SE125 Machine Learning

Recommender Systems

Yue Ding

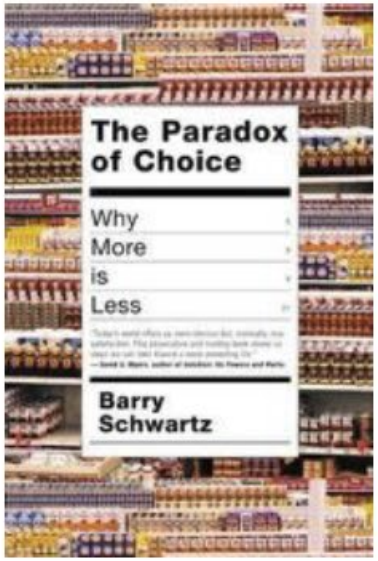
School of Software, Shanghai Jiao Tong University
dingyue@sjtu.edu.cn

References and Acknowledgement

- CCF Advanced Disciplines Lectures, 2021, Recommender Systems

The tyranny of choice

Information overload

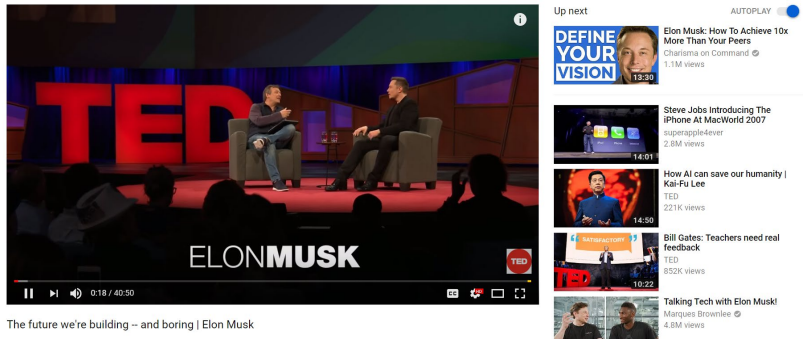


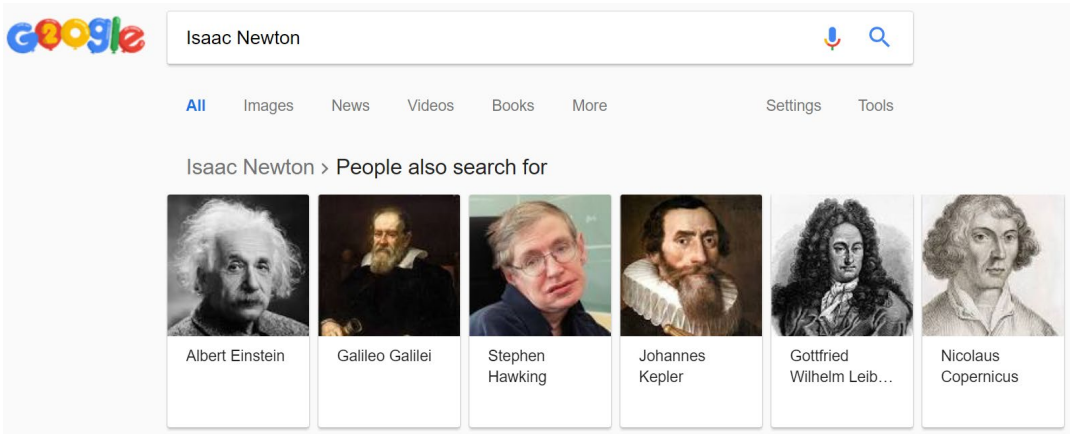
"People read around 10 MB worth of material a day, hear 400 MB a day, and see 1 MB of information every second" - The Economist, November 2006

In 2015, consumption will raise to 74 GB a day - UCSD Study 2014

Recommender Systems

- Recommender Systems are a particular type of personalized Web-based applications that provide to users personalized recommendations about content they may be interested in.





Recommender systems

推荐技术是商业化最成功的AI技术

音乐

视频

电子商务

个性化阅读/信息流

社交网络好友推荐

位置服务推荐

广告推荐

App分发

主播推荐

1. **改善用户体验，提升产品竞争力**

你关心的，就是头条

2. **流量变现**

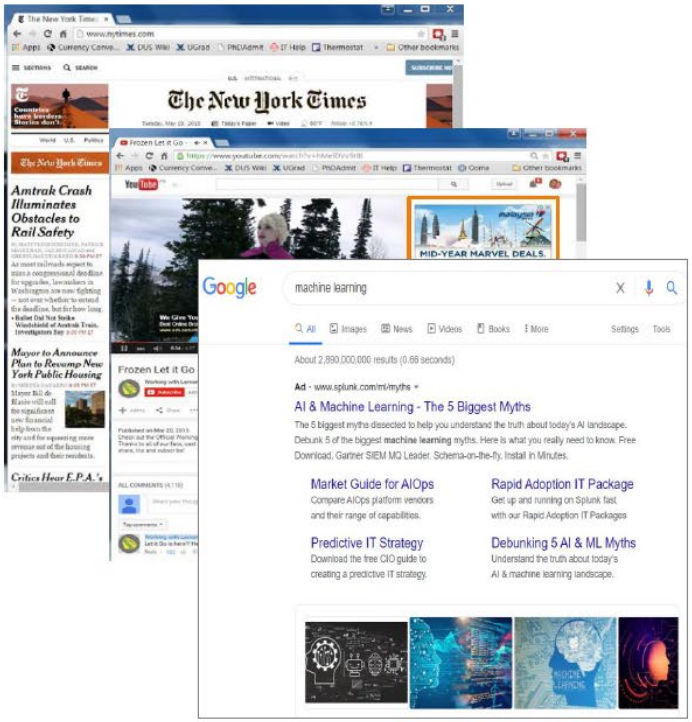
信息创造价值

The value of recommendations

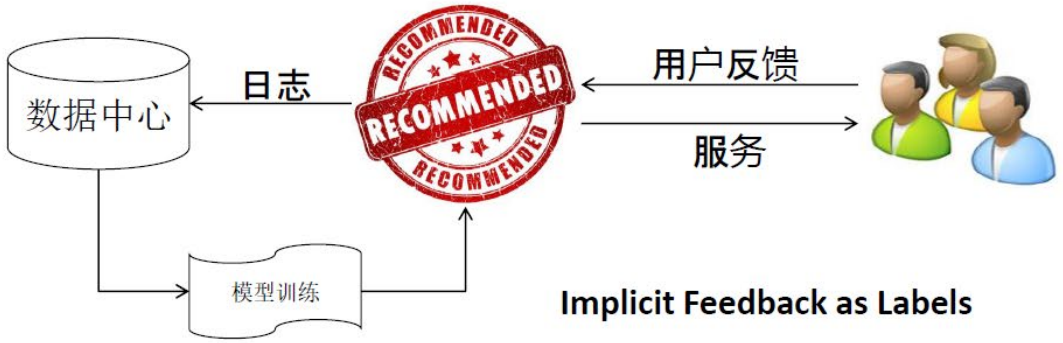
- 35% of the purchases on Amazon are the result of their recommender system, according to **McKinsey**.
- Recommendations are responsible for 70% of the time people spend watching videos on **YouTube**.
- 75% of what people are watching on Netflix comes from recommendations, according to **McKinsey**.

2018年天猫双十一达成2135亿元交易，背后**产生了453亿次AI个性化推荐**，使消费者在海量商品中**更容易找到真正需要**的商品

交互式系统：Implicit Feedback in Closed-loop



推荐/搜索/广告



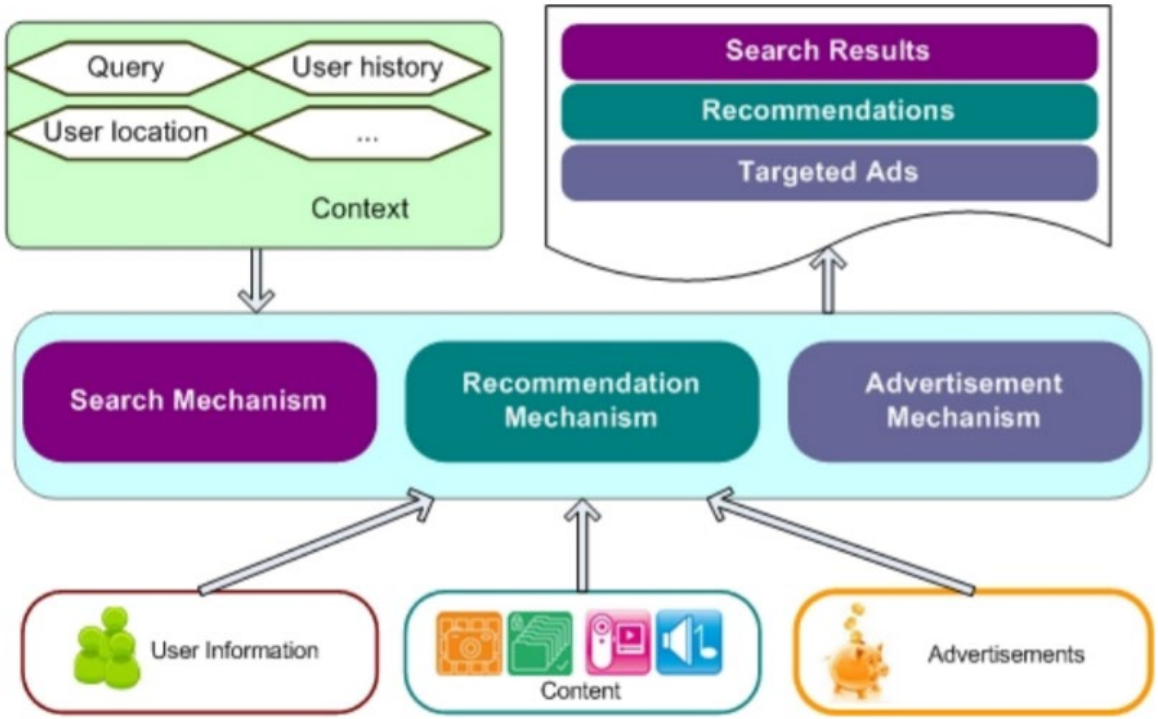
Implicit Feedback as Labels

- 通过服务用户，获得大数据
- 通过学习大数据，提升服务能力

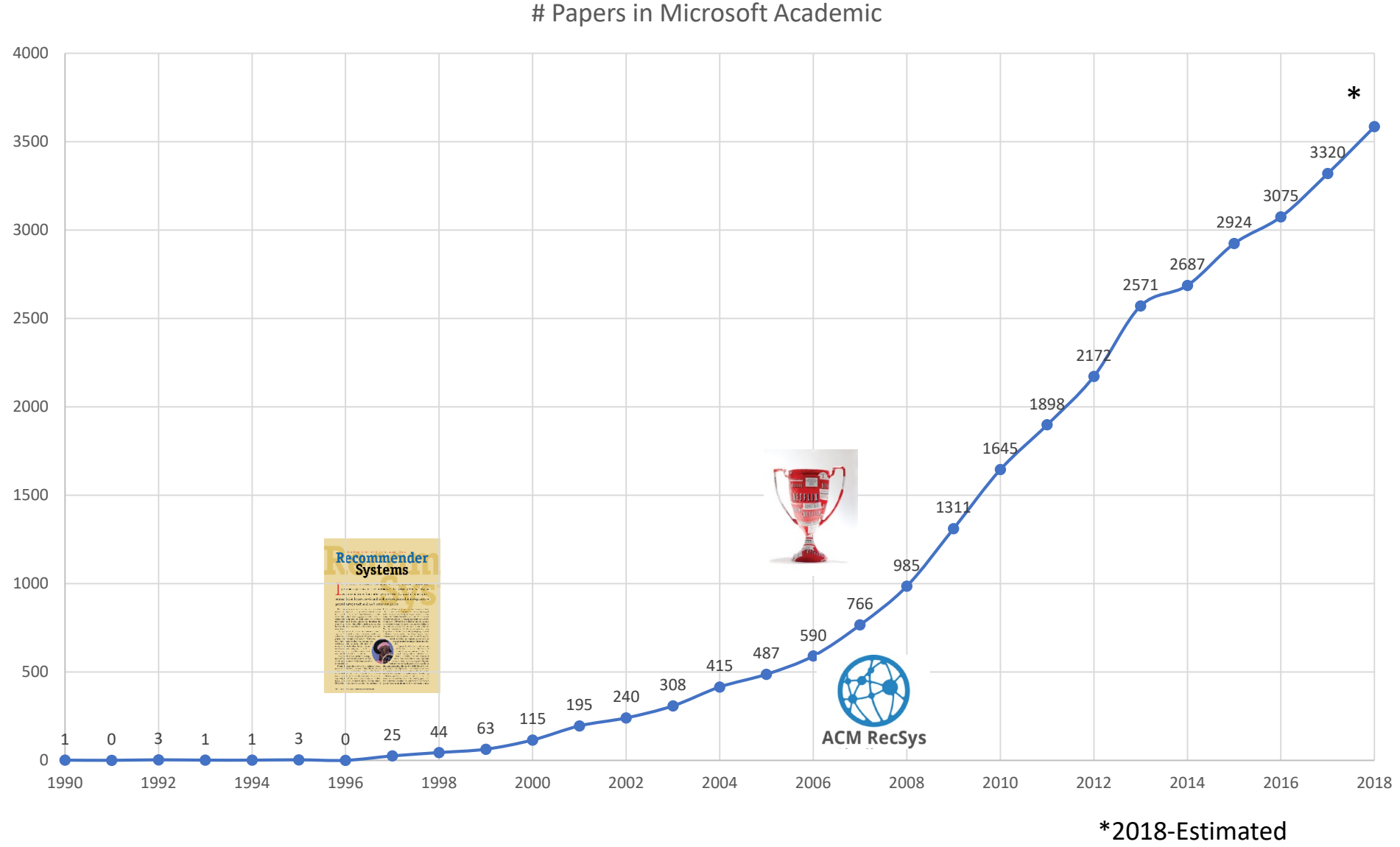
- 失误非致命性、场景封闭化
- 大量用户隐式反馈（Label数据）
- 商业模式清晰

搜索、推荐和广告

- 搜索：有明确的搜索意图，搜索出来的结果和用户的搜索词相关。
- 推荐：不具有目的性，依赖用户的历史行为和画像数据进行个性化推荐。
- 广告：借助搜索和推荐技术实现广告的精准投放，可以将广告理解成搜索推荐的一种应用场景，技术方案更复杂，涉及到智能预算控制、广告竞价等。

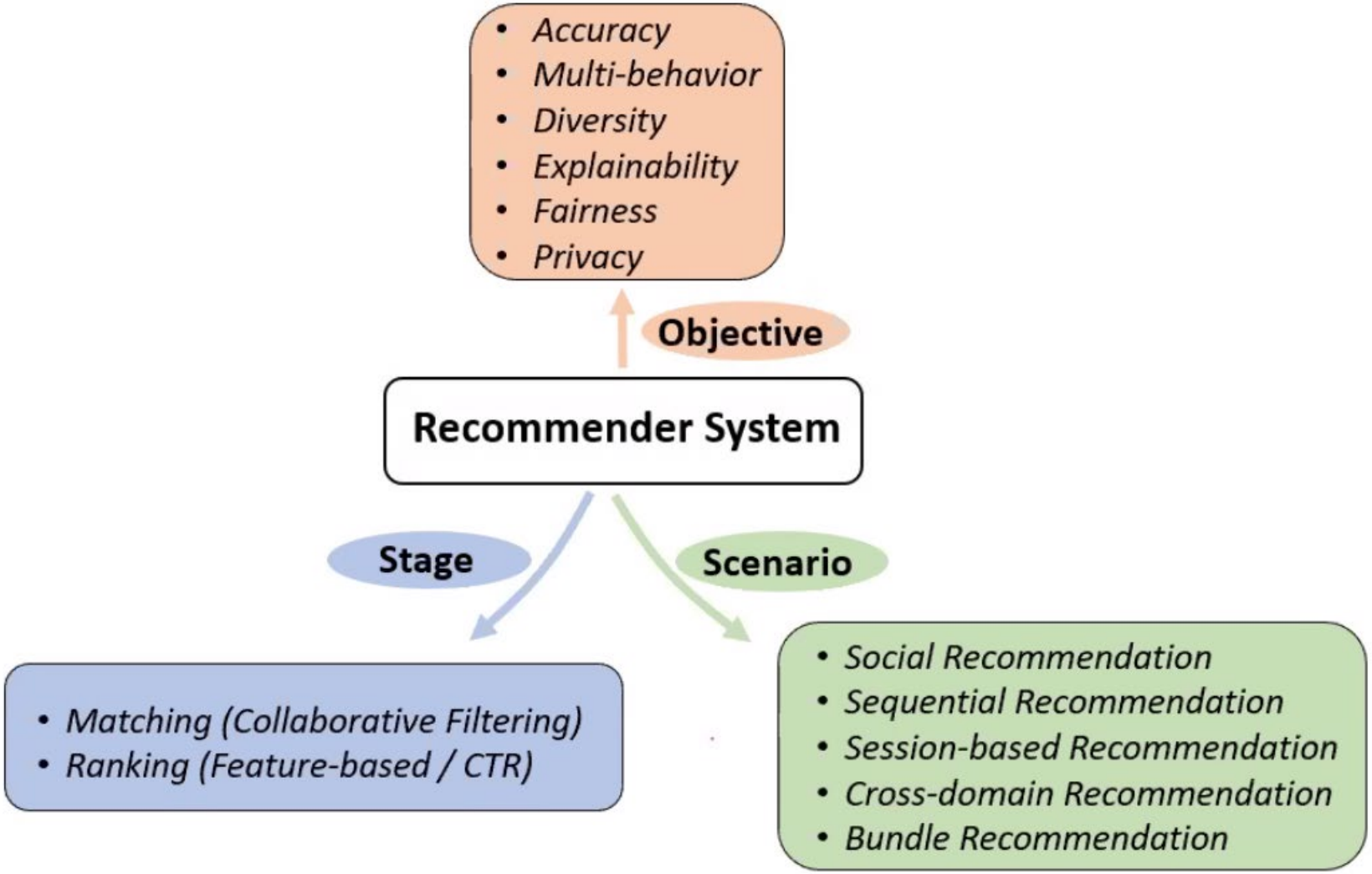


The Rise of Recommender Systems

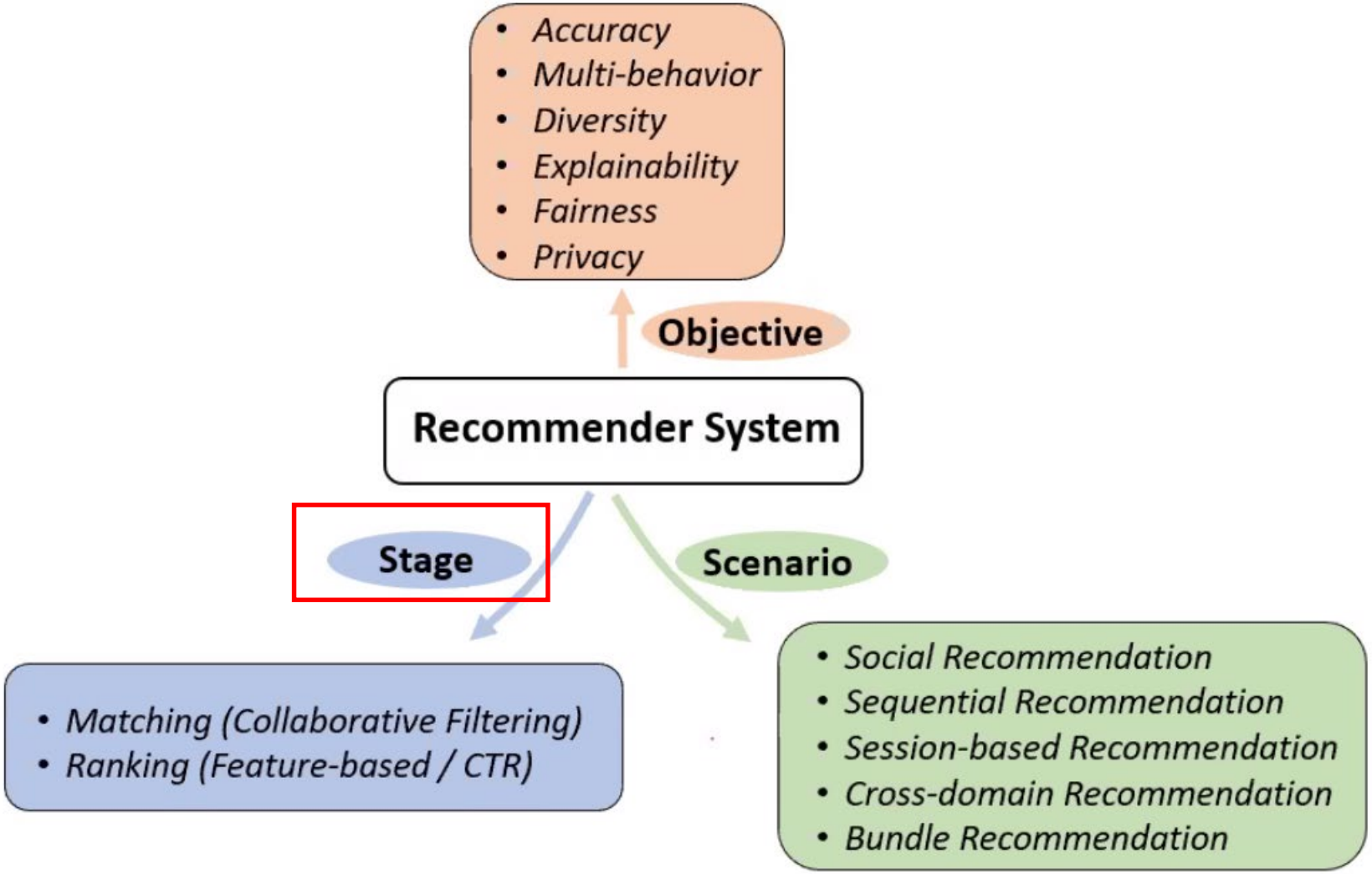


*2018-Estimated

Recommender Systems

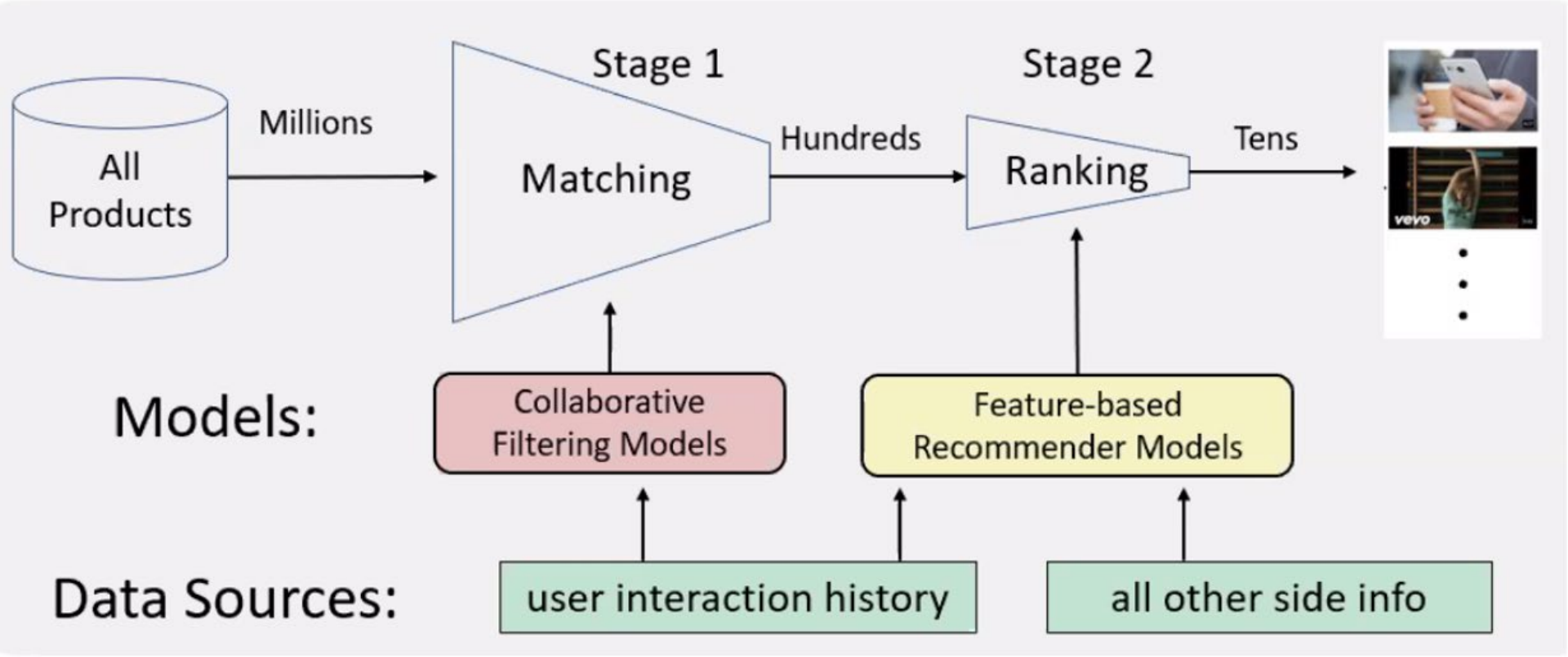


Recommender Systems



Recommender Systems

- Stages
  - Matching: recall items from all-item pool
  - Collaborative-filtering models



➢ Collaborative filtering

items

users				
	1	0	0	1
	0	1	0	0
	1	1	0	0
	1	0	0	1

0/1 Interaction matrix

OR

items

users				
	5	-	-	3
	-	2	-	-
	4	1	-	-
	3	-	-	3

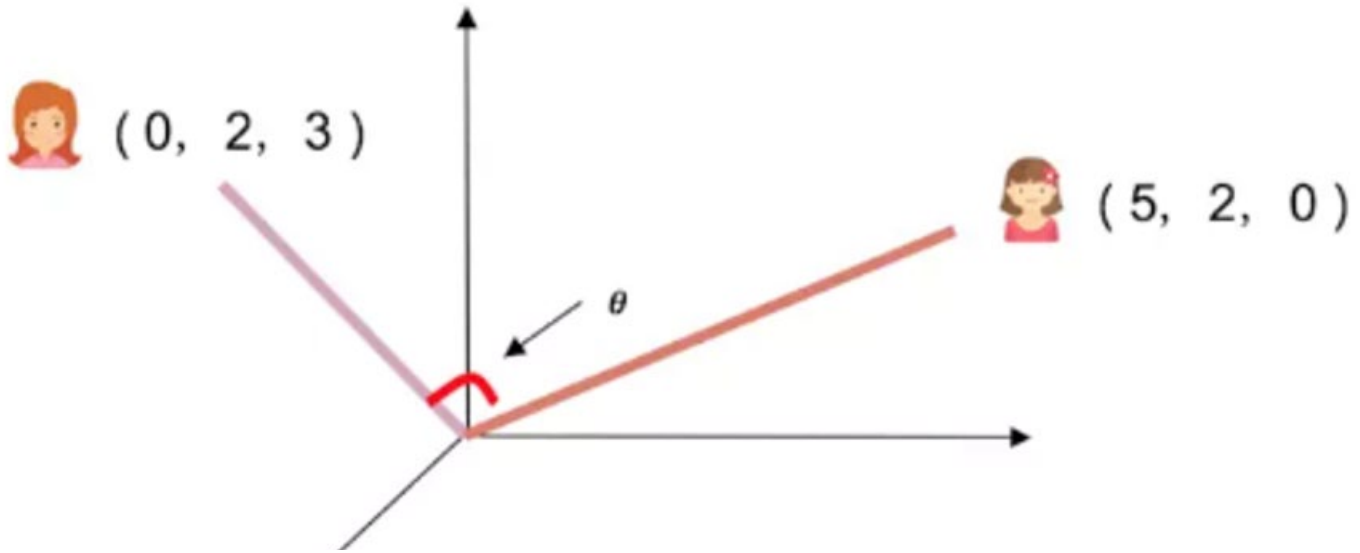
Rating matrix

- Implicit CF
- Scenario: e-commerce, ads, etc.
- Data: an interaction matrix
- Task: estimate positive position
- Measurement: Ranking metrics

- Explicit CF
- Scenario: movie, POI, etc.
- Data: a rating matrix (e.g. 1-5)
- Task: estimate ratings on unknown positions
- Measurement: Regression metrics

User Collaborative Filtering

$$\cos(\theta) = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} (p_i)^2} \times \sqrt{\sum_{i=1}^{n} (q_i)^2}}$$

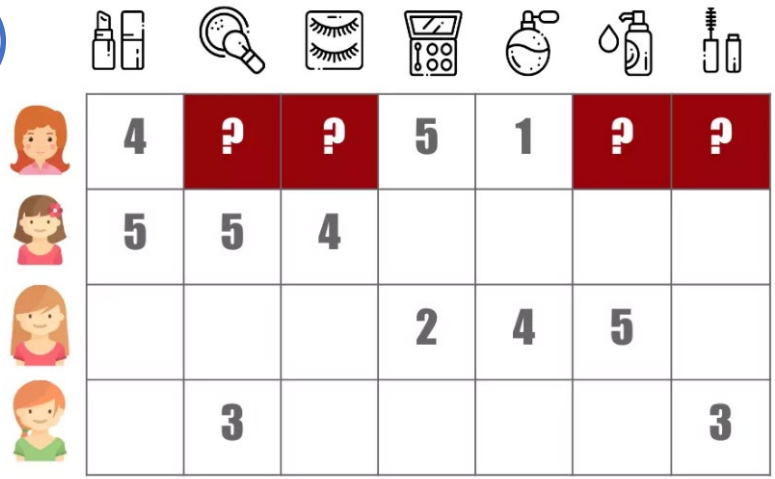


Result:

cos( ) = $\frac{0*5+2*2+3*0}{\text{sqrt}(0\text{^}2+2\text{^}2+3\text{^}2) * \text{sqrt}(5\text{^}2+2\text{^}2+0\text{^}2)}$ = **0.21**

User Collaborative Filtering

1



4	?	?	5	1	?	?
5	5	4				
			2	4	5	
	3					3

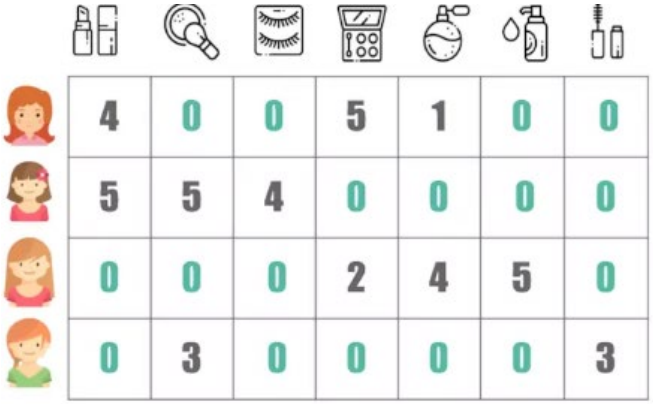
1 Assign average score to each person

Person 1: average score = (4+5+1) = 10/3

3

4	10/3	10/3	5	1	10/3	10/3
5	5	4	14/3	14/3	14/3	14/3
11/3	11/3	11/3	2	4	5	11/3
11/3	3	11/3	11/3	11/3	11/3	3

2



4	0	0	5	1	0	0
5	5	4	0	0	0	0
0	0	0	2	4	5	0
0	3	0	0	0	0	3

P1&P2:
Cosine similarity = **0.38**
(4*5 + 0*5 + 0*4 + 5*0 + 1*0 + 0*0 + 0*0) /
[(SQRT(4^2+0^2+ 0^2+ 5^2+1^2+0^2+0^2)*
(SQRT(5^2+5^2+ 4^2+ 0^2+0^2+0^2+0^2)]

P1&P3:
Cosine similarity = **0.32**
(4*0 + 0*0 + 0*0 + 5*2 + 1*4 + 0*5 + 0*0) /
[(SQRT(4^2+0^2+ 0^2+ 5^2+1^2+0^2+0^2)*
(SQRT(0^2+0^2+ 0^2+ 2^2+4^2+5^2+0^2)]

P1&P4:
Cosine similarity = **0.0**
(4*0 + 0*3 + 0*0 + 5*0 + 1*0 + 0*0 + 0*3) /
[(SQRT(4^2+0^2+ 0^2+ 5^2+1^2+0^2+0^2)*
(SQRT(0^2+3^2+ 0^2+ 0^2+0^2+0^2+3^2)]

4

2 Normalize

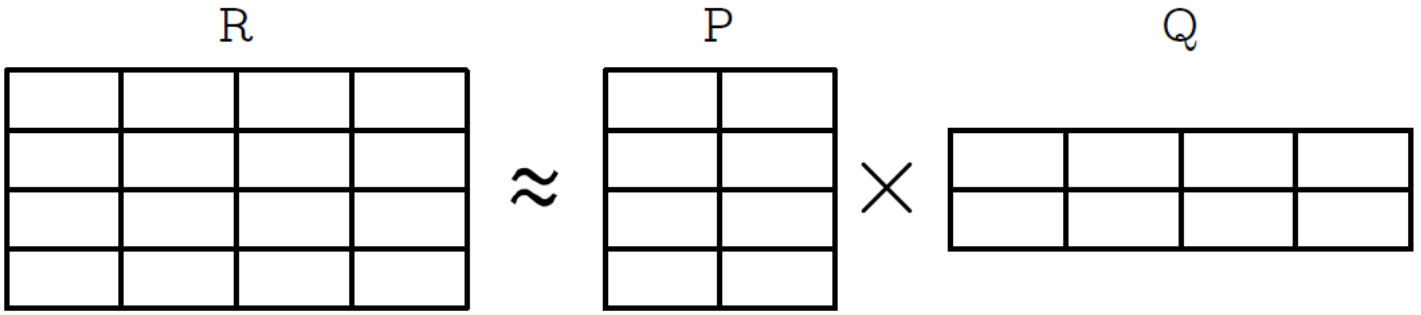
Formular: Score - Average(Scores)

Eg: Person 1 score: 4 - 10/3 = 0.7

0.7	0.0	0.0	1.7	-2.3	0.0	0.0
0.3	0.3	-0.7	0.0	0.0	0.0	0.0
0.0	0.0	0.0	-1.7	0.3	1.3	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0

Matrix Completion

- The completion is driven by a factorization



- Associate a latent factor vector with each user and each item
- Missing entries are estimated through the dot product

$$r_{ij} \approx p_i q_j$$

- Feature-based Recommender Models
  - Also known as Click-Through Rate Prediction
- Input: user/item attributes (ID can regarded as a kind of attribute)

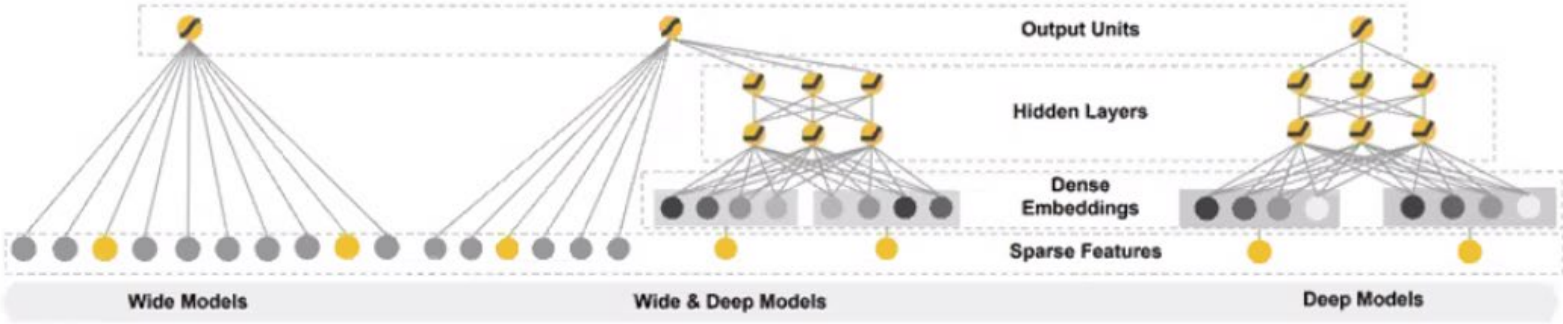
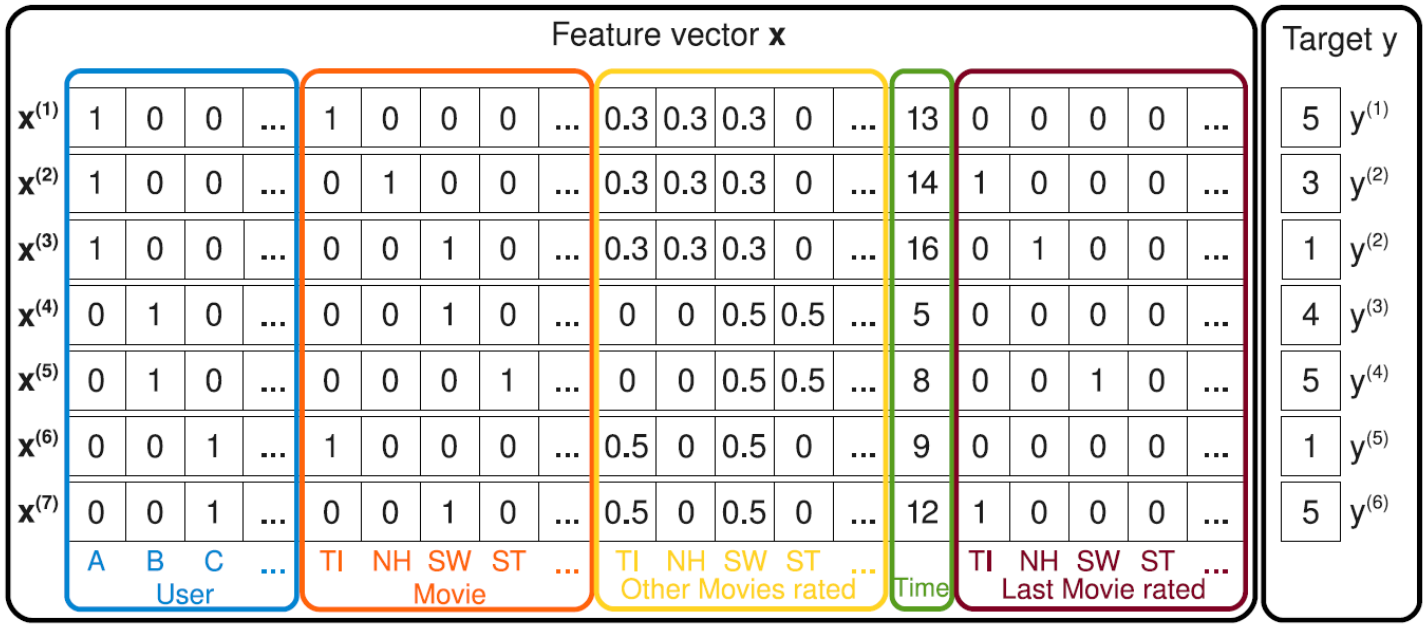


Figure from:
Cheng, H. T et al. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems

Factorization Machines



$$\mathrm{FM}: \hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j$$

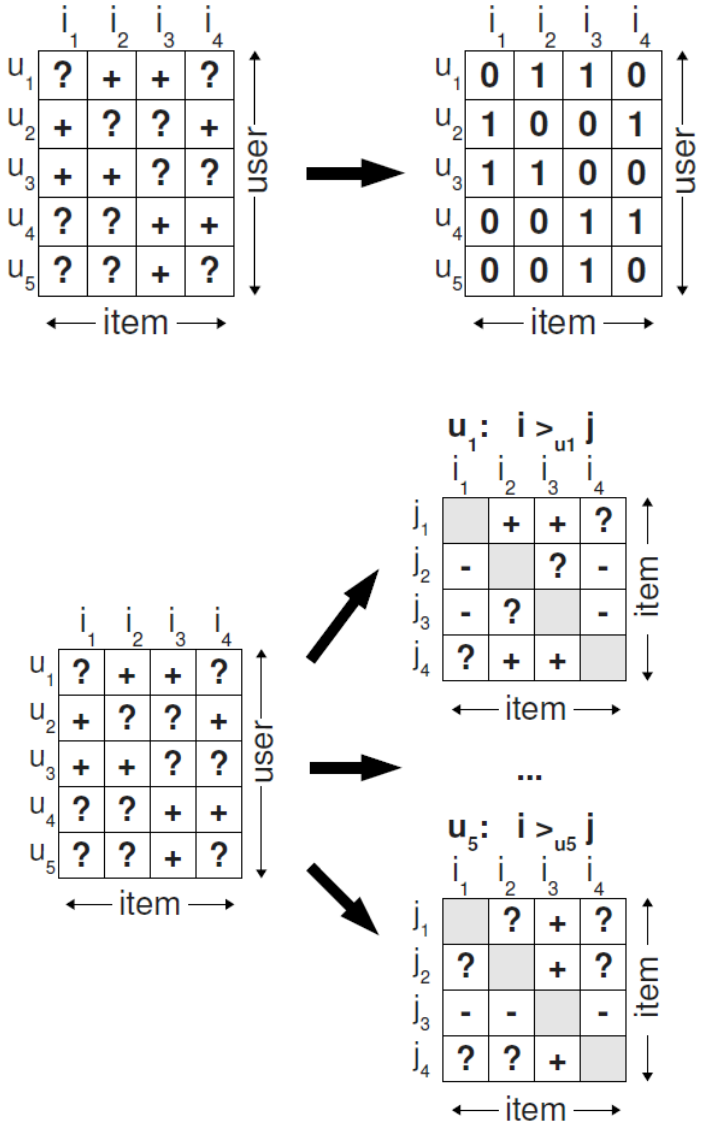
LR模型　　Dense化两两特征

$$\langle v_i, v_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f} =$$

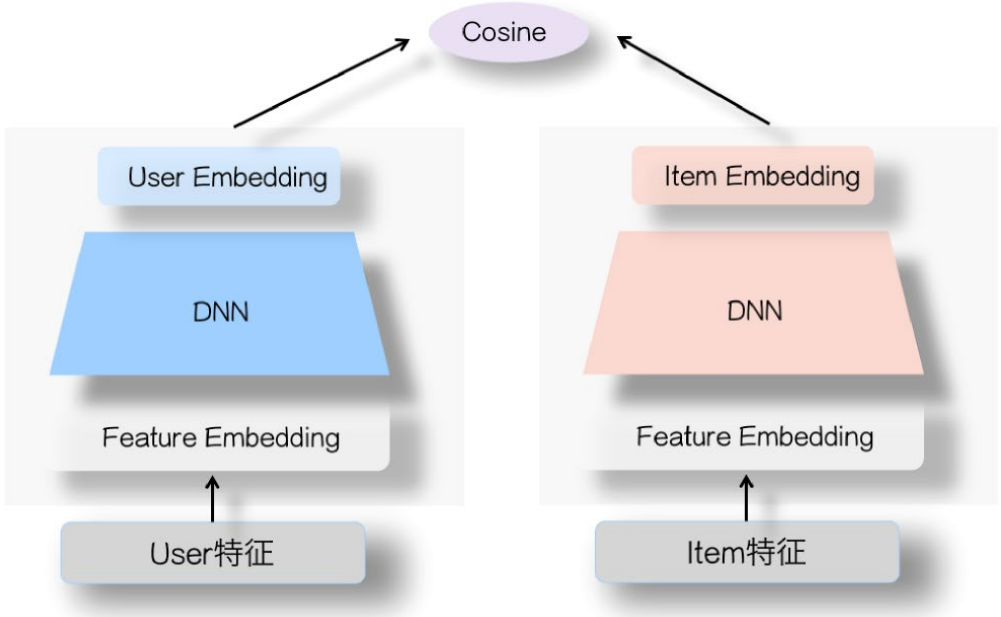
v_i: 0.1 0.2 0.6 0.8 0.1 0.2

v_j: 0.4 0.2 0.4 0.2 0.4 0.2

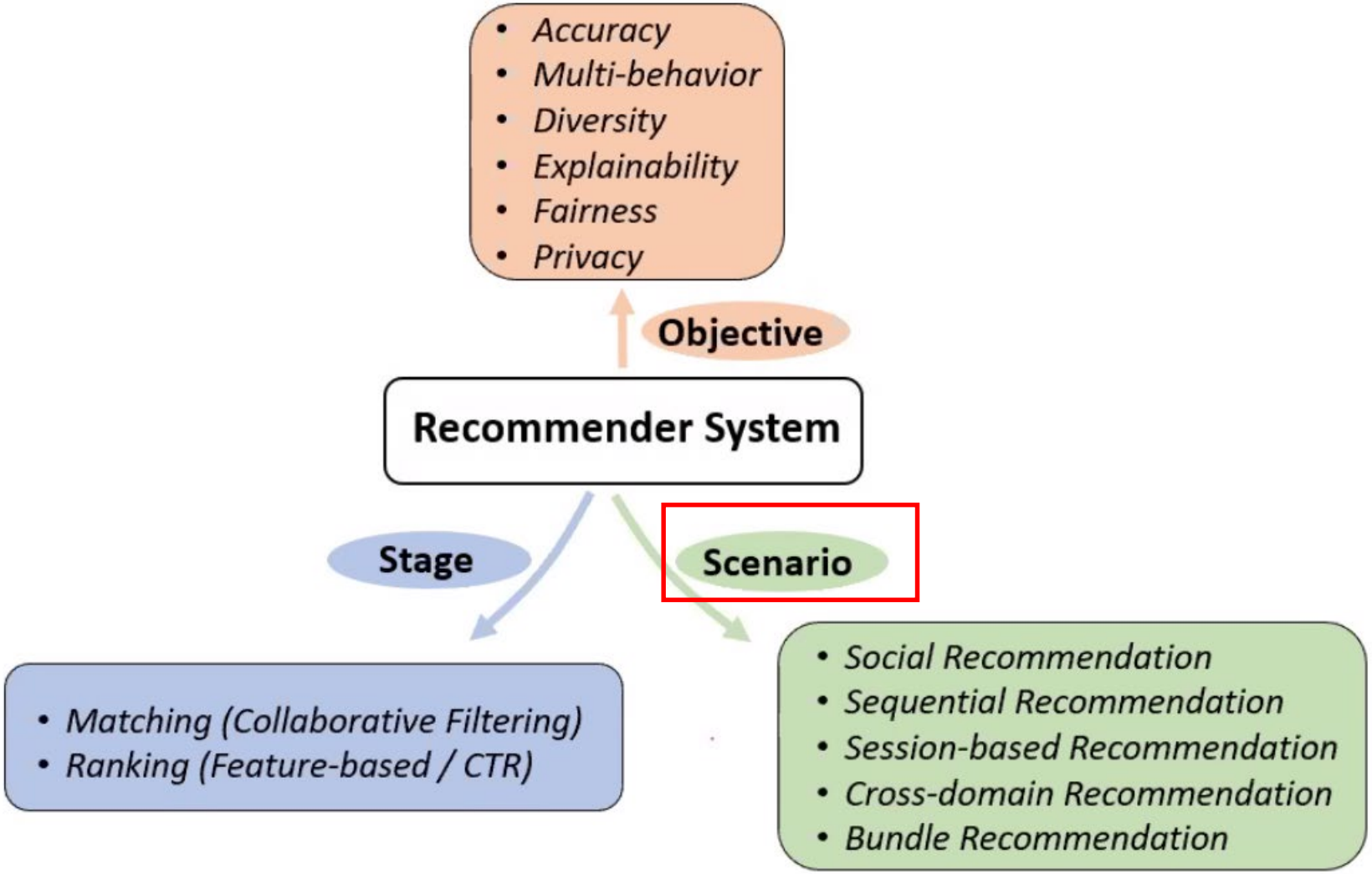
Bayesian Personalized Ranking



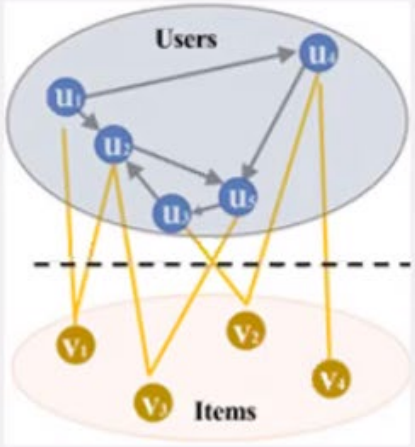
双塔模型-BPR训练



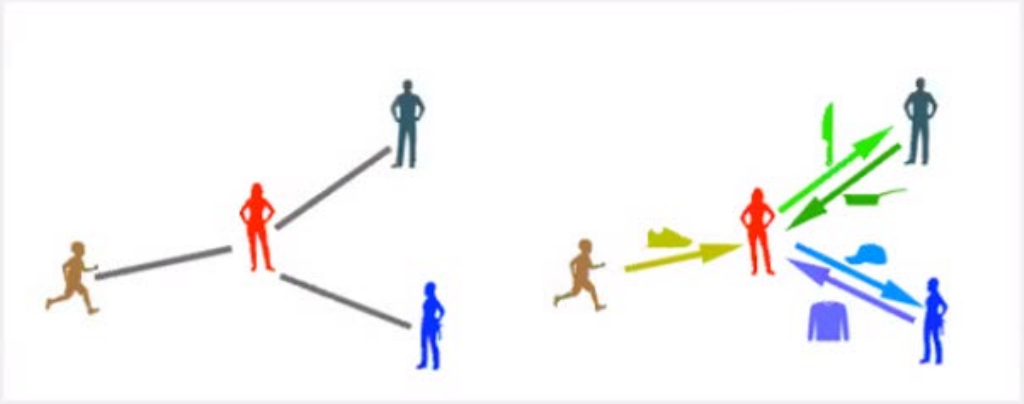
Recommender Systems



- Scenario: social recommendation
  - Definition: Improve recommendation with social network
  - *Social-trust* assumption: friends tend to have similar interests
  - **Input**: user interaction data + social relation data
  - **Output**: user-item interaction probability



Social Recommendation



Traditional Social RecSys v.s Social E-Commerce RecSys, such as Pinduoduo

Figures are from:
Wu et al. DiffNet++: A Neural Influence and Interest Diffusion Network for Social Recommendation. TKDE 2020
Lin et al. Recommender Systems with Characterized Social Regularization. CIKM 2018

- Scenario: sequential recommendation
  - Definition: predict user's next interaction based on historical sequential interactions
  - **Input**: user-item interactions at timestamps $t_1, t_2, \ldots, t_n$
  - **Output**: user-item interaction at timestamp t_{n+1}

- Scenario: session-based recommendation
  - Definition: predict next interaction based on anonymous short sequences
  - **Input**: *anonymous* behavior sessions
  - **Output**: next interaction of a given session

  - *Difference with Sequential Recommendation*
    - **Anonymous** (No user ID)
    - **Repetitive** items in one session
    - **Shorter** (as is collected in a short period)

- Scenario: cross-domain recommendation
  - Definition: recommendation with multi-domain datasets
  - Improve the target domain's performance with the auxiliary domain
  - **Input**: user-item historical interactions in multiple domains
  - **Output**: user-item interaction probability at target domain(s)
  - Challenges
    - Different user behaviors
    - Different data distribution
    - No overlapped user/item

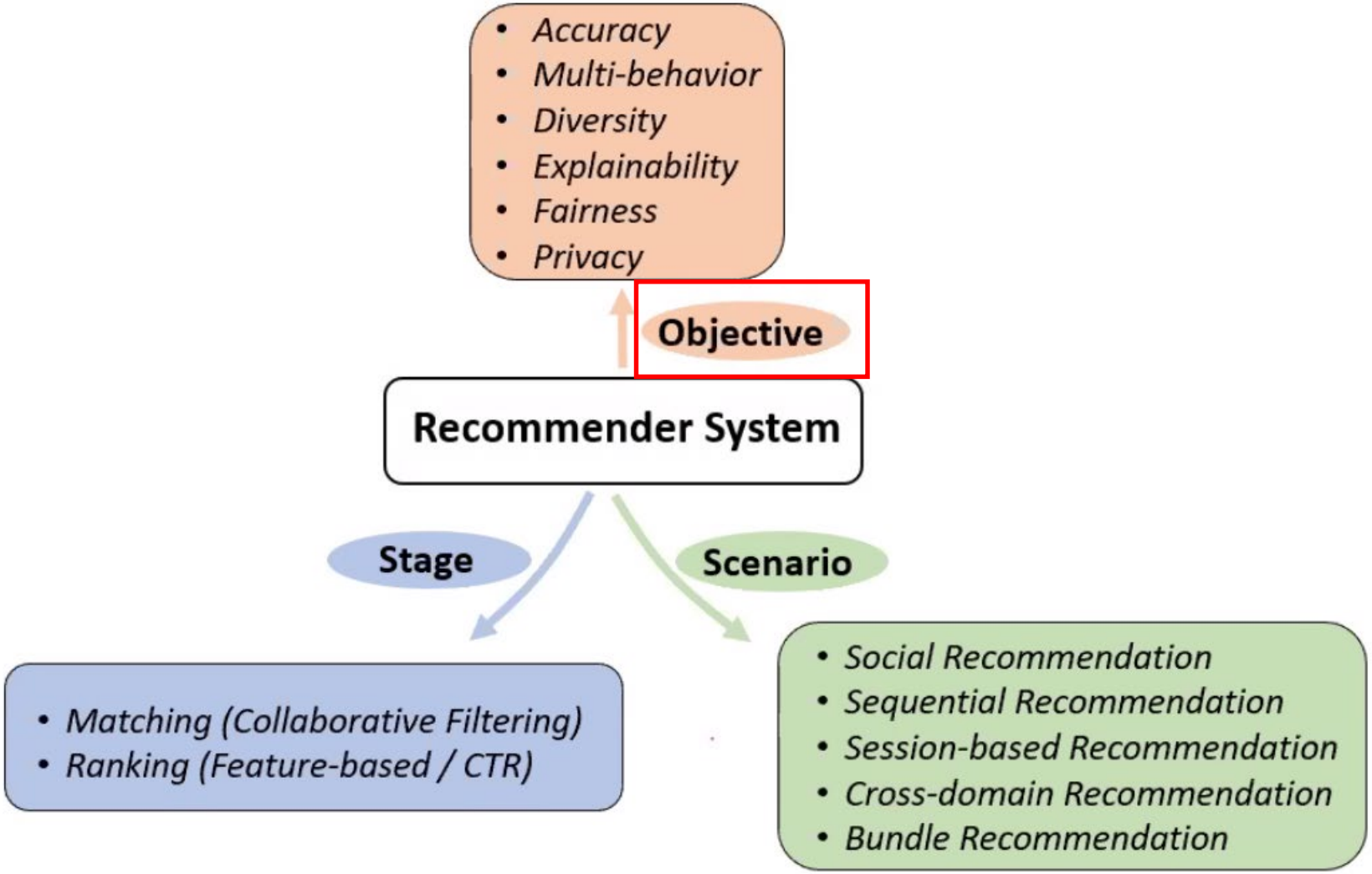
- Scenario: bundle/list recommendation
  - Definition: Recommend a bundle that is made with several items to user
  - **Input**: user-item/bundle historical interactions
  - **Output**: user-bundle interaction probability

App Bundle

Suit Bundle

Game Bundle

Recommender Systems



- Objective: accuracy (the most important)
  - Generally, it can be understood the ***whether the recommended items match with ground truth***
  - Top-K metrics
    - Hit Ratio (HR), Recall, NDCG, MRR, etc.
  - More metrics
    - AUC, GAUC, LogLoss, etc.
  - Most existing recommender systems are designed towards achieving high recommendation accuracy
    - *High accuracy → high CTR/CVR*

      → better user experience and higher business profit

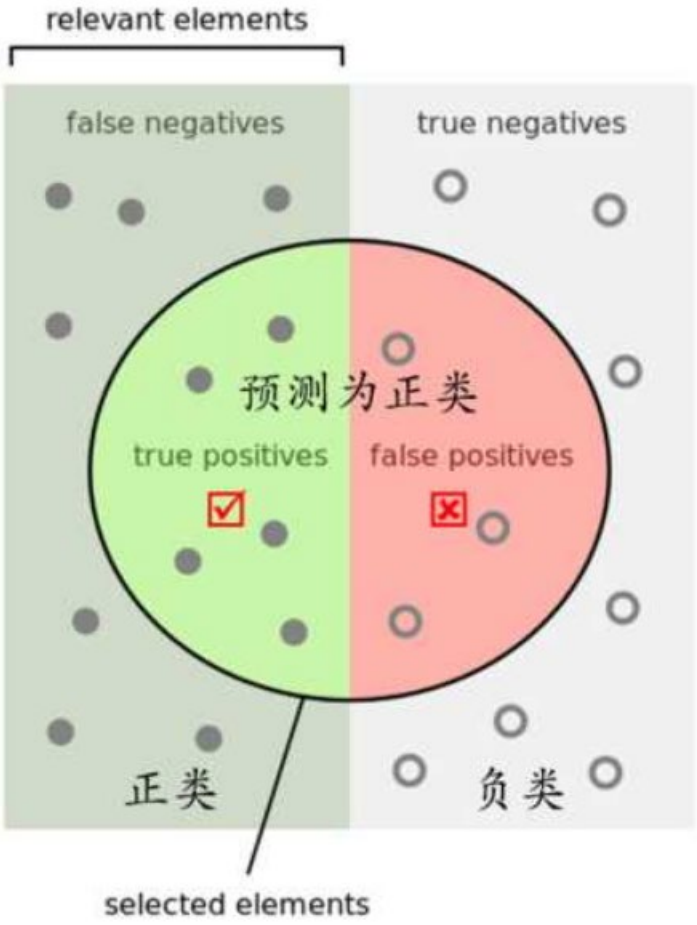
Evaluation Metrics

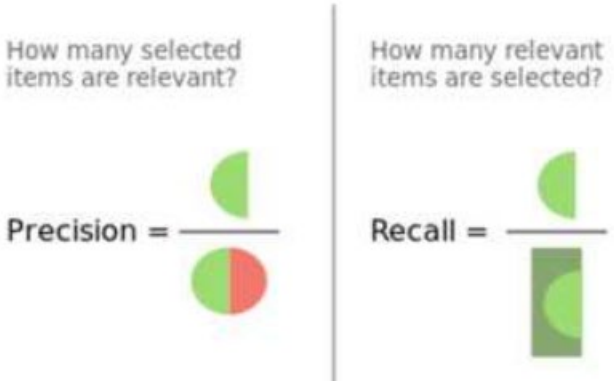
- **Recall** attempts to answer the following question:
  - What proportion of actual positives was identified correctly?

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Precision** attempts to answer the following question:
  - What proportion of positive identifications was actually correct?

$$\text{Precision} = \frac{TP}{TP + FP}$$





- Objective: diversity
  - Recommend diverse items to user while keeping high recommendation accuracy
  - Motivation: only pursuing high accuracy

    → the recommendation list become redundant

    → user can only be recommended certain categories of item
  - Metrics (always defined on **item category**)
    - Gini, entropy, coverage, etc.
    - Accuracy should be also considered of course

accurate but redundant

accurate and diverse

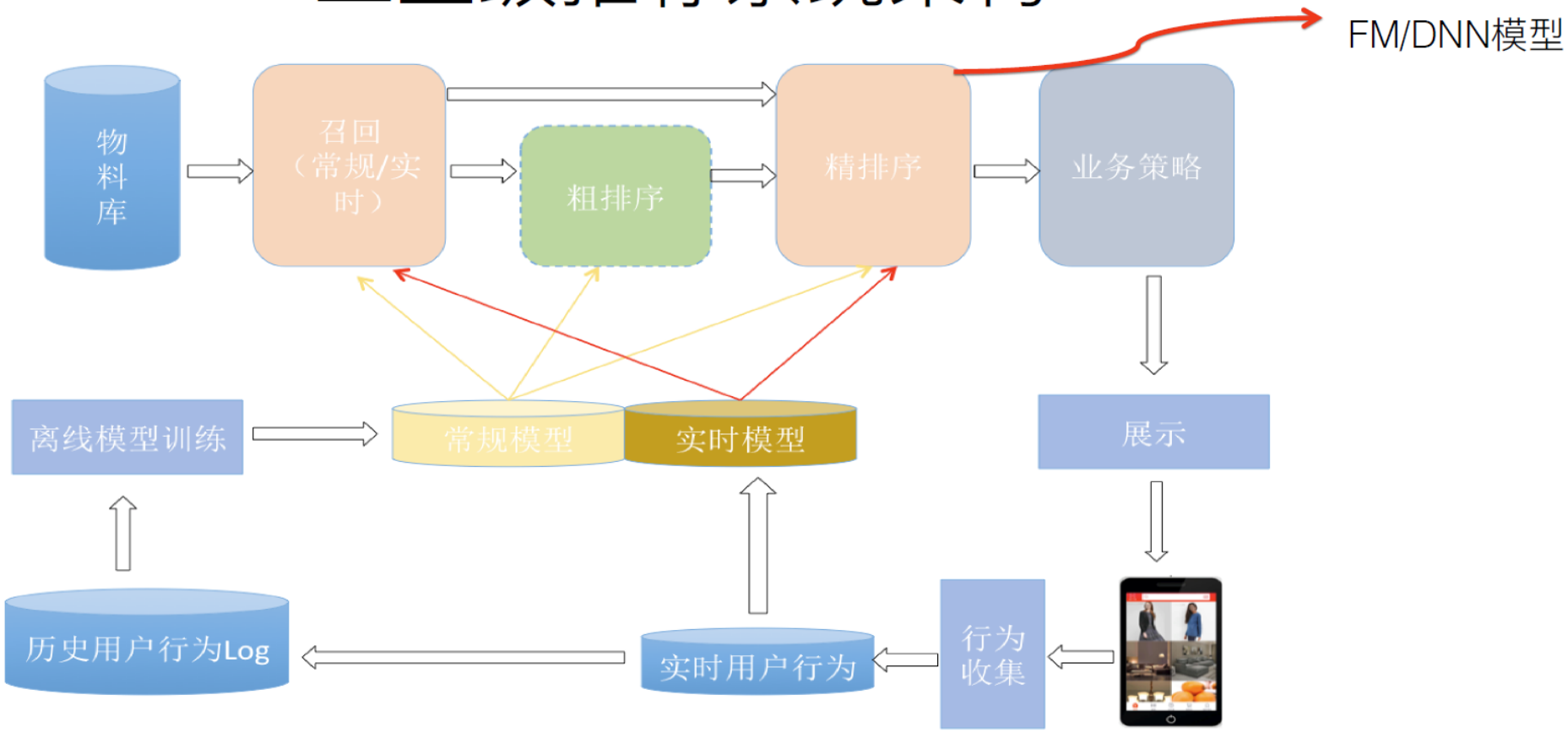
- Objective: explainability
  - What to explain
    - Two folds: explain 1) the model or 2) recommendation results
  - How to explain the model
    - Design explainable model
    - Such as attention modules, explicit feature-interaction, etc.
  - How to explain the results
    - User/Item-based explanation (CF effect / Social-trust)
    - Textual explanation (such as key words in reviews)
    - Knowledge-graph based explanation (via meta-path in KG)

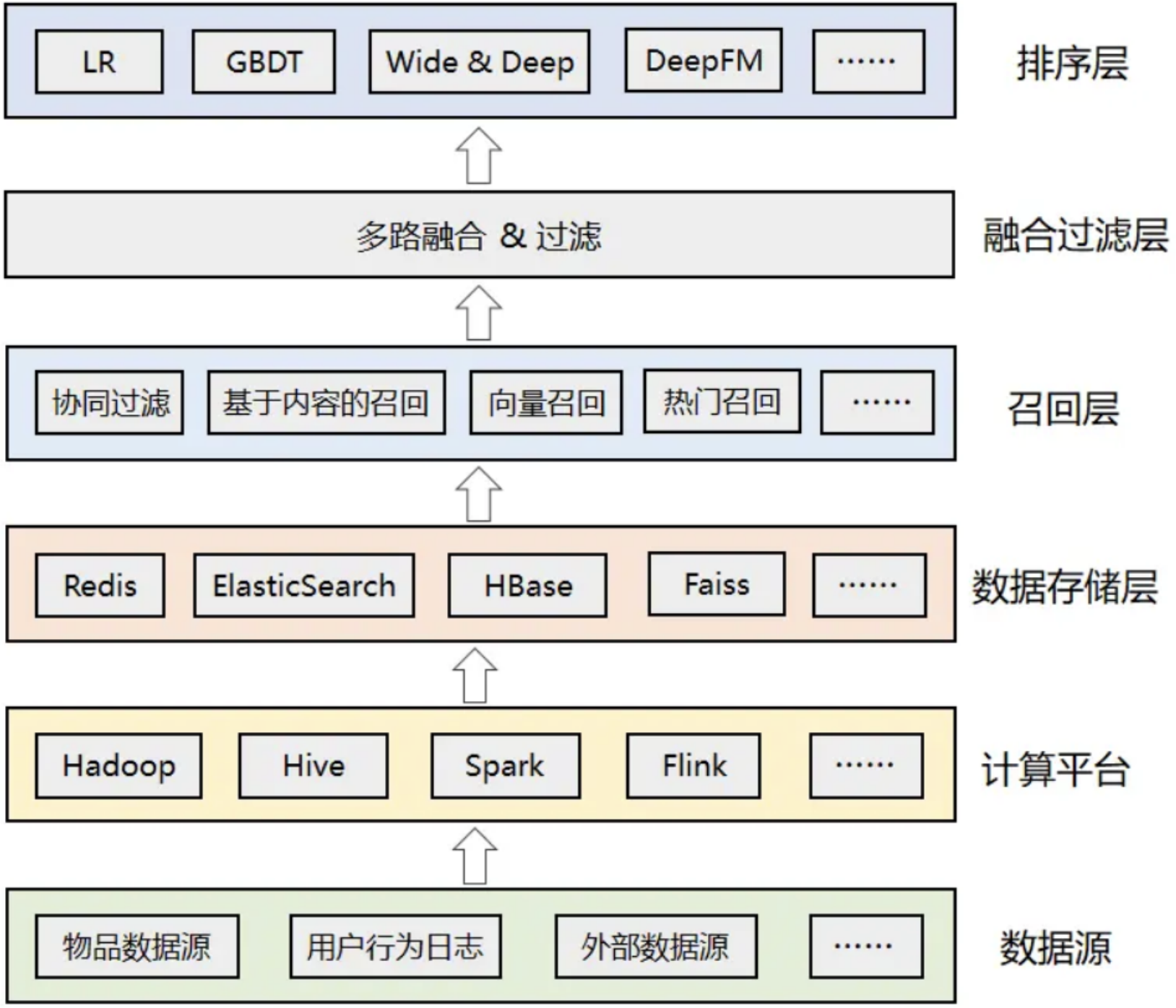
- Objective: fairness
  - Motivation: users' demand on to be fairly treated by RecSys

  - For item-side, it can also be similarity defined
    - "Does items fairly recommended/exposed/assigned to target users?"

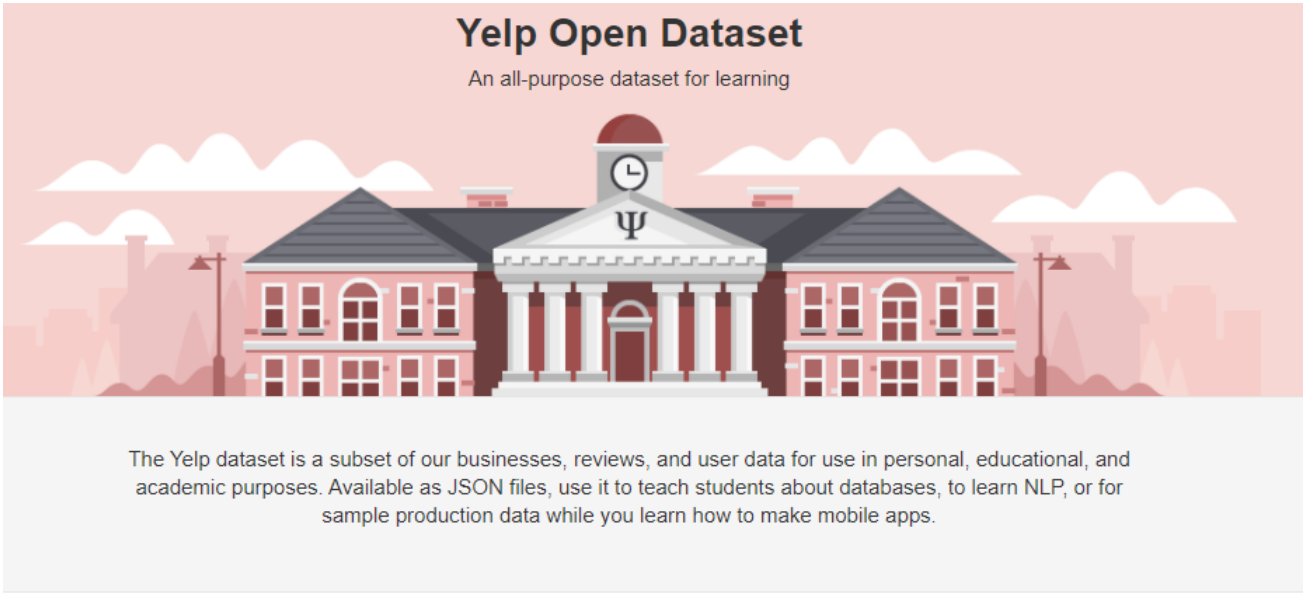
- Objective: privacy
  - When and where the privacy is highly concerned
    - Data collection: recommender may be the attacker
    - Data/model sharing: target company may be the attacker
    - Model/Results public-release: any third-party may be the attacker
  - Representative solutions
    - Transferring/sharing nosensitive model parameters
      - Distributed machine learning model
      - Sharing item-side information
    - Data protection mechanism
      - Data perturbations such as differential privacy-based ones
    - Federated learning

工业级推荐系统架构


FM/DNN模型
物料库
召回（常规/实时）
粗排序
精排序
业务策略
离线模型训练
常规模型
实时模型
展示
历史用户行为Log
实时用户行为
行为收集


LR
GBDT
Wide & Deep
DeepFM
……
排序层
多路融合 & 过滤
融合过滤层
协同过滤
基于内容的召回
向量召回
热门召回
……
召回层
Redis
ElasticSearch
HBase
Faiss
……
数据存储层
Hadoop
Hive
Spark
Flink
……
计算平台
物品数据源
用户行为日志
外部数据源
……
数据源

The Yelp Challenge



- https://www.yelp.com/dataset/documentation/main
- https://www.yelp.com/biz/garaje-san-francisco

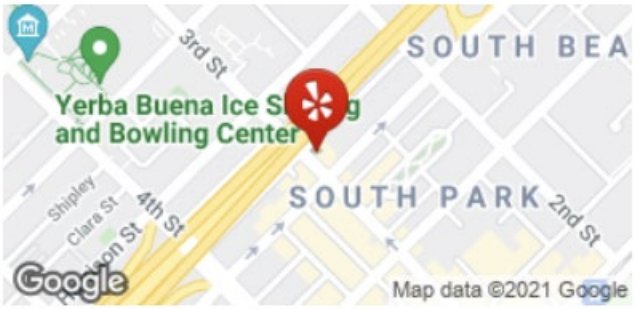
Garaje

1625 reviews

Claimed • $$ • Mexican, Burgers, Beer Bar Edit

Closed 12:00 PM - 2:00 PM, 5:00 PM - 9:00 PM Hours updated 3 weeks ago

Location & Hours



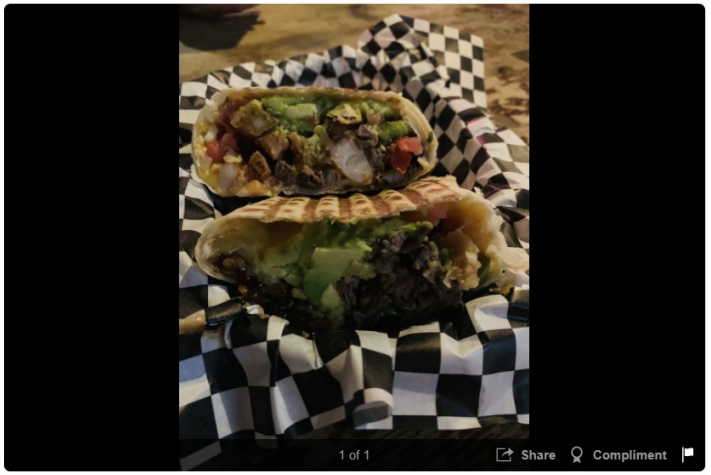
475 3rd St
San Francisco, CA 94107
James Lick Fwy &
Stillman St
South Beach

Get directions

Mon	12:00 PM - 2:00 PM 5:00 PM - 9:00 PM	Closed now
Tue	12:00 PM - 2:00 PM 5:00 PM - 9:00 PM	
Wed	12:00 PM - 2:00 PM 5:00 PM - 9:00 PM	
Thu	12:00 PM - 2:00 PM 5:00 PM - 9:00 PM	
Fri	12:00 PM - 2:00 PM 5:00 PM - 9:00 PM	
Sat	5:00 PM - 9:00 PM	
Sun	Closed	

Garaje > Photos & Videos

1 photo from Alex N. of Garaje



Alex N. Elite 2021
San Francisco, CA
163 30 35

10/12/2021

1 photo

I came at 8pm on a Monday with a group of 7. There was no line to order, however it took 30 minutes for everyone to get their food. Considering we all ordered zapatos and it wasn't exactly crowded there, this felt way too long.

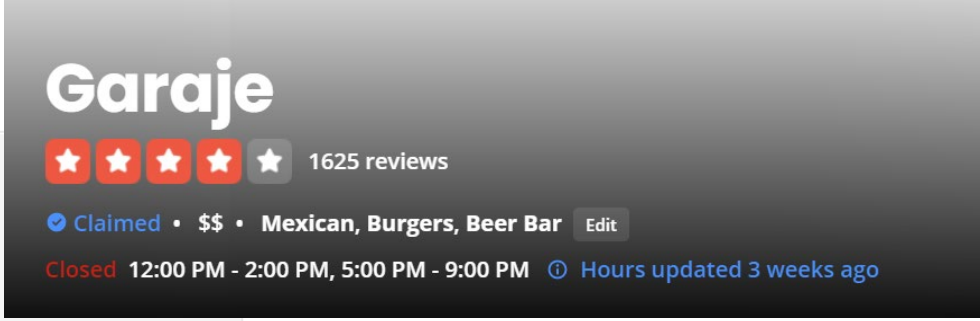
Food:
Rolls Royce Zapato - Unfortunately, it sounded more exciting than it actually was. The steak and prawns were cooked well but lacked flavor. A zapato is a "burrito without rice", but California burritos never have rice anyway. This is essentially a surf n turn cali burrito, but without the sauces you normally get. The pico is not enough when there are fries in there and it tasted rather dry

1995 The OG Zapto - I enjoyed this a lot more than the rolls royce. After eating they dry rolls royce, the OG with its bbq sauce tasted great. The carne asada again either doesnt have that much flavor or it's overshadowed by the bbq sauce.

Overall, the food was mediocre and the wait for food was long. It's fine if you're here to watch sporting events, but it's a pass on casual eating.

business.json

Contains business data including location data, attributes, and categories.




```
{
    // string, 22 character unique string business id
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // string, the business's name
    "name": "Garaje",

    // string, the full address of the business
    "address": "475 3rd St",

    // string, the city
    "city": "San Francisco",

    // string, 2 character state code, if applicable
    "state": "CA",

    // string, the postal code
    "postal code": "94107",

    // float, latitude
    "latitude": 37.7817529521,

    // float, longitude
    "longitude": -122.39612197,

    // float, star rating, rounded to half-stars
    "stars": 4.5,

    // integer, number of reviews
    "review_count": 1198,

    // integer, 0 or 1 for closed or open, respectively
    "is_open": 1,

    // object, business attributes to values. note: some attribute values m
    "attributes": {
        "RestaurantsTakeOut": true,
        "BusinessParking": {
            "garage": false,
            "street": true,
            "validated": false,
            "lot": false,
            "valet": false
        },
    },

    // an array of strings of business categories
    "categories": [
        "Mexican",
        "Burgers",
        "Gastropubs"
    ],

    // an object of key day to value hours, hours are using a 24hr clock
    "hours": {
        "Monday": "10:00-21:00",
        "Tuesday": "10:00-21:00",
        "Friday": "10:00-21:00",
        "Wednesday": "10:00-21:00",
        "Thursday": "10:00-21:00",
        "Sunday": "11:00-18:00",
        "Saturday": "10:00-21:00"
    }
}
```

Alex N. Elite 2021
San Francisco, CA
163 30 35

10/12/2021

1 photo

I came at 8pm on a Monday with a group of 7. There was no line to orde
minutes for everyone to get their food. Considering we all ordered zapat
crowded there, this felt way too long.

Food:
Rolls Royce Zapato - Unfortunately, it sounded more exciting than it actu
prawns were cooked well but lacked flavor. A zapato is a "burrito withou
burritos never have rice anyway. This is essentially a surf n turn cali burr
you normally get. The pico is not enough when there are fries in there a

1995 The OG Zapto - I enjoyed this a lot more than the rolls royce. After
the OG with its bbq sauce tasted great. The carne asada again either doe
or it's overshadowed by the bbq sauce.

Overall, the food was mediocre and the wait for food was long. It's fine i
sporting events, but it's a pass on casual eating.

review.json

Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

```
{
    // string, 22 character unique review id
    "review_id": "zdSx_SD6obEhz9VrW9uAWA",

    // string, 22 character unique user id, maps to the user in user.json
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // integer, star rating
    "stars": 4,

    // string, date formatted YYYY-MM-DD
    "date": "2016-03-09",

    // string, the review itself
    "text": "Great place to hang out after work: the prices are decent, and

    // integer, number of useful votes received
    "useful": 0,

    // integer, number of funny votes received
    "funny": 0,

    // integer, number of cool votes received
    "cool": 0
}
```

Alex N.

From San Francisco, CA

163 Friends · 30 Reviews · 35 Photos

Elite 2021 What is Yelp Elite?

Add friend
Compliment
Send message
Follow Alex N.
Similar Reviews

Alex's Profile

- Profile Overview
- Friends
- Reviews
- Business Photos
- Compliments
- Bookmarks
- Collections

Report this profile

Reviews

Sort by: Date

Black Sugar - Temp. CLOSED
Tea Rooms, Bubble Tea, Coffee & Tea
320 O'Farrell St
San Francisco, CA 94102

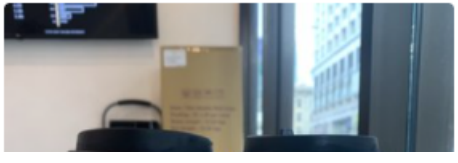
10/27/2021

This place uses plastic straws in case that is a factor for anyone. The cups seal pretty well even though it is a lid and you can shake it pretty rigorously with minimal leakage. I came at 4pm on a Tuesday and got my drink within 2 minutes of ordering.

House milk tea (75% sweet, less ice) -
* 2 stars
This tea tasted more floral with some chocolatey notes. If i had to guess, I'd say this was some combination of teas. My main complaint with this drink is that it was not creamy enough. While the tea flavor was decently strong, it definitely felt like there was not enough milk or creamer added.

Sesame Matcha Latte (75% sweet, less ice) -
* 4 stars
I'm a huge fan of black sesame and this tasted fantastic. While the sesame flavor is strong, it overpowers the matcha and I couldn't really make it out in the drink.

Forewarned, black sesame drinks, including this one, usually have a good amount of pulp in it and is not totally smooth. There were a good amount of the sesame bits still in the drink it was almost thick like a milkshake. Overall, i still thoroughly enjoyed this drink and would recommend the black sesame fans out there to give it a shot!



About Alex N.

Rating Distribution

5 stars 3
4 stars 14
3 stars 11
2 stars 2
1 star 0

View more graphs

Review Votes

Useful **42**
Funny **7**
Cool **16**

Stats

Bookmarks **12**
Followers **1**

4 Compliments

1 · 2 · 1

Location

San Francisco, CA

Yelping Since

June 2014

user.json

User data including the user's friend mapping and all the metadata as

```
{
    // string, 22 character unique user id, maps to the
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, the user's first name
    "name": "Sebastien",

    // integer, the number of reviews they've written
    "review_count": 56,

    // string, when the user joined Yelp, formatted lik
    "yelping_since": "2011-01-01",

    // array of strings, an array of the user's friend
    "friends": [
        "wqoXYLWmpkEH0YvTmHBsJQ",
        "KUXLLiJGrjtSsapmxmpvTA",
        "6e9rJKQC3n0RSKyHLViL-Q"
    ],

    // integer, number of useful votes sent by the user
    "useful": 21,

    // integer, number of funny votes sent by the user
    "funny": 88,

    // integer, number of cool votes sent by the user
    "cool": 15,

    // integer, number of fans the user has
    "fans": 1032,

    // array of integers, the years the user was elite
    "elite": [
        2012,
        2013
    ],

    // float, average rating of all reviews
    "average_stars": 4.31,

    // integer, number of hot compliments received by the user
    "compliment_hot": 339,

    // integer, number of more compliments received by the user
    "compliment_more": 668,

    // integer, number of profile compliments received by the user
    "compliment_profile": 42,

    // integer, number of cute compliments received by the user
    "compliment_cute": 62,

    // integer, number of list compliments received by the user
    "compliment_list": 37,

    // integer, number of note compliments received by the user
    "compliment_note": 356,

    // integer, number of plain compliments received by the user
    "compliment_plain": 68,

    // integer, number of cool compliments received by the user
    "compliment_cool": 91,

    // integer, number of funny compliments received by the user
    "compliment_funny": 99,

    // integer, number of writer compliments received by the user
    "compliment_writer": 95,

    // integer, number of photo compliments received by the user
    "compliment_photos": 50
}
```

checkin.json

Checkins on a business.

```
{
    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg"

    // string which is a comma-separated list of timestamps for each checki
    "date": "2016-04-26 19:49:16, 2016-08-30 18:36:57, 2016-10-15 02:45:18,
}
```

tip.json

Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

```
{
    // string, text of the tip
    "text": "Secret menu - fried chicken sando is da bombbbbbb Their zapato

    // string, when the tip was written, formatt
    "date": "2013-09-20",

    // integer, how many compliments it has
    "compliment_count": 172,

    // string, 22 character business id, maps to
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // string, 22 character unique user id, maps
    "user_id": "49JhAJh8vSQ-vM4Aourl0g"
}
```

photo.json

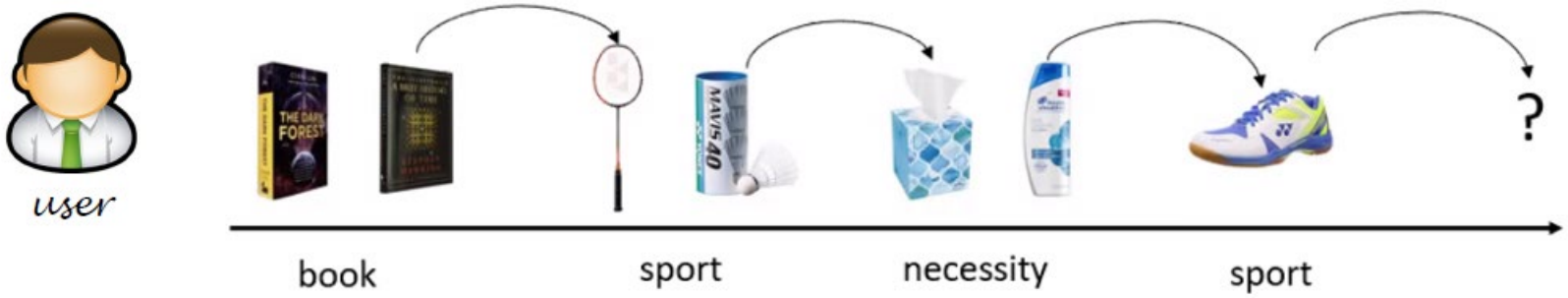
Contains photo data including the caption and classification (one of "food", "drink", "menu", "inside" or "outside").

```
{
    // string, 22 character unique photo id
    "photo_id": "_nN_DhLXkfwEkwPNxne9hw",
    // string, 22 character business id, maps to business in business.json
    "business_id" : "tnhfDv5Il8EaGSXZGiuQGg",
    // string, the photo caption, if any
    "caption" : "carne asada fries",
    // string, the category the photo belongs to, if any
    "label" : "food"
}
```

讨论并设计预测模型

- 假设我们拿到了10000个不同用户的Yelp点评数据，经过处理后发现，这些点评记录的时间跨度为2年，每个用户都有50条店铺消费记录，依次按照时间顺序排列，每笔消费都写了评论并上传了图片。

- 问题：预测用户之后会去哪家店铺消费并为每个用户给出推荐店铺。



提示和建议

- 1. 把问题先定义好 (next-item prediction, next-k?)
- 2. 数据集划分
- 3. 特征嵌入
- 4. 预测模型
- 5. 训练
- 6. 评估和测试 （评价指标）