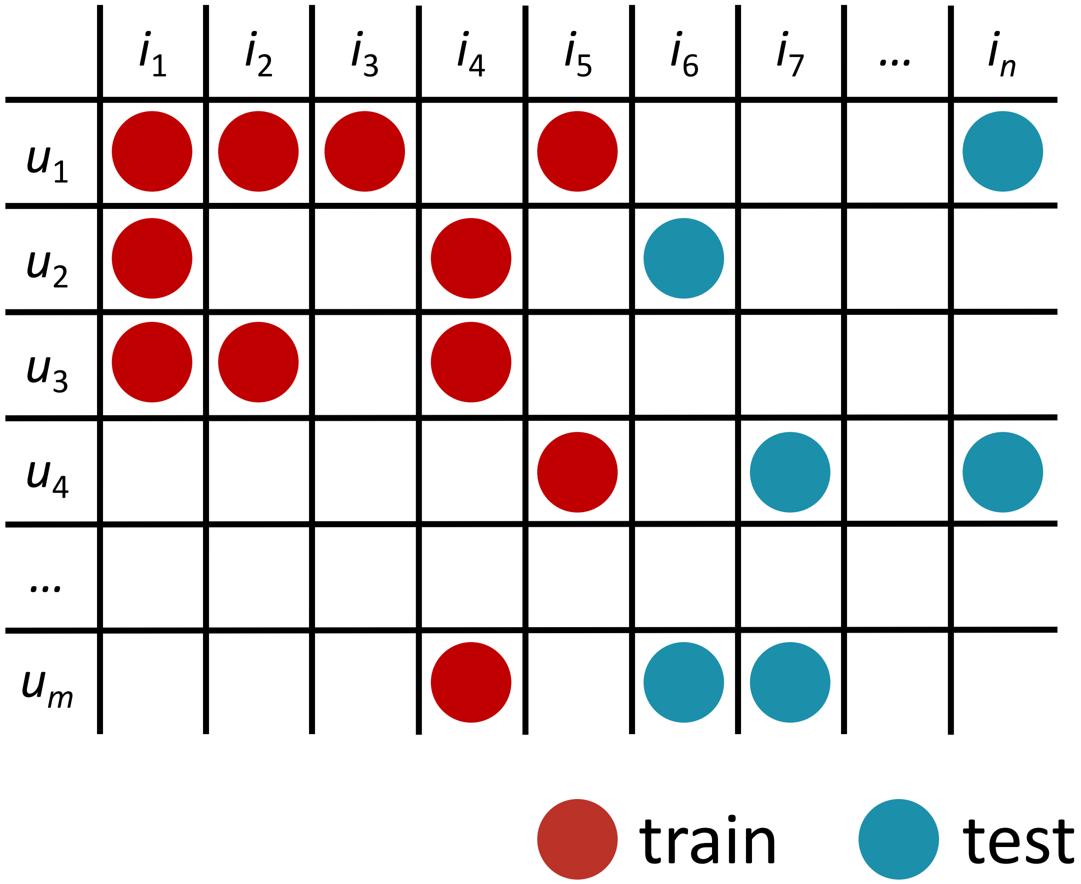
Split users

- Learn from some users
- Predict for others

Problem

- Test cases are user cold-start (i.e., users have no training)

Splitting by item



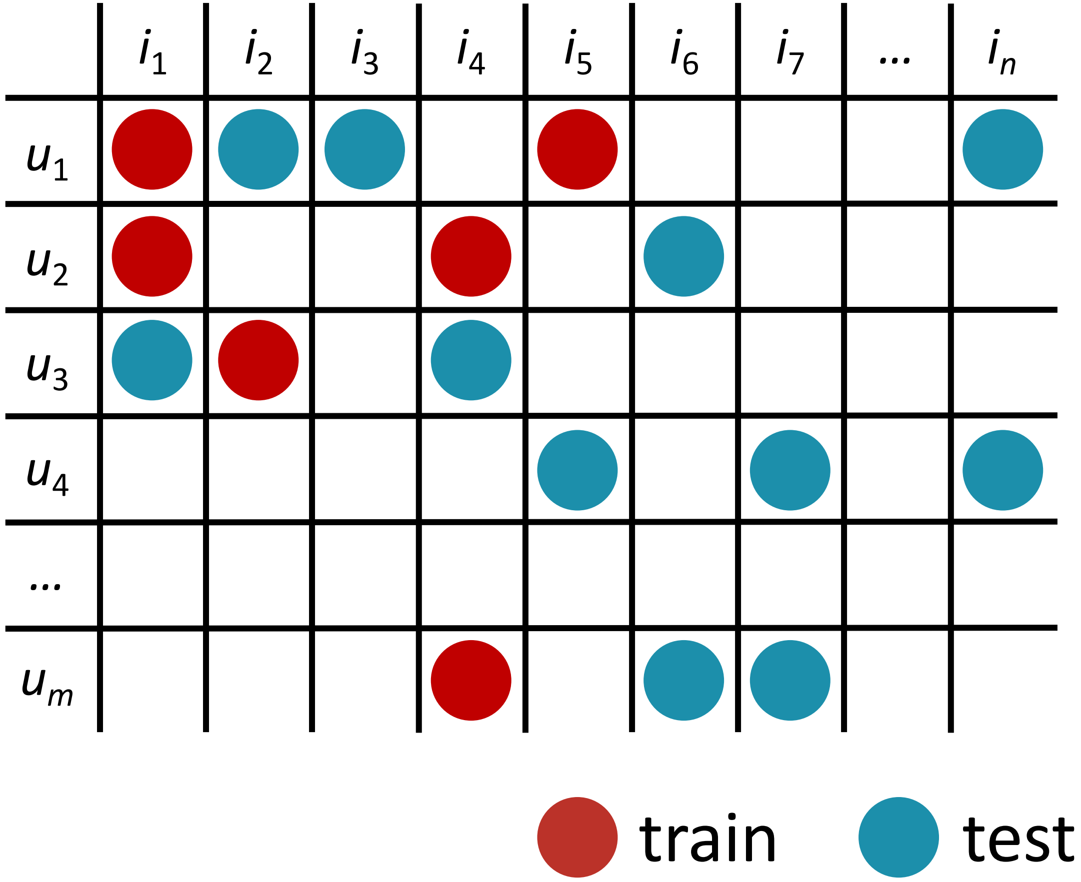
Split items

- Learn from some items
- Predict for others

Problem

- Test cases are item cold-start (i.e., items have no training)

Splitting by interaction (randomly)



Split interactions

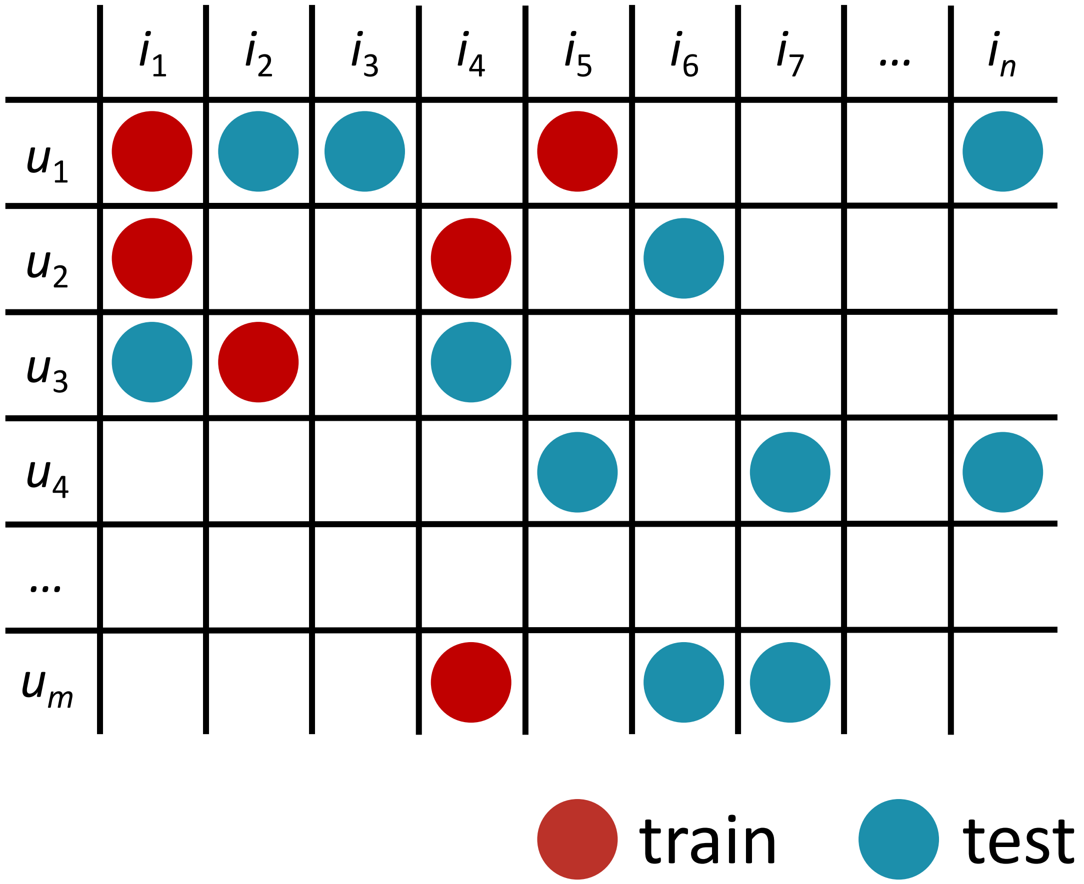
- Learn from partial profiles of both users and items
- Predict hidden interactions

Advantage

- Realistic mix of cold and non-cold start test cases

Problem?

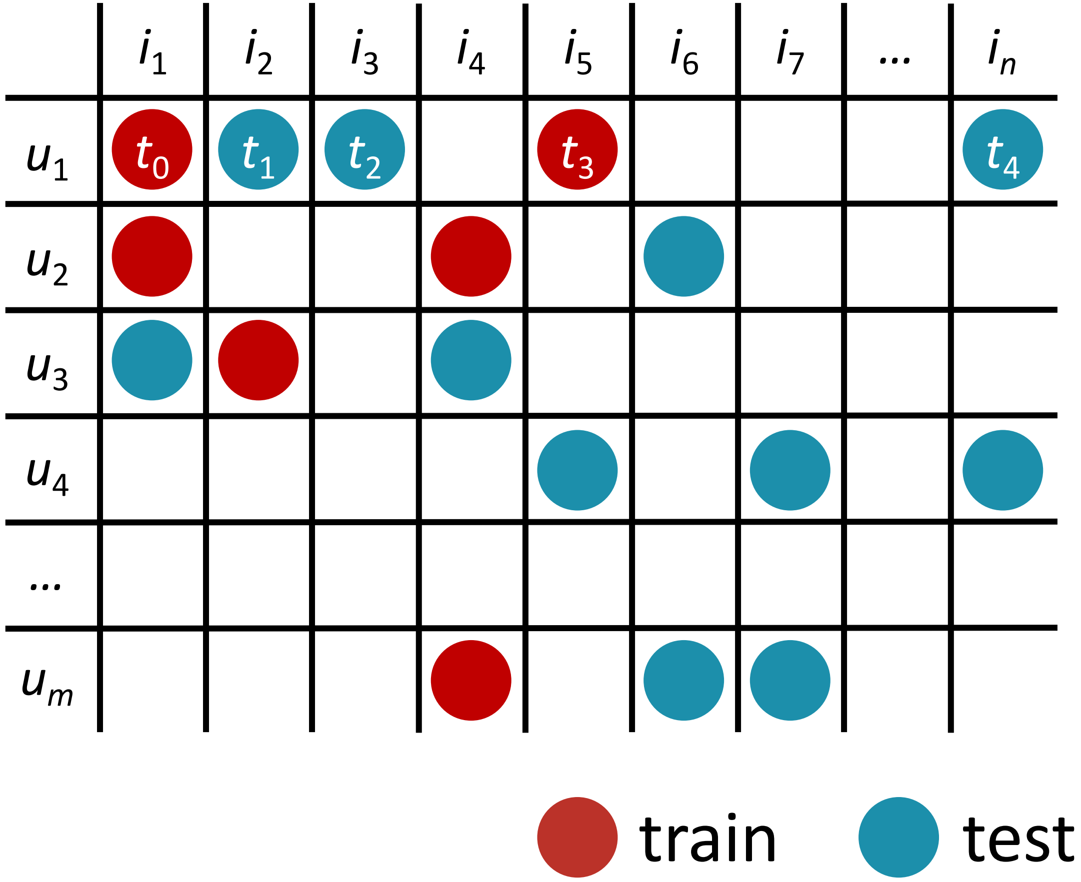
Splitting by interaction (randomly)



Problem

- Interactions aren’t truly i.i.d.
- Future interactions may leak into the training set

Splitting by interaction (randomly)



Problem

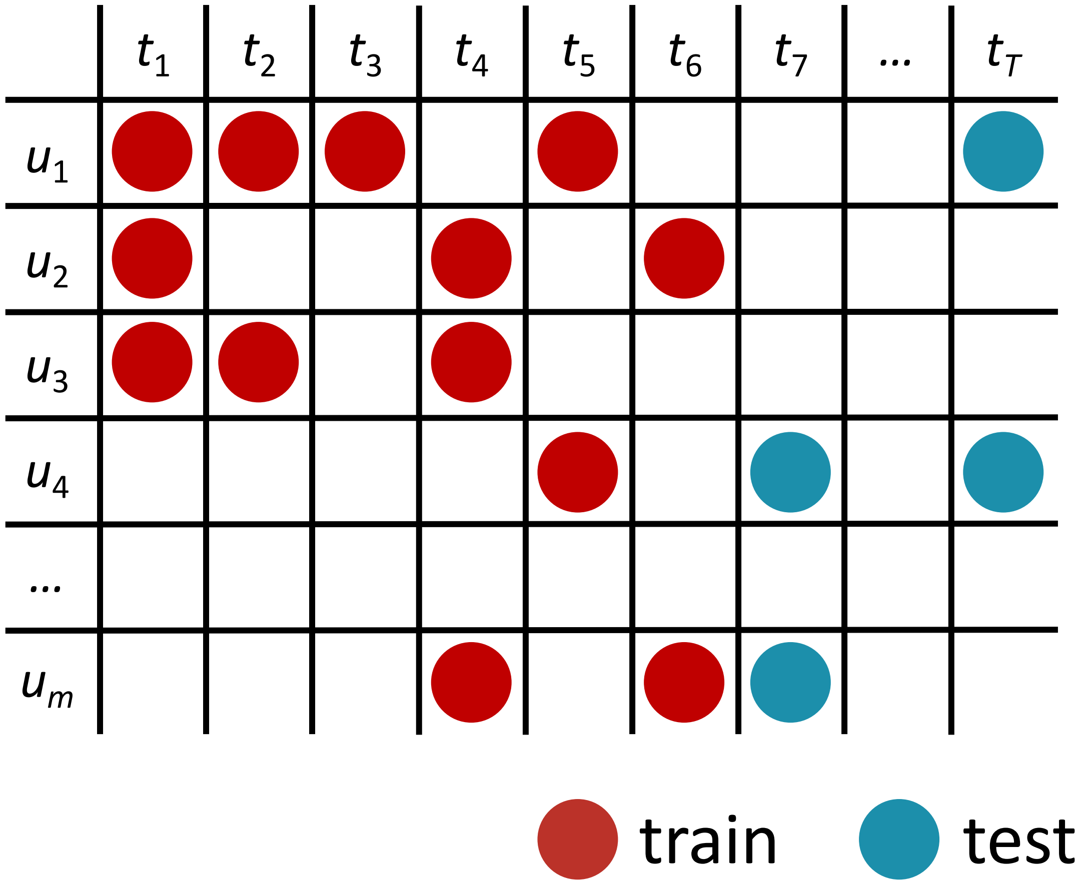
- Interactions aren't truly i.i.d.
- Future interactions may leak into the training set
  - t_0: western
  - t_1, t_2: sci-fi
  - t_3: sci-fi (leaked to improve t_1, t_2)

Splitting by interaction (randomly)

Other popular yet unjustified protocols

- Global: choose 10% of the overall interactions for test
- Given-k: choose k items from each user for training
- All-but-k: choose k items from each user for test

Splitting by interaction (temporally)



Advantages

- Realistic mix of cold and non-cold start test cases
- No future data leaking
- Could test with sliding windows with multiple cutting points to counter seasonal effects

Summary

Evaluating recommenders is hard

- Offline evaluation doubly so

Online evaluation not always an option

- Costly access to user base, exploration risk

Need to design tests around goals

- Different methods can achieve different results

References

Recommender Systems: An Introduction (Ch. 7)

Recommender Systems Handbook (Ch. 8)

Recommender Systems: The Textbook (Ch. 7)

Statistical Methods for Recommender Systems (Ch. 4)