

# Hybrid Recommendation System

Chien-Te Lee

2020/8/20

# Outline

---

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

# Outline

---

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

# Problem definition

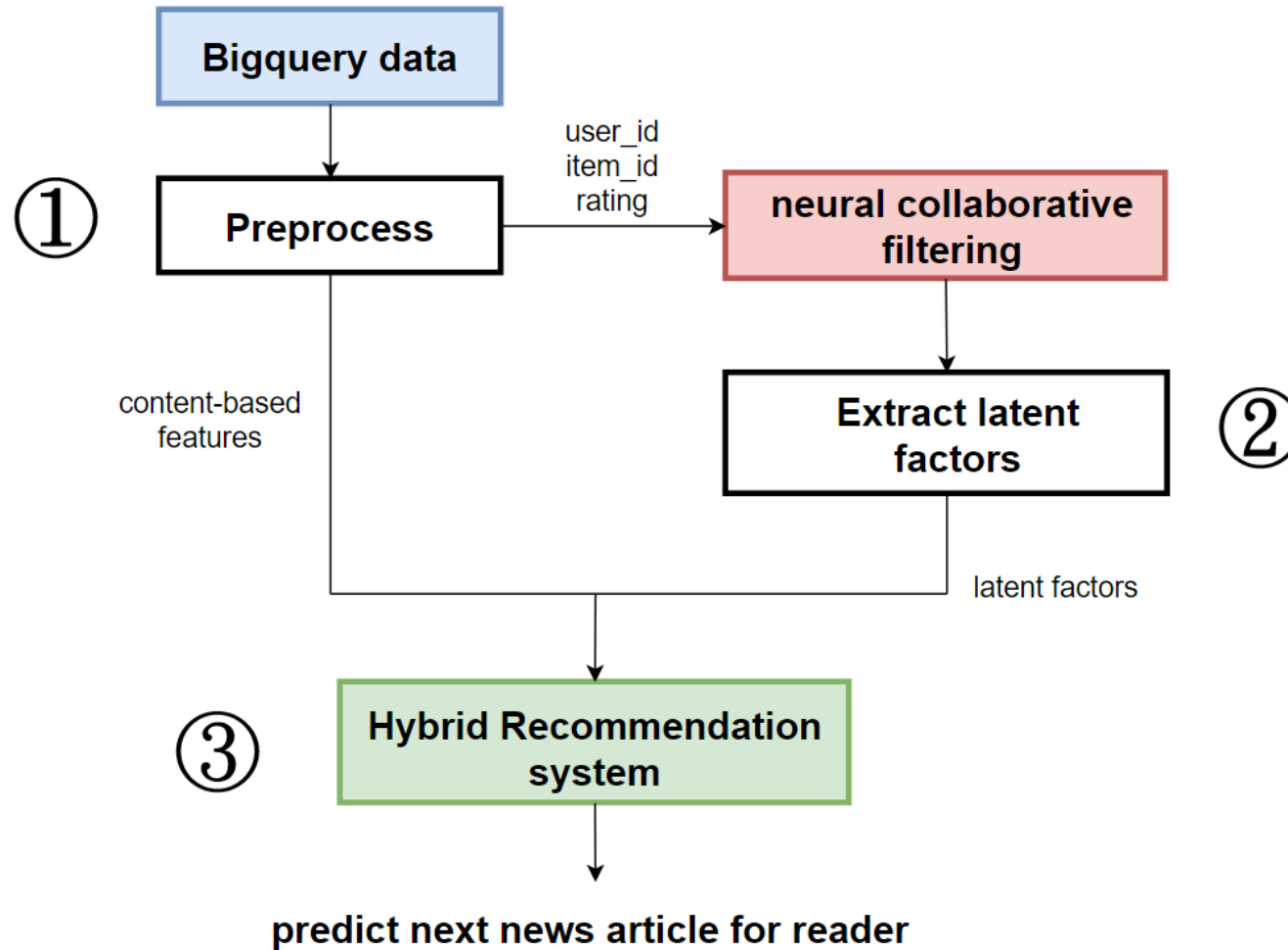
---

- Given a reader is reading an article on the news website, how to figure out what the reader would like to read for the next article?



# Workflow

- Preprocess dataset → Extract latent factors → Train hybrid model



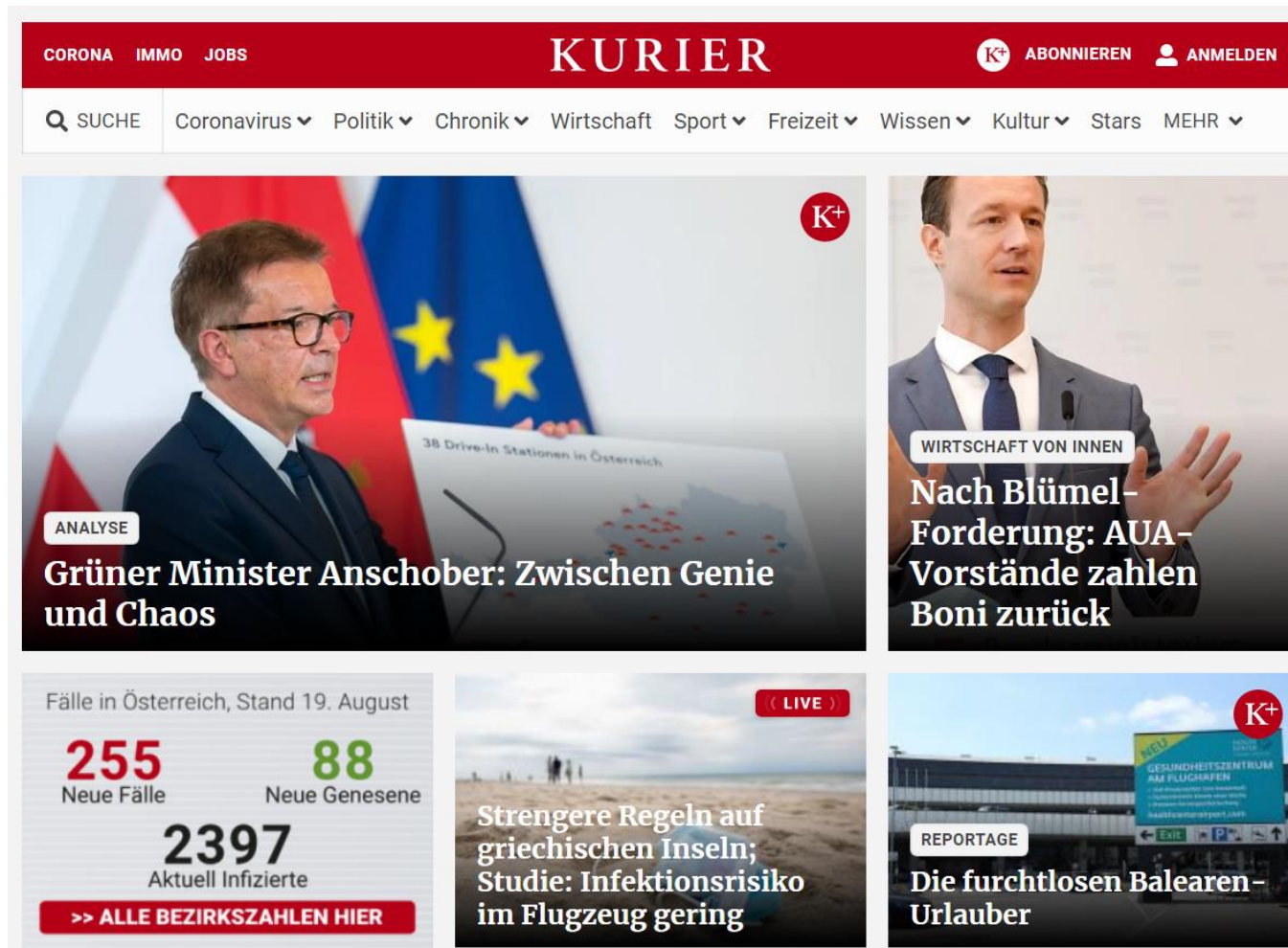
# Outline

---

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

# Preprocess BigQuery dataset

- “cloud-training-demos.GA360\_test.ga\_sessions\_sample” is the Google Analytic data from Austrian news website Kurier.at.



# Preprocess BigQuery dataset

- Use standard SQL to query the public BigQuery dataset, and select customDimensions as content-based features.

The screenshot displays the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'hybrid-recsys-gcp', and a search bar. The left sidebar contains navigation links for BigQuery, FEATURES & INFO, and SHORTCUT. The main area is divided into a Query editor and Query results section.

**Query editor:** The query is written in SQL and is as follows:

```
1 WITH user_count_table AS (  
2   SELECT  
3     fullVisitorId as user_id,  
4     Count(*) - 1 AS interactions  
5  
6   FROM cloud-training-demos.GA360_test.ga_sessions_sample,  
7     UNNEST(hits) AS hits  
8 )
```

Below the query editor, there are buttons for Run, Save query, Save view, Schedule query, and More. A status message indicates: "This query will process 432.1 MB when run."

**Query results:** The results section shows the query is complete (7.1 sec elapsed, 432.1 MB processed). The results are displayed in a table with the following columns: Row, user\_id, item\_id, title, author, category, device\_brand, article\_year, and article\_mon.

Row	user_id	item_id	title	author	category	device_brand	article_year	article_mon
1	1000163602560555666	299853016	Schröcksnadel gegen Werdenigg: Keine Aussprache	unknown	News	Apple	2017	11
2	1000163602560555666	298888038	Investment kann jetzt jeder!	unknown	News	Apple	2017	11
3	1000163602560555666	299814775	Meghan Markle: Verlobungsring aus Dianas Brosche	Maria Zelenko	Lifestyle	Apple	2017	11
4	1000163602560555666	299772450	Kritik an Prinz Harry: "Ein verzogener Rotzlöffel"	Elisabeth Spitzer	Stars & Kultur	Apple	2017	11
5	1000163602560555666	299824032	YouTube: Schwere Probleme mit verstörenden Kindervideos	Georg Leyrer	Stars & Kultur	Apple	2017	11
6	1000163602560555666	299809748	Glamour-Auftritt von Charlene & Albert bei Rugby Gala	Christina Michlits	Stars & Kultur	Apple	2017	11
7	1000163602560555666	299826775	Auf Bank ausgeruht: Pensionist muss Strafe zahlen	Marlene Patsalidis	Lifestyle	Apple	2017	11
8	1000163602560555666	299907204	Britische Urlauberin filmte in Australien eigenen Krokodil-Angriff	unknown	News	Apple	2017	11
9	1000163602560555666	299907204	Britische Urlauberin filmte in Australien eigenen Krokodil-Angriff	unknown	News	Apple	2017	11
10	1000163602560555666	299933565	Koalitionsverhandler vor Konsens bei Krankenkassen-Einigung	Peter Tomel	News	Apple	2017	11

At the bottom, there is a pagination bar showing "Rows per page: 100" and "1 - 100 of 155720".



# Preprocess BigQuery dataset

- Selected features:

	user_id	item_id	title	author	category	device_brand	article_year	article_month	rating	next_item_id	fold
0	1000196974485173657	299910994	Direktorensprecherin Isabella Zins: So könnte ...	Ute Brühl	News	unknown	2017	11	1.000000	299899819	0
1	1000196974485173657	299930679	Wintereinbruch naht: Erster Schnee im Osten mö...	Daniela Wahl	News	unknown	2017	11	1.000000	299972194	0
2	1004209053768679755	18976804	Heimskandal - Brigitte Wanker: Die Landesverrä...	Georg Hönigsberger	News	Huawei	2013	7	1.000000	299695400	0
3	1004555043399129313	299837992	Das erste TV-Interview von Prinz Harry & Megha...	Christina Michlits	Stars & Kultur	unknown	2017	11	0.979912	299824032	0
4	1004555043399129313	299836841	ÖVP will Studiengebühren FPÖ in Verhandlungen...	Raffaella Lindorfer	News	unknown	2017	11	1.000000	299899819	0

- Use  $0.5 \times \text{time} / \text{median\_time}$  as rating
- Use  $\text{ABS}(\text{MOD}(\text{FarmFingerprint}(\text{visitor\_id} + \text{visit\_time}), 10))$  as hash\_id

# Outline

---

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

# Extract Latent Factors

- Collaborative filtering use matrix factorization to split rating matrix into user matrix and item matrix.
- User and item latent factors are variables which represent similarities in high dimensional space.

		Item			
		W	X	Y	Z
User	A		4.5	2.0	
	B	4.0		3.5	
	C		5.0		2.0
	D		3.5	4.0	1.0

Rating Matrix

=

A	1.2	0.8
B	1.4	0.9
C	1.5	1.0
D	1.2	0.8

User Matrix

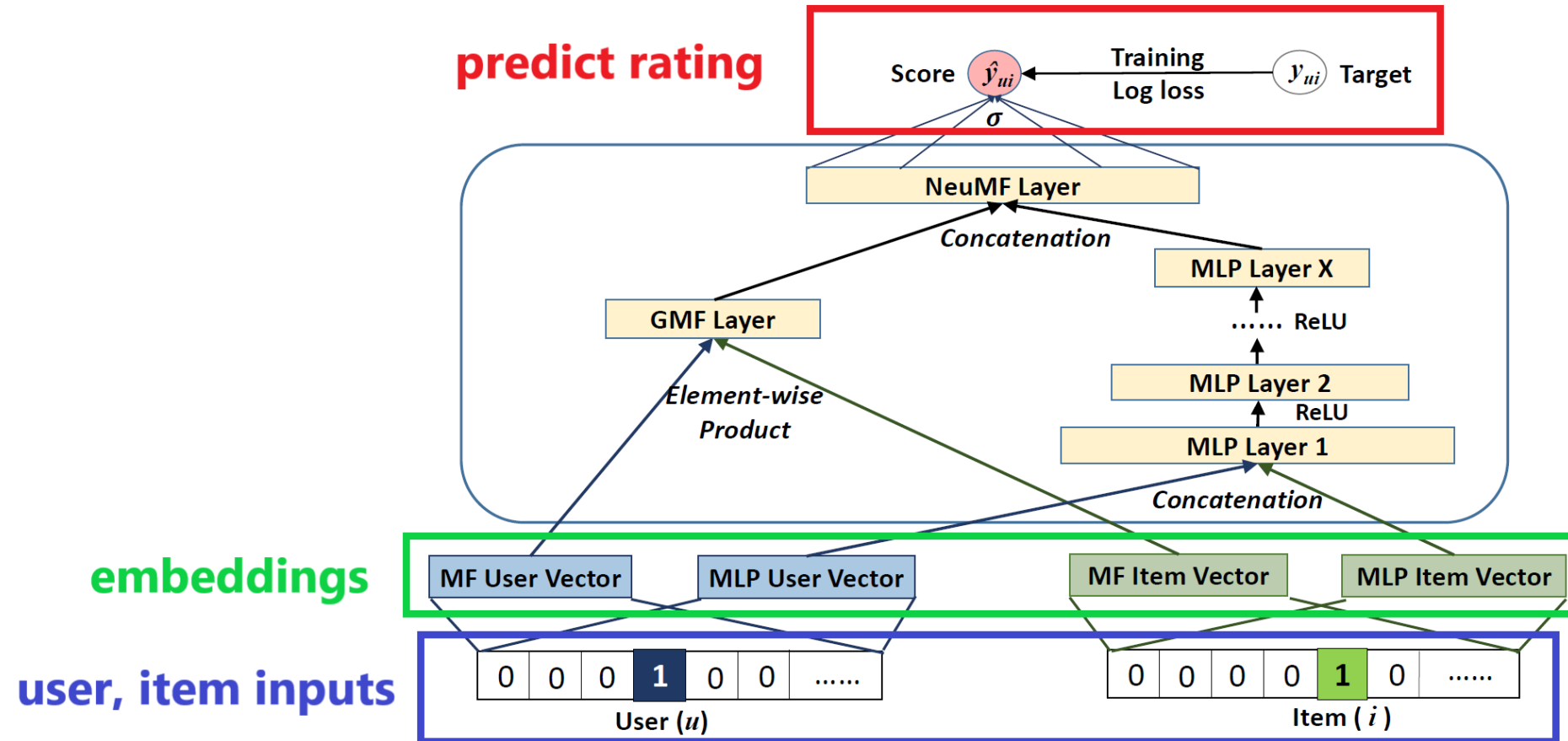
X

	W	X	Y	Z
	1.5	1.2	1.0	0.8
	1.7	0.6	1.1	0.4

Item Matrix

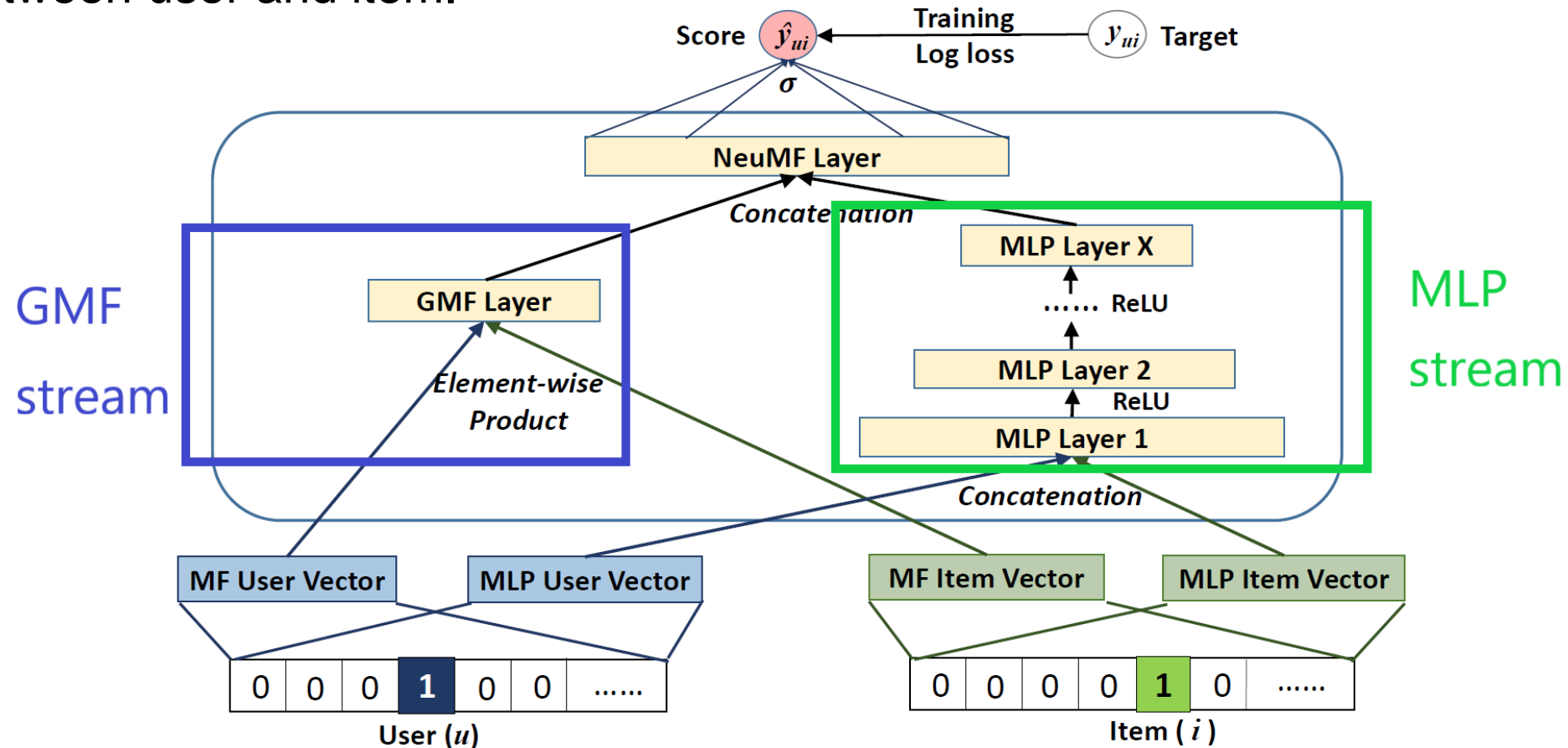
# Extract Latent Factors

- Use neural collaborative filtering to predict rating
- Extract embedding as latent factors



# Extract Latent Factors

- The Generalized Matrix Factorization (GMF) stream represents the matrix factorization.
- The Multi-Layer Perceptron (MLP) stream captures the non-linear relation between user and item.



# Extract Latent Factors

- User latent factors

	user_id	u_latent_0	u_latent_1	u_latent_2	u_latent_3	u_latent_4	u_latent_5	u_latent_6	u_latent_7	u_latent_8	...	u_latent_10
0	1000163602560555666	-0.223890	0.013522	-0.197976	0.217497	-0.053681	-0.006720	-0.117004	0.113265	0.187496	...	-0.043796
1	1000196974485173657	-0.033306	0.020547	0.104502	-0.003414	0.063732	0.086023	-0.062370	0.030699	-0.115149	...	-0.034464
2	1002090131595000997	-0.179897	-0.139295	0.073862	-0.047588	0.047952	-0.000489	0.117391	0.058213	-0.077938	...	-0.012818
3	1002109532017576768	-0.079408	-0.174885	0.014121	-0.081578	0.140167	-0.137453	0.088288	0.162533	-0.106551	...	0.016511
4	1004209053768679755	-0.000192	-0.134218	0.076557	-0.169822	-0.072396	0.000815	-0.026878	-0.070867	0.092746	...	0.002242

- Item latent factors

	item_id	i_latent_0	i_latent_1	i_latent_2	i_latent_3	i_latent_4	i_latent_5	i_latent_6	i_latent_7	i_latent_8	...	i_latent_10	i_latent_11	i_latent_12
0	100170790	-0.044401	-0.054478	-0.024215	-0.095297	0.030977	-0.051534	-0.087727	0.066595	-0.116718	...	0.027045	-0.022293	-0.038569
1	100292889	0.044174	0.018957	-0.020329	0.005043	-0.066686	-0.046977	-0.011907	0.023122	-0.024344	...	0.048059	-0.057492	-0.052943
2	100735153	0.004435	-0.092585	-0.101787	-0.067878	0.077632	0.000198	-0.068222	0.012467	-0.053971	...	-0.001989	0.011519	0.056933
3	100915139	0.009406	0.015461	-0.031027	-0.006515	0.015776	-0.004458	0.006125	-0.020394	-0.046054	...	0.044172	0.006846	-0.025532
4	101092112	-0.063698	0.059048	0.049322	-0.023419	0.039215	0.036990	0.013302	-0.031852	-0.001982	...	0.043176	-0.017732	-0.049217

# Extract Latent Factors

---

- Concatenate latent factors with preprocessed content-based features
- Each specific user\_id matches with specific user\_latent (INNER JOIN)
- Each specific item\_id matches with specific item\_latent (INNER JOIN)

	user_id	item_id	content-based features	user_latent	item_latent
0					
1					
2					

# Outline

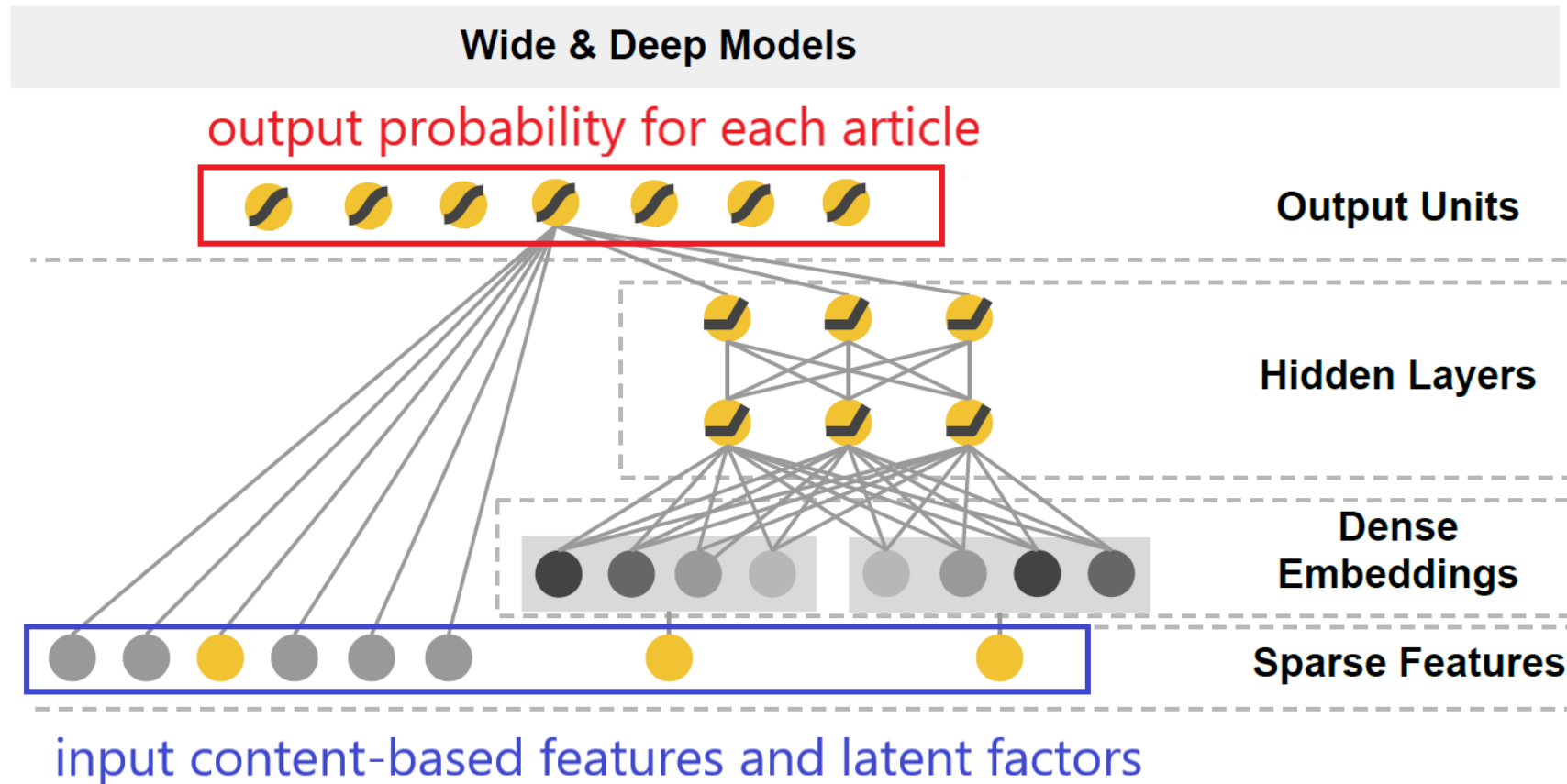
---

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**



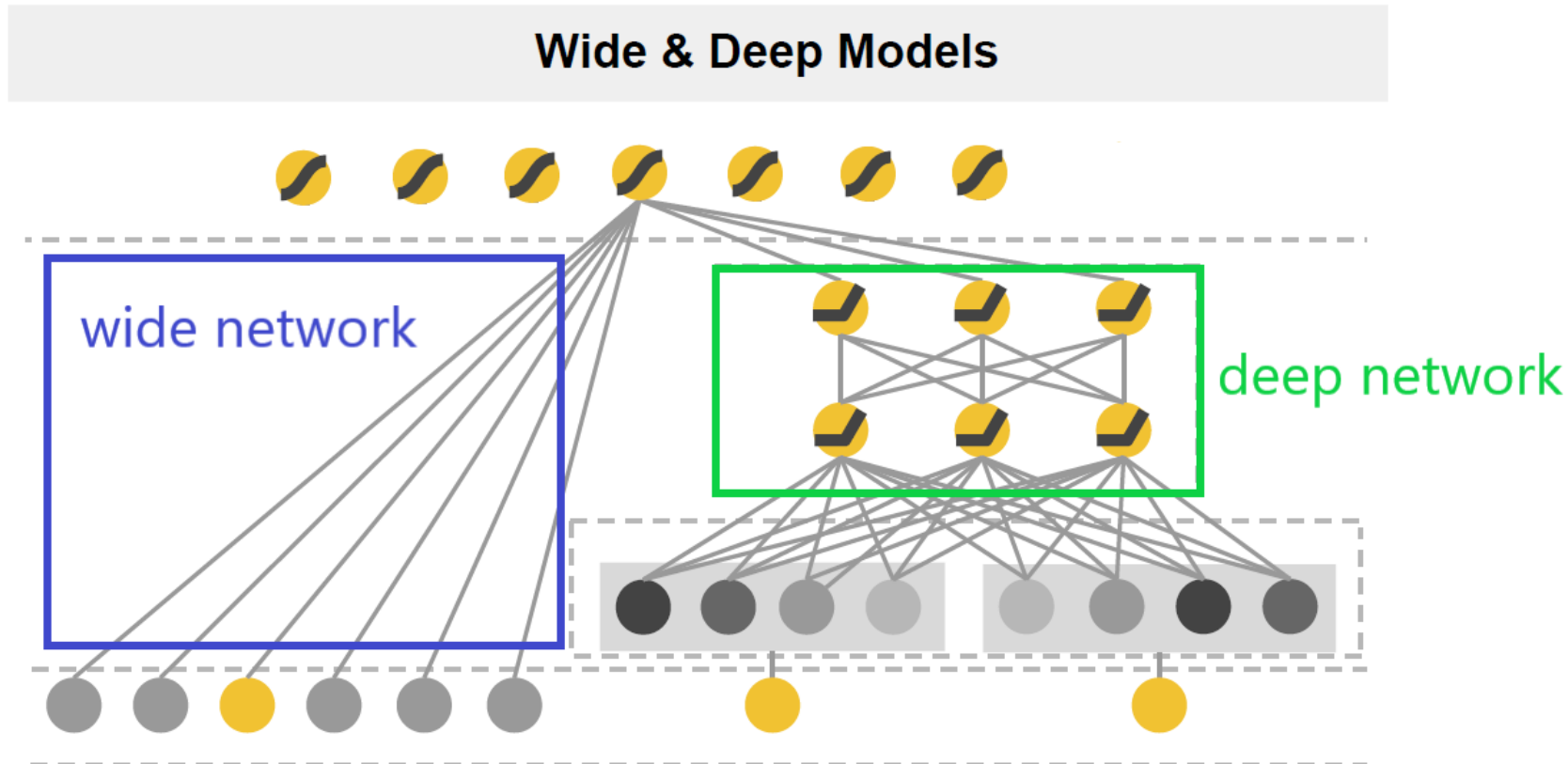
# Hybrid Recommendation System

- Apply wide & deep network for hybrid model.
- Use content-based features and user, item latent factors as input.
- Predict the probability for each article as next item.



# Hybrid Recommendation System

- The deep network takes dense embedding features, and the wide network takes sparse features.
- Dense embedding feature: user\_id, item\_id, author, device\_brand, title (NNLM)
- Sparse feature: author, cross\_date, category, device\_brand



# Hybrid Recommendation System

- Result: Train [bc\_loss: 4.05, acc: 11.85, top\_10\_acc: 47.35]  
Test [bc\_loss: 4.47, acc: 9.71, top\_10\_acc: 42.09]

The screenshot displays the Google Cloud Platform (GCP) console interface for a project named 'hybrid-recsys-gcp'. The top navigation bar includes the GCP logo, project name, a search bar, and user profile information. The left sidebar contains navigation icons for various GCP services. The main content area shows a list of logs for a specific job, filtered by the resource label 'job\_id=hybrid\_recsys\_train\_job\_200816\_180754' and a timestamp range from 2020-08-16T18:07:57Z onwards. The logs are displayed in a table format with columns for timestamp, log level, source, and message. The logs show the progress of training epochs, with metrics such as loss, accuracy, and top-10 accuracy. The training process concludes with a 'Module completed; cleaning up.' message, followed by 'Clean up finished.' and 'Task completed successfully.' messages. The final log entry indicates 'Job completed successfully.' at 2020-08-17 02:30:26.584 HKT. The bottom of the log view shows a message 'No newer entries found matching current filter.' and a 'Load newer logs' button.

Google Cloud Platform hybrid-recsys-gcp Search products and resources

CLASSIC CREATE METRIC CREATE SINK SAVE SEARCH SHOW LIBRARY

1 resource.labels.job\_id="hybrid\_recsys\_train\_job\_200816\_180754" timestamp>="2020-08-16T18:07:57Z"

Submit Filter No limit Jump to now

Showing logs from all time (HKT) Download logs View Options

2020-08-17 02:28:14.430 HKT	master-replica-0	Epoch 25, train[loss: 4.195067, acc: 10.283843, top_10_acc: 44.652580], Test[loss: 4.460537, acc: 8.933158, top_10_acc: 41.486385]
2020-08-17 02:28:14.430 HKT	master-replica-0	Epoch 26, train[loss: 4.184915, acc: 10.303751, top_10_acc: 44.901104], Test[loss: 4.461580, acc: 9.098693, top_10_acc: 41.674755]
2020-08-17 02:28:14.430 HKT	master-replica-0	Epoch 27, train[loss: 4.172222, acc: 10.390444, top_10_acc: 45.134857], Test[loss: 4.462777, acc: 8.990239, top_10_acc: 41.503510]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 28, train[loss: 4.163945, acc: 10.527871, top_10_acc: 45.268433], Test[loss: 4.460226, acc: 9.178606, top_10_acc: 41.760372]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 29, train[loss: 4.152080, acc: 10.687773, top_10_acc: 45.531082], Test[loss: 4.463778, acc: 9.098693, top_10_acc: 41.703293]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 30, train[loss: 4.139294, acc: 10.762908, top_10_acc: 45.841251], Test[loss: 4.465705, acc: 9.167190, top_10_acc: 41.760372]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 31, train[loss: 4.127554, acc: 10.894554, top_10_acc: 45.994091], Test[loss: 4.467805, acc: 9.212854, top_10_acc: 41.726128]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 32, train[loss: 4.120628, acc: 11.023632, top_10_acc: 46.083355], Test[loss: 4.459177, acc: 9.287060, top_10_acc: 42.102859]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 33, train[loss: 4.111029, acc: 11.164268, top_10_acc: 46.300411], Test[loss: 4.461469, acc: 9.269936, top_10_acc: 41.977280]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 34, train[loss: 4.105287, acc: 11.313254, top_10_acc: 46.487926], Test[loss: 4.468490, acc: 9.184315, top_10_acc: 42.177067]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 35, train[loss: 4.092279, acc: 11.333804, top_10_acc: 46.593243], Test[loss: 4.468904, acc: 9.538216, top_10_acc: 42.040070]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 36, train[loss: 4.088718, acc: 11.441690, top_10_acc: 46.761494], Test[loss: 4.470200, acc: 9.355556, top_10_acc: 42.017239]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 37, train[loss: 4.079213, acc: 11.596456, top_10_acc: 46.922039], Test[loss: 4.468232, acc: 9.492551, top_10_acc: 41.943035]
2020-08-17 02:28:14.431 HKT	master-replica-0	Epoch 38, train[loss: 4.071007, acc: 11.703699, top_10_acc: 46.961212], Test[loss: 4.476312, acc: 9.549632, top_10_acc: 41.977280]
2020-08-17 02:28:14.432 HKT	master-replica-0	Epoch 39, train[loss: 4.063299, acc: 11.757642, top_10_acc: 47.091572], Test[loss: 4.471290, acc: 9.424054, top_10_acc: 42.314060]
2020-08-17 02:28:14.432 HKT	master-replica-0	Epoch 40, train[loss: 4.052914, acc: 11.855253, top_10_acc: 47.350372], Test[loss: 4.471430, acc: 9.715167, top_10_acc: 42.097153]
2020-08-17 02:28:15.039 HKT	master-replica-0	Module completed; cleaning up.
2020-08-17 02:28:15.040 HKT	master-replica-0	Clean up finished.
2020-08-17 02:28:15.040 HKT	master-replica-0	Task completed successfully.
2020-08-17 02:30:26.584 HKT	service	Job completed successfully.

No newer entries found matching current filter. Load newer logs

# Hybrid Recommendation System

- The model has 42.09% chance to correctly predict the next news article the reader would like to view if our model provide 10 recommended items.
- If randomly picking 10 items from total 2421 news articles, the top 10 accuracy would be only 0.413%. Our hybrid model has around 100 times better top 10 accuracy than random picking.



# Outline

---

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

# Reference

---

- [Neural Collaborative Filtering](#)
- [Wide & Deep Learning for Recommender Systems](#)
- [Recommendation Systems with TensorFlow on GCP](#)
- [End-to-end Machine Learning with TensorFlow on GCP](#)
- [Collaborative Filtering using Deep Neural Networks \(in Tensorflow\)](#)
- [Get started with TensorBoard](#)
- [Deploying models](#)
- [Method: projects.predict](#)
- [Simple Matrix Factorization example on the Movielens dataset using Pyspark](#)

Thank you for your attention!!