

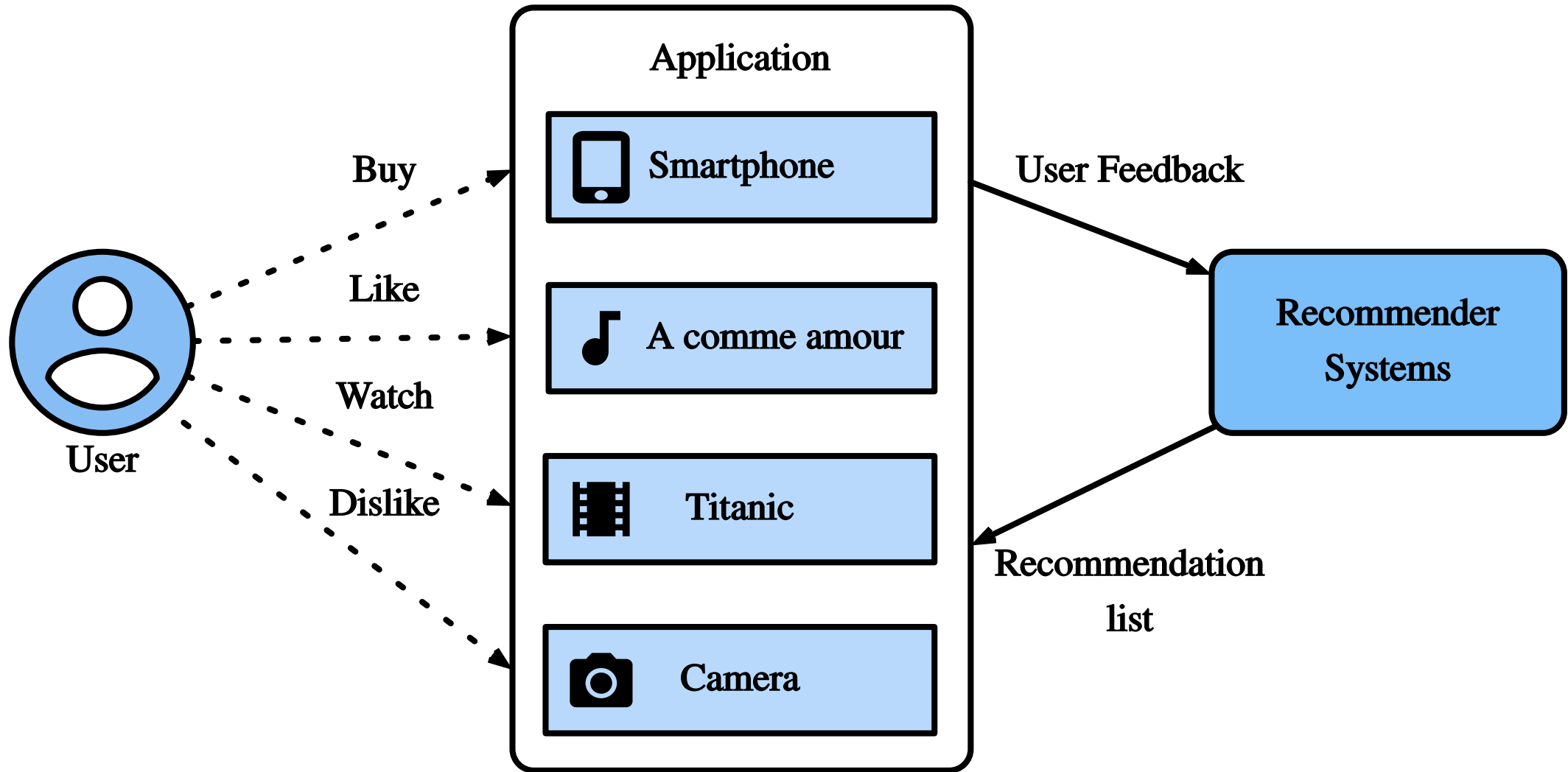
COMP 4332 / RMBI 4310

Big Data Mining (Spring 2022)

Project 3 Rating Prediction

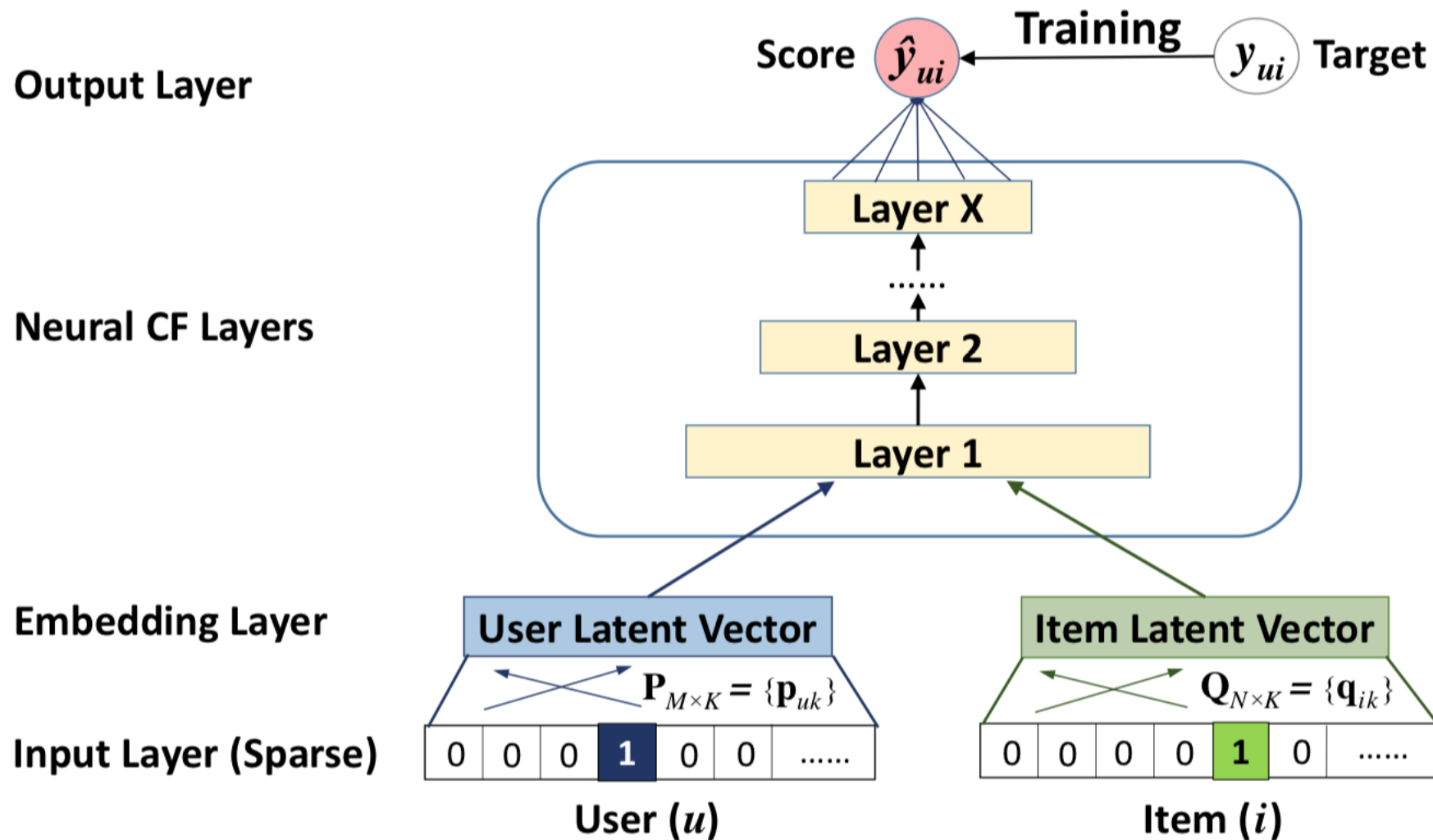
TA: Jiaxin Bai (jbai@connect.ust.hk)

Recommendation Systems



In Previous Tutorial

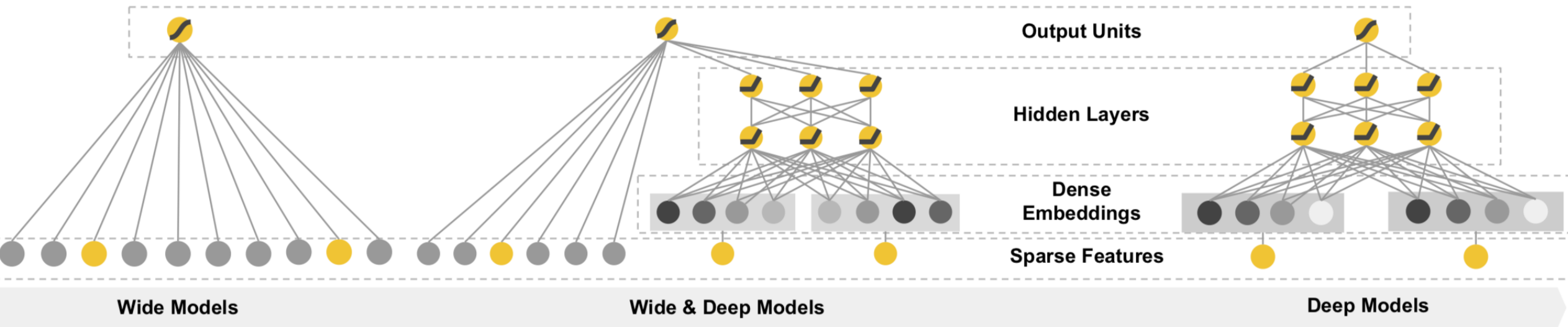
Neural CF



Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu and Tat-Seng Chua (2017). [Neural Collaborative Filtering](#). In Proceedings of WWW '17, Perth, Australia, April 03-07, 2017.

In Previous Tutorial

Wide & Deep Learning










Memorization

Generalization

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM, 7–10.

Rating Prediction

- Predict users' ratings on items given some known ratings. The prediction would be evaluated by Root Mean Squared Error (RMSE)

	i_1	i_2	i_3	i_4	i_5	i_6
 U1	4	?	3	?	5	?
 U2	?	2	?	?	4	1
 U3	?	?	1	?	2	5
 U4	?	?	3	?	?	1
 U5	1	4	?	?	2	5
 U6	5	?	2	1	?	4
 U7	?	2	3	?	4	5

Dataset

- User ratings
- Extra user information
- Extra business information

User ratings:

	user_id	business_id	stars
0	ec8f38aa91755dcf5837020d022ad384	ecaa90564e18dca1c7b653038f71d6bf	1.0
1	64fe4dd0a489c9b96a3e8d7fbd337888	ef118bb0ae1fc369e1f47d1b34f6acee	5.0
2	a49909b39426ebb3538aa837b5b88840	e8b182a923810d52981aa02d56dde799	5.0
3	a56726d5676d647e42e2aca54f21b075	250040e979eae9ef5912aa5a1d285e4e	5.0
4	3e19d8260e655ba87bea0922bac92266	e02880faf4d42fe1df7bd370fb1c787b	4.0

Extra user information

Techniques for using this information through Wide and Deep Learning model will be introduced in tutorial 8

```
{
  "average_stars":3.63,
  "compliment_cool":1,
  "compliment_cute":0,
  "compliment_funny":1,
  "compliment_hot":1,
  "compliment_list":0,
  "compliment_more":0,
  "compliment_note":0,
  "compliment_photos":0,
  "compliment_plain":0,
  "compliment_profile":0,
  "compliment_writer":0,
  "cool":16,
  "elite": "",
  "fans":4,
  "funny":22,
  "name":"Jenna",
  "review_count":33,
  "useful":48,
  "user_id":"88422913727e71e88611fdfe3512fa03",
  "yelping_since":"2013-02-21 22:29:06"
}
```


Extra business information

Techniques for using this information through Wide and Deep Learning model will be introduced in tutorial 8

```
{  
  "address": "4075 S Durango Dr, Ste 105B",  
  "attributes": {  
    "business_id": "c7d693d13177b9839d89f277e5280315",  
    "categories": "Mobile Phones, Mobile Phone Repair, Shopping",  
    "city": "Las Vegas",  
    "hours": {  
      "is_open": 1,  
      "latitude": 36.115305,  
      "longitude": -115.280737,  
      "name": "Computer Doctor BG",  
      "postal_code": "89147",  
      "review_count": 211,  
      "stars": 5.0,  
      "state": "NV"  
    }  
  }  
}
```

We provide:

- Rating data (rating scale is 1.0-5.0) :
 - 'train.csv' : 60080 ratings
 - 'valid.csv' : 7510 ratings
 - 'test.csv' : 7510 ratings (entries of 'stars' column in 'test.csv' are all set to 0.0)
- User information :
 - 'user.csv': 2980 users
- Business information
 - 'business.csv': 5964 businesses
- Code for evaluating predictions: 'evaluate.py'

Submission

- Predictions on **test data** (please make sure you can successfully evaluate your validation predictions on the validation data with the help of evaluate.py)
- Report (1~2 pages)
- Code (Frameworks and even programming languages are not restricted.)
- DDL: 11:59 pm, May 23, 2022
- Submission:
 - Each **team leader** is required to submit the groupNo.zip file that contains pre.csv and your team's code on canvas.
 - Each **student** is required to submit **his/her own project report individually** (All members in a group **can choose to submit the same project report**. But the **submission still need to be done individually**)
- we will check your report with your code and the RMSE.

Grading Rule

Grade	Model (80%)	Report (20%)	Baseline (RMSE on test set)
60%		submission	1.20
80%	an easy baseline that most students can outperform	detailed explanation	1.15
90%	a competitive baseline that about half students can surpass	detailed explanation and analysis	1.12
100%	a very competitive baseline	excellent visualization and analysis	1.09

Other information:

1. You are welcome to use any methods to make the prediction.
2. The methods taught in the class/tutorial (including previous ones) + some parameter tuning + some feature engineering are enough for you to get the full marks.
3. Late submission policy is the same as project 1.
4. Peer evaluation is **not** required.

Thank You