Research Progress on Data Intelligence in AI Graduate Program (AI 대학원 데이터 지능 연구 성과)

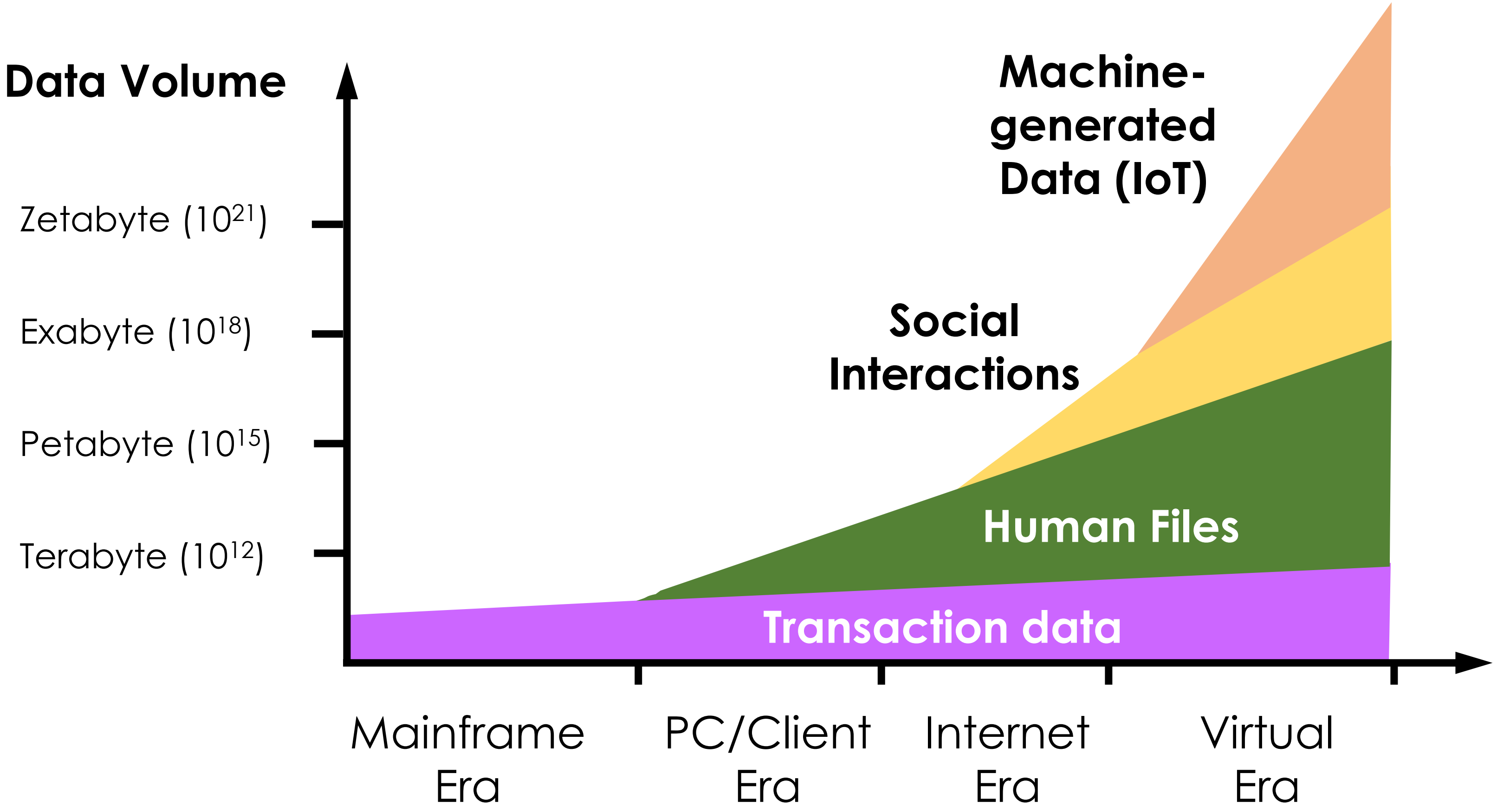
Jongwuk Lee (이종욱)

Sungkyunkwan University (성균관대)

What is Data Intelligence?

Explosive Growth of Data

- **Data exist everywhere!**



Source: *https://medium.com/@melodyucros/ladyboss-heres-why-you-should-study-big-data-721b04b8a0ca*

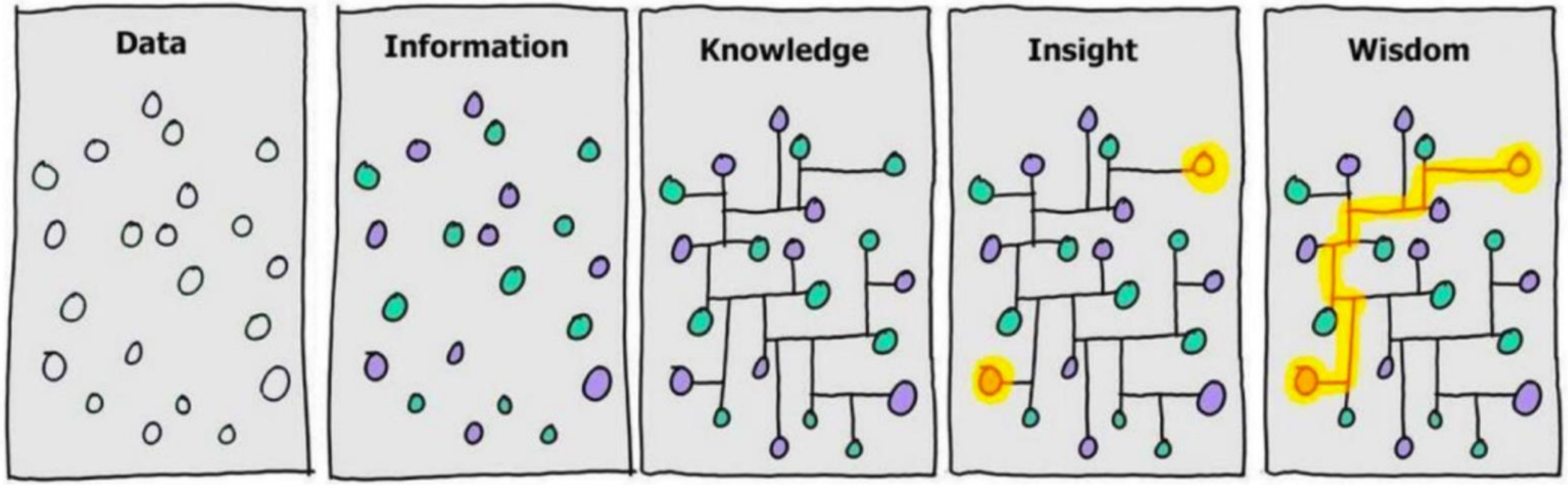
Data Flooding and Overload

- Drowning in data but starving for knowledge!

What is Data Mining (DM)?

- **Data mining (knowledge discovery from data)**
  - Extraction of **interesting** (**non-trivial**, **implicit**, **unknown** and **potentially useful**) **patterns** or **knowledge** from data

- **Refers to data intelligence for business/real-world objective.**



Source: https://www.theifactory.com/news/gaining-wisdom-from-data/

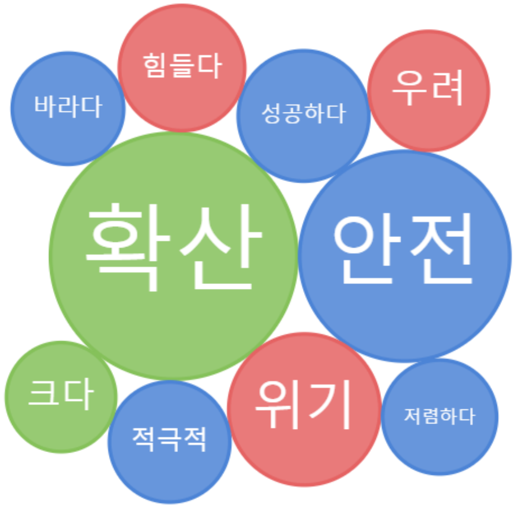
Example: Word Association Map

- **Finding relevant words for “코로나19” from a news corpus**

2020년 11월 4주차

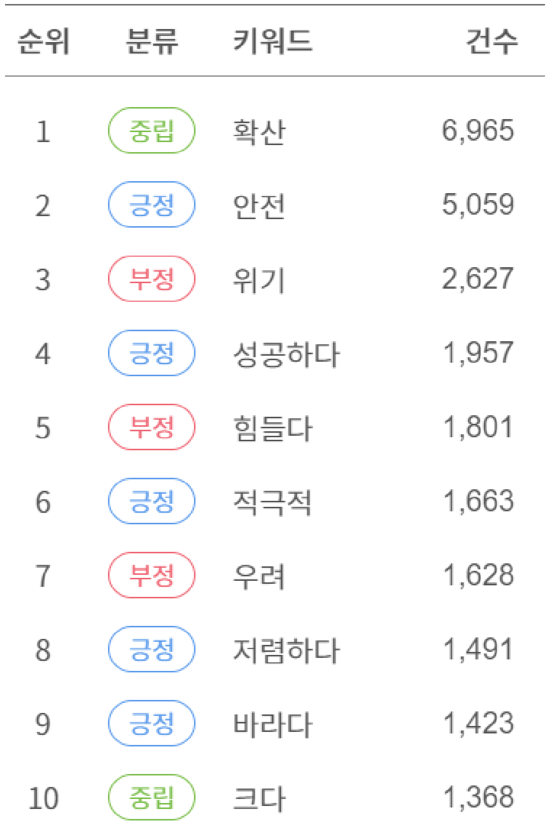
(2020.11.23 ~ 2020.11.25)

"코로나19" 감성 연관어 TOP 10



긍정 45%　부정 23%　중립 32%

감성어 랭킹



순위	분류	키워드	건수
1	중립	확산	6,965
2	긍정	안전	5,059
3	부정	위기	2,627
4	긍정	성공하다	1,957
5	부정	힘들다	1,801
6	긍정	적극적	1,663
7	부정	우려	1,628
8	긍정	저렴하다	1,491
9	긍정	바라다	1,423
10	중립	크다	1,368

http://www.socialmetrics.co.kr/

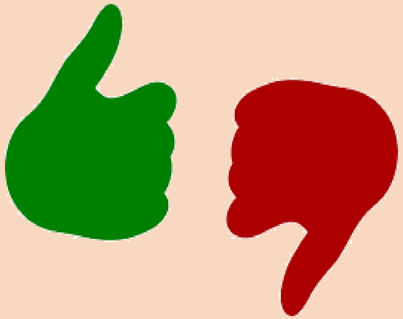
Our Achievements in AI Graduate Program

- **20+ top-tier papers** were published in WWW, KDD, SIGIR, ICDM, CIKM, ACL, EMNLP, COLING, and VLDB.

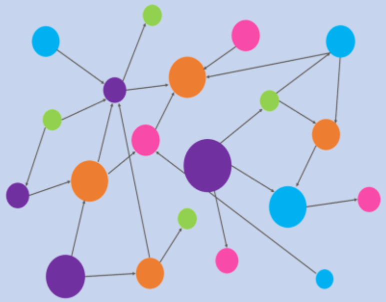
Our Achievements in AI Graduate Program

- **Three main research topics are**

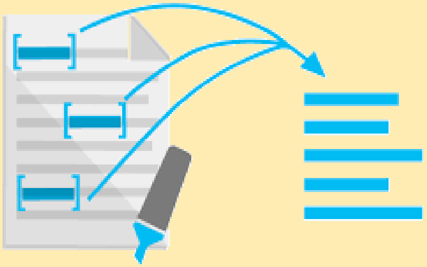
Rec systems



Graph mining



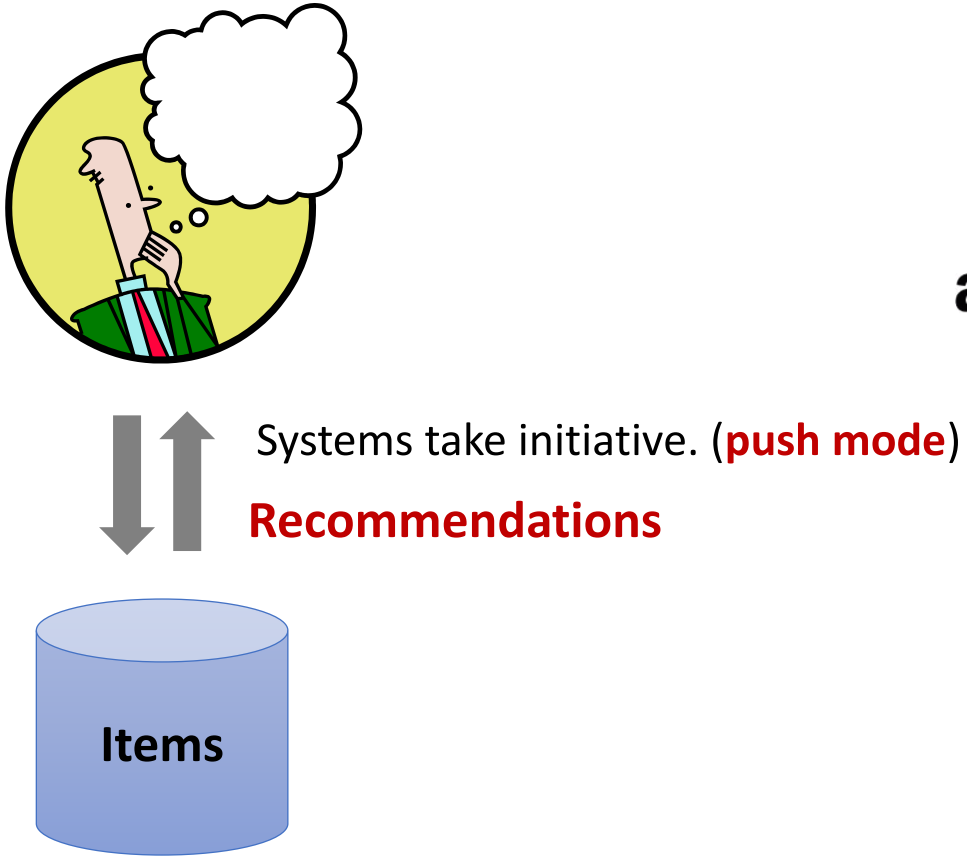
Text mining and Understanding



Recommender Systems

Case 1: Recommender Systems

- **How to provide relevant items to users?**



Systems take initiative. (**push mode**)

Recommendations

products, movies, music, news, …

LinkedIn

PANDORA

amazon.com

del.icio.us

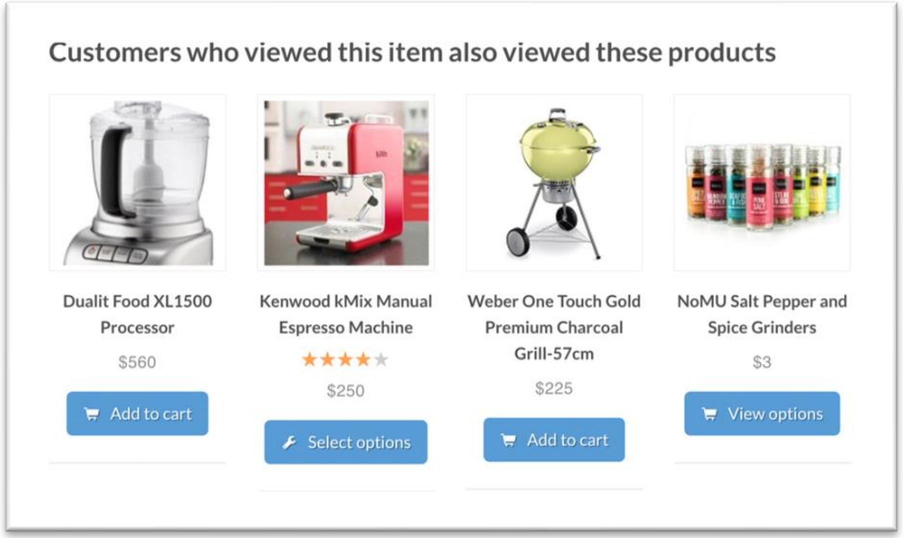
NETFLIX

Spotify

Google News

YouTube

Case 1: Recommender Systems

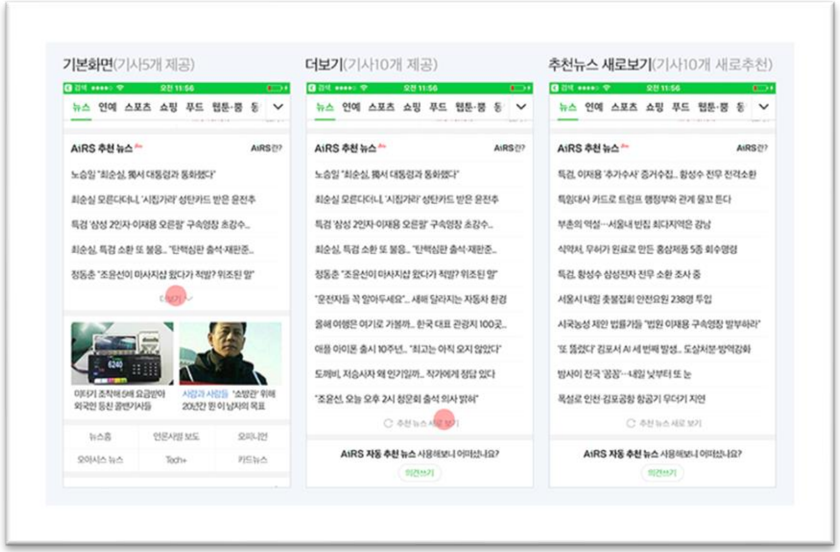

Customers who viewed this item also viewed these products

Dualit Food XL1500 Processor
$560

Kenwood kMix Manual Espresso Machine
★★★★☆
$250

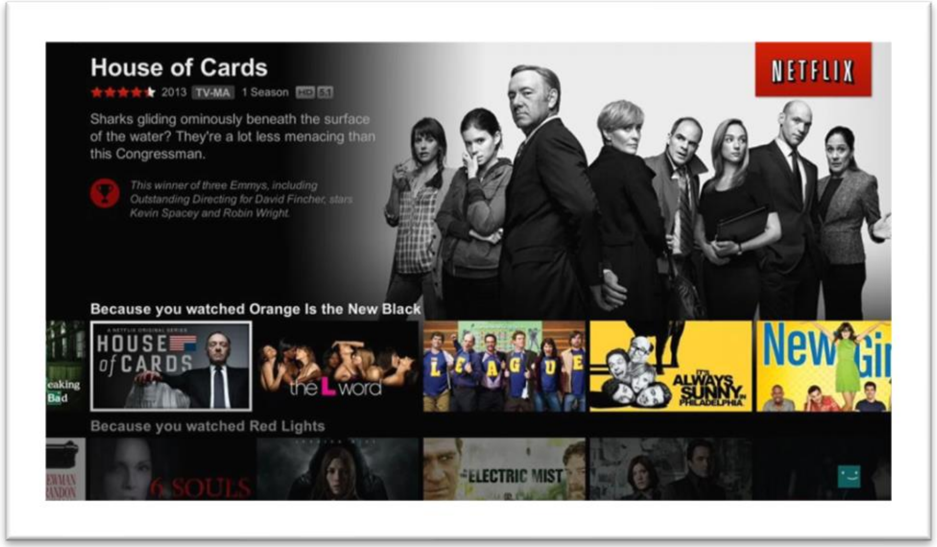
Weber One Touch Gold Premium Charcoal Grill-57cm
$225

NoMU Salt Pepper and Spice Grinders
$3


기본화면(기사5개 제공)

더보기(기사10개 제공)

추천뉴스 새로보기(기사10개 새로추천)


House of Cards

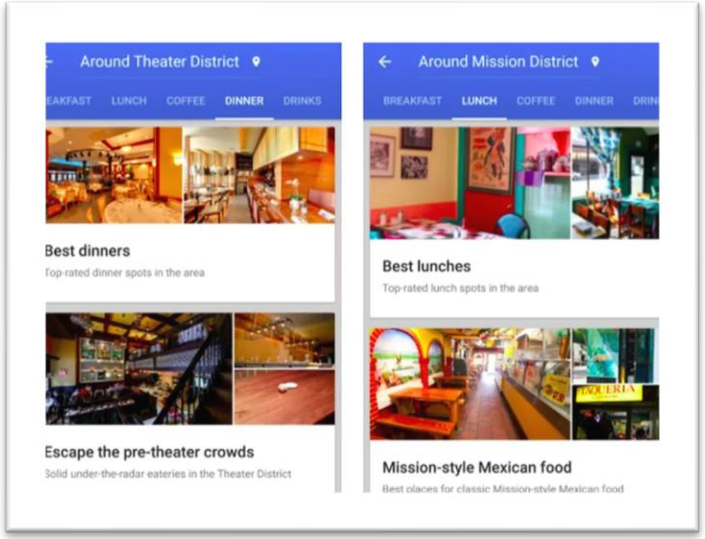
★★★★★ 2013 TV-MA 1 Season HD 5.1

Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.

This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.

Because you watched Orange Is the New Black

Because you watched Red Lights


Around Theater District

Best dinners
Top-rated dinner spots in the area

Escape the pre-theater crowds
Solid under-the-radar eateries in the Theater District

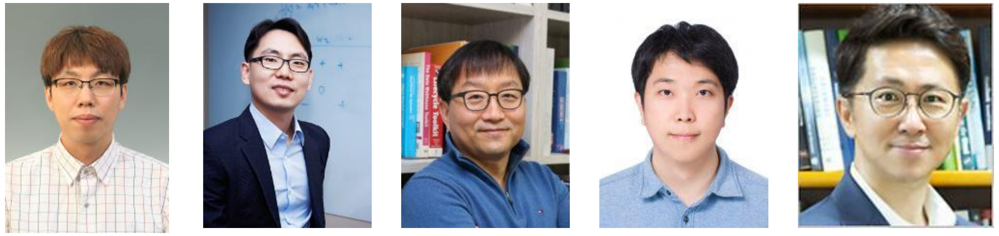
Around Mission District

Best lunches
Top-rated lunch spots in the area

Mission-style Mexican food
Best places for classic Mission-style Mexican food

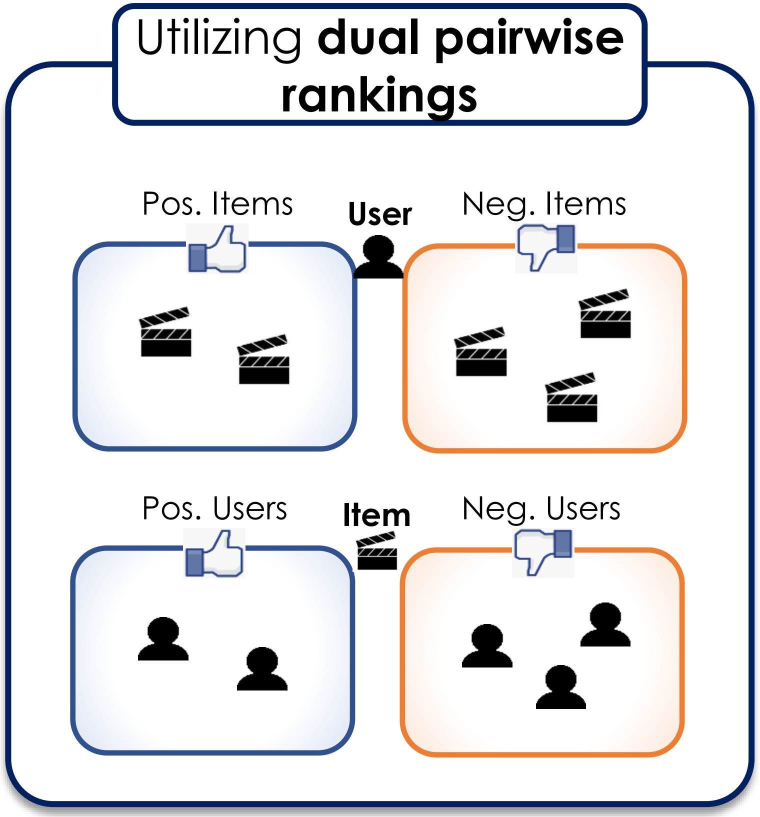
Our Achievements in RecSys

- **6 papers in WWW, SIGIR, ICDM, and CIKM**
  - Dual Neural Personalized Ranking, **WWW** 2019 (**SKKU**)
  - Collaborative Distillation for Top-N Recommendation, **ICDM** 2019 (**SKKU**)
  - AR-CF: Augmenting Virtual Users and Items in Collaborative Filtering for Addressing Cold-Start Problem, **SIGIR** 2020 (**Hanyang Univ.**)
  - Interest Sustainability-Aware Recommender System, **ICDM** 2020 (**POSTECH**)
  - DE-RRD: A Knowledge Distillation Framework for Recommender System, **CIKM** 2020 (**POSTECH**)
  - News Recommendation with Topic-Enriched Knowledge Graphs, **CIKM** 2020 (**Yonsei Univ.**)

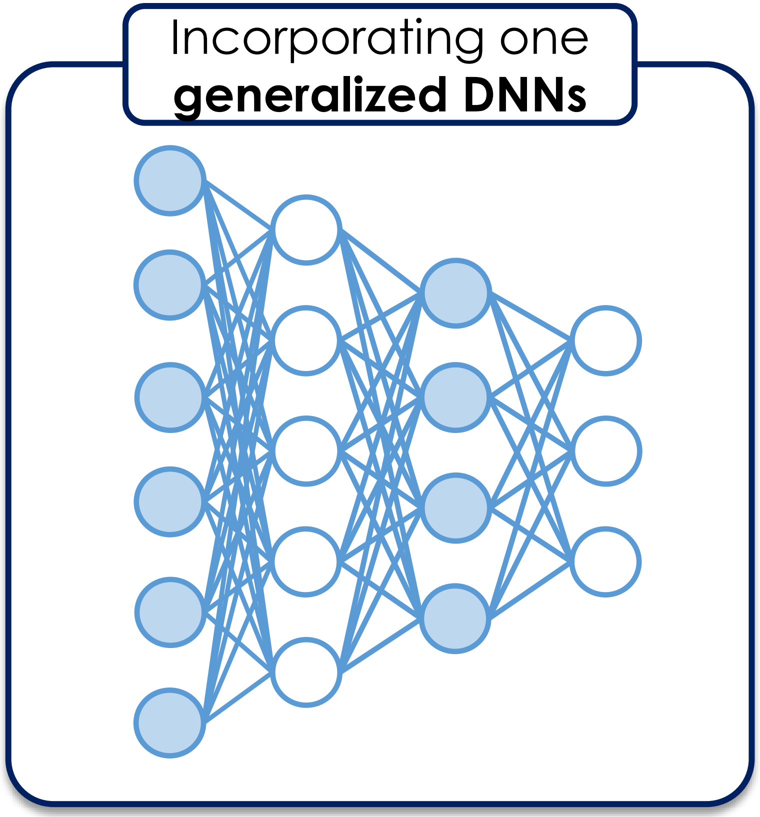


Dual Pairwise Ranking

- Utilizing both user- and item-side pairwise rankings over a neural architecture



Utilizing **dual pairwise rankings**

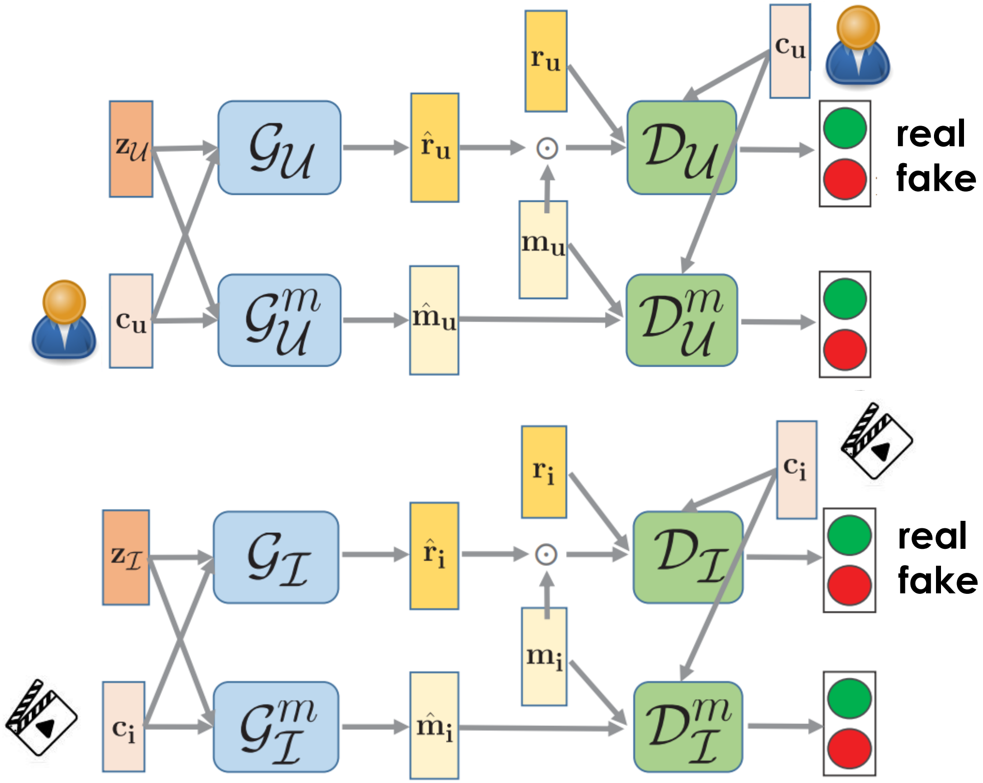


Incorporating one **generalized DNNs**

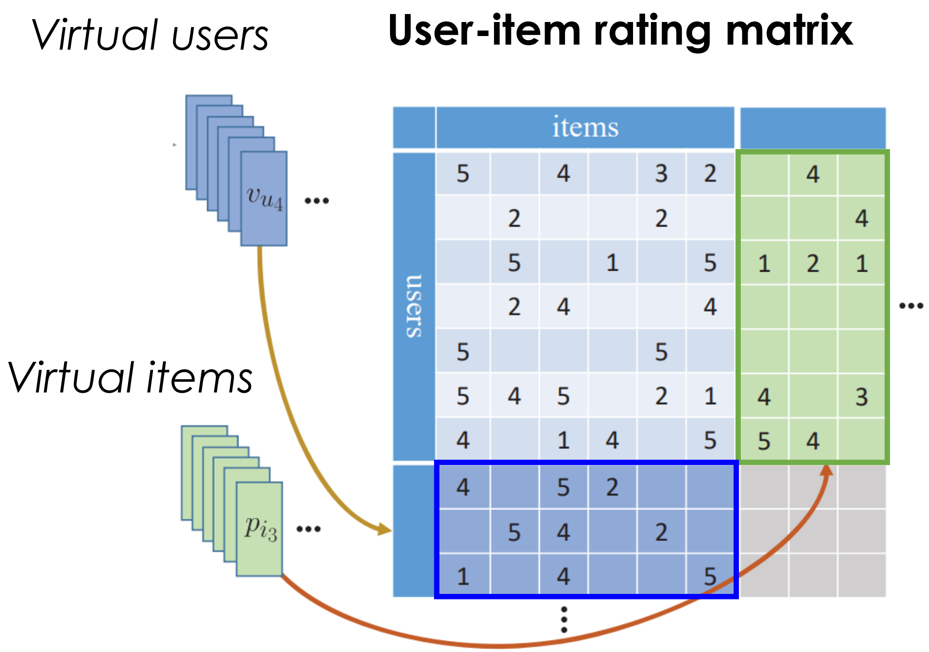
Dual Neural Personalized Ranking, WWW 2019 (SKKU)

AR-CF with Virtual Users/Items

- **Adopting GANs to address the cold-start problem**


r_u
c_u
$z_\mathcal{U}$
$\mathcal{G}_\mathcal{U}$
$\hat{r}_u$
$\odot$
$\mathcal{D}_\mathcal{U}$
real
fake
m_u
c_u
$\mathcal{G}^m_\mathcal{U}$
$\hat{m}_u$
$\mathcal{D}^m_\mathcal{U}$
r_i
c_i
$z_\mathcal{I}$
$\mathcal{G}_\mathcal{I}$
$\hat{r}_i$
$\odot$
$\mathcal{D}_\mathcal{I}$
real
fake
m_i
c_i
$\mathcal{G}^m_\mathcal{I}$
$\hat{m}_i$
$\mathcal{D}^m_\mathcal{I}$

Training CGANs to generate virtual users/items



User-item rating matrix

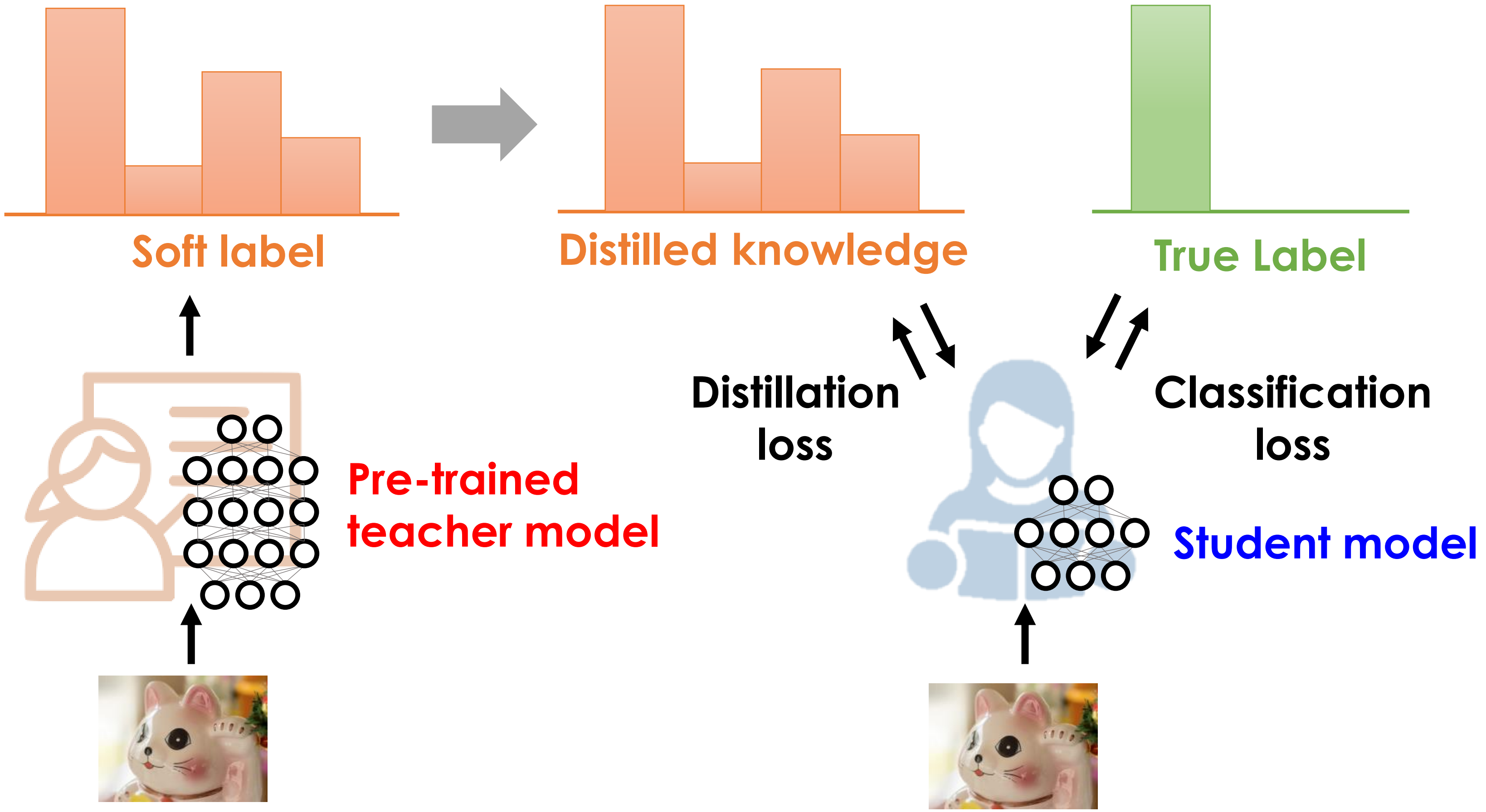
	items								
users	5		4		3	2		4	
		2			2				4
		5		1		5	1	2	1
		2	4			4			
	5				5				
	5	4	5		2	1	4		3
	4		1	4		5	5	4	
	4		5	2					
		5	4		2				
	1		4			5			

Augmenting the original user-item matrix

AR-CF: Augmenting Virtual Users and Items in Collaborative Filtering for Addressing Cold-Start Problem, SIGIR 2020 (Hanyang Univ.)

Knowledge Distillation (KD)

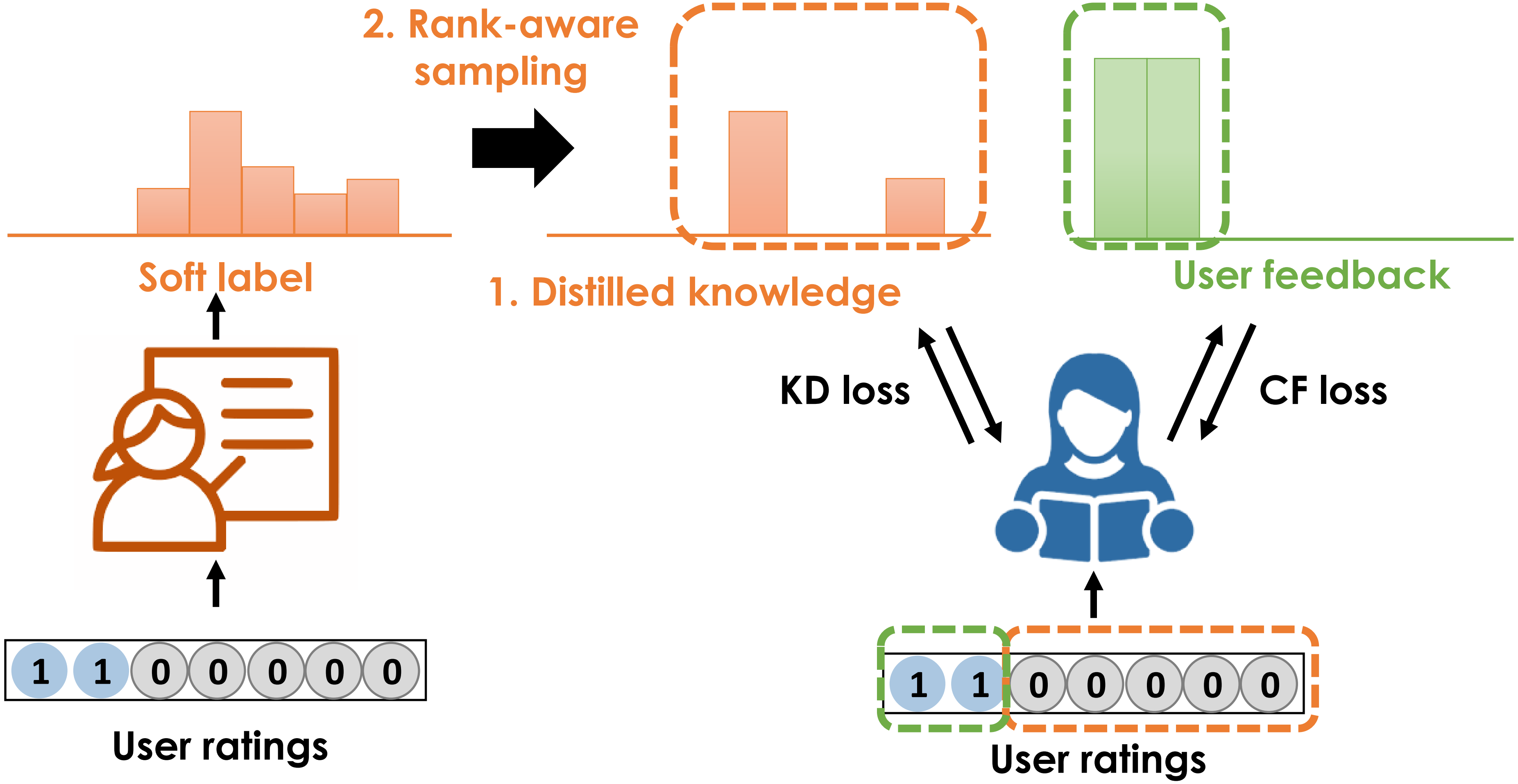
- A small **student model** is trained to mimic a pre-trained and large **teacher model**.



Distilling the Knowledge in a Neural Network, NIPS Workshop 2014

Collaborative Distillation (CD)

- **Applying KD for recommender models**



2. Rank-aware sampling

Soft label

1. Distilled knowledge

User feedback

KD loss

CF loss

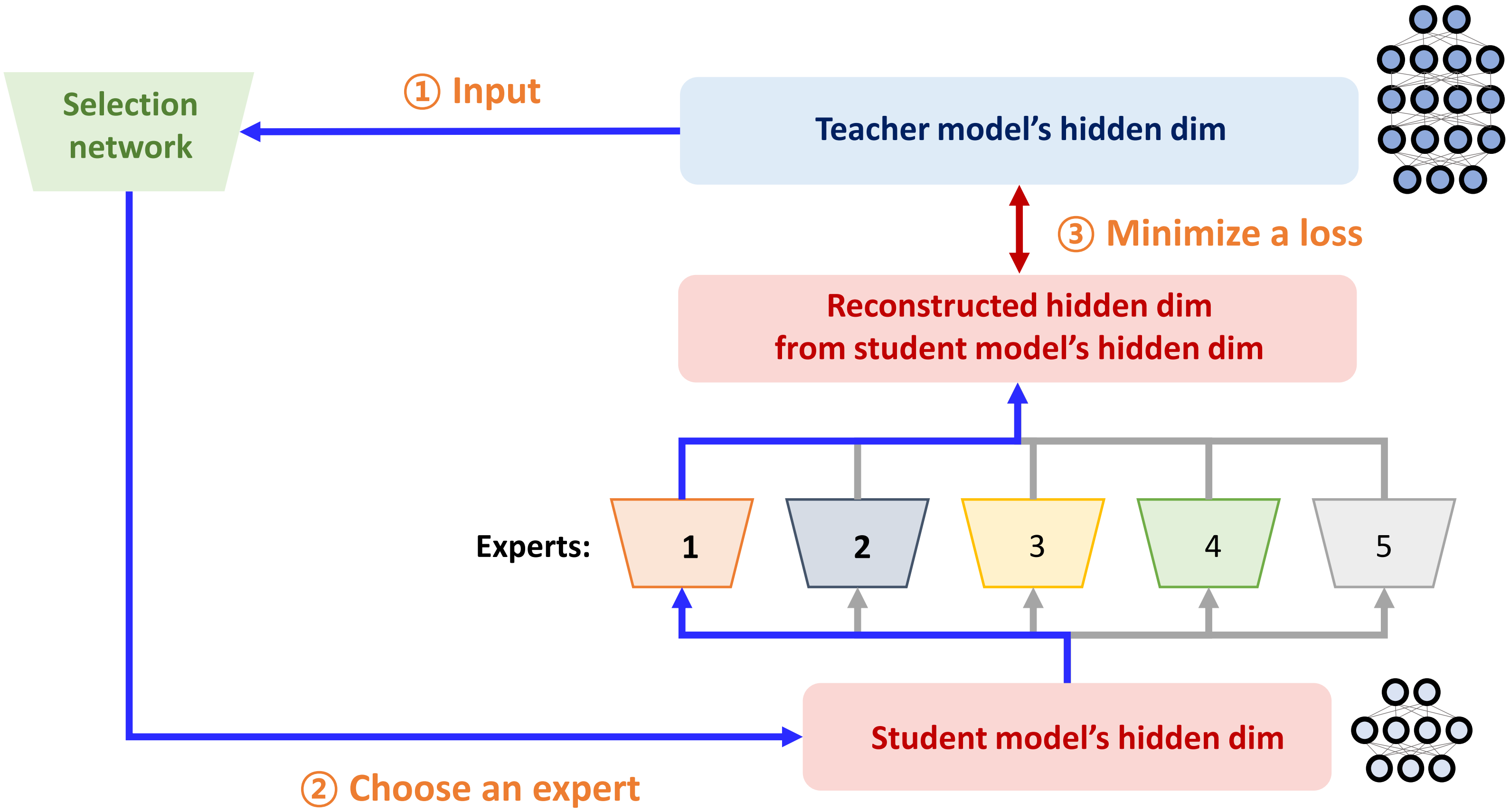
User ratings

User ratings

Collaborative Distillation for Top-N Recommendation, ICDM 2019 (SKKU)

Distillation Experts for Compression

- Propose distillation experts (DE) to transfer latent knowledge from the teacher model.

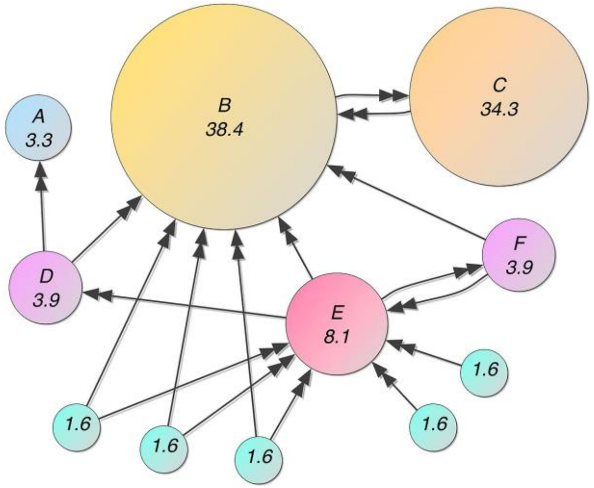

Selection network
① Input
Teacher model's hidden dim
③ Minimize a loss
Reconstructed hidden dim from student model's hidden dim
Experts:
1
2
3
4
5
Student model's hidden dim
② Choose an expert

DE-RRD: A Knowledge Distillation Framework for Recommender System, CIKM 2020 (POSTECH)

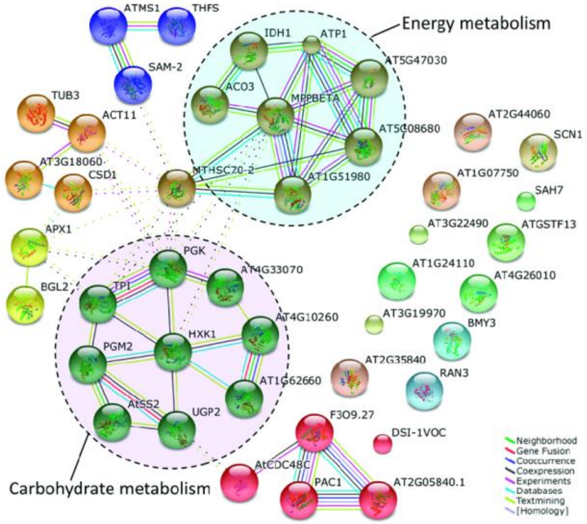
Graph Mining

Case 2: Graph Analysis and Learning

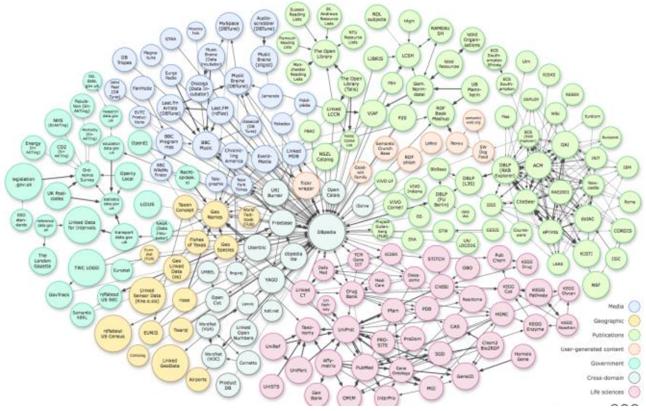
- Finding frequent subgraphs and substructures
- Learning graph embedding
- Predicting/classifying node and edges



5B+ Web pages



6k+ proteins



Knowledge graph

Case 2: Graph Analysis and Learning

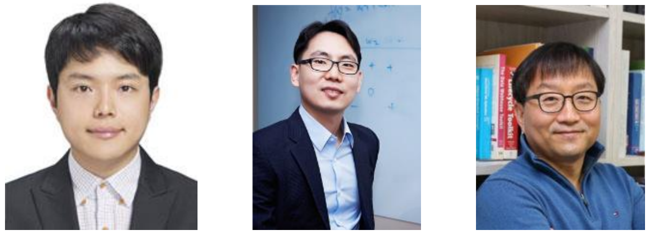
- **Nodes as people and edges as friendship**

facebook
December 2010

2.5B+ users in Facebook

Our Achievements in Graph Mining

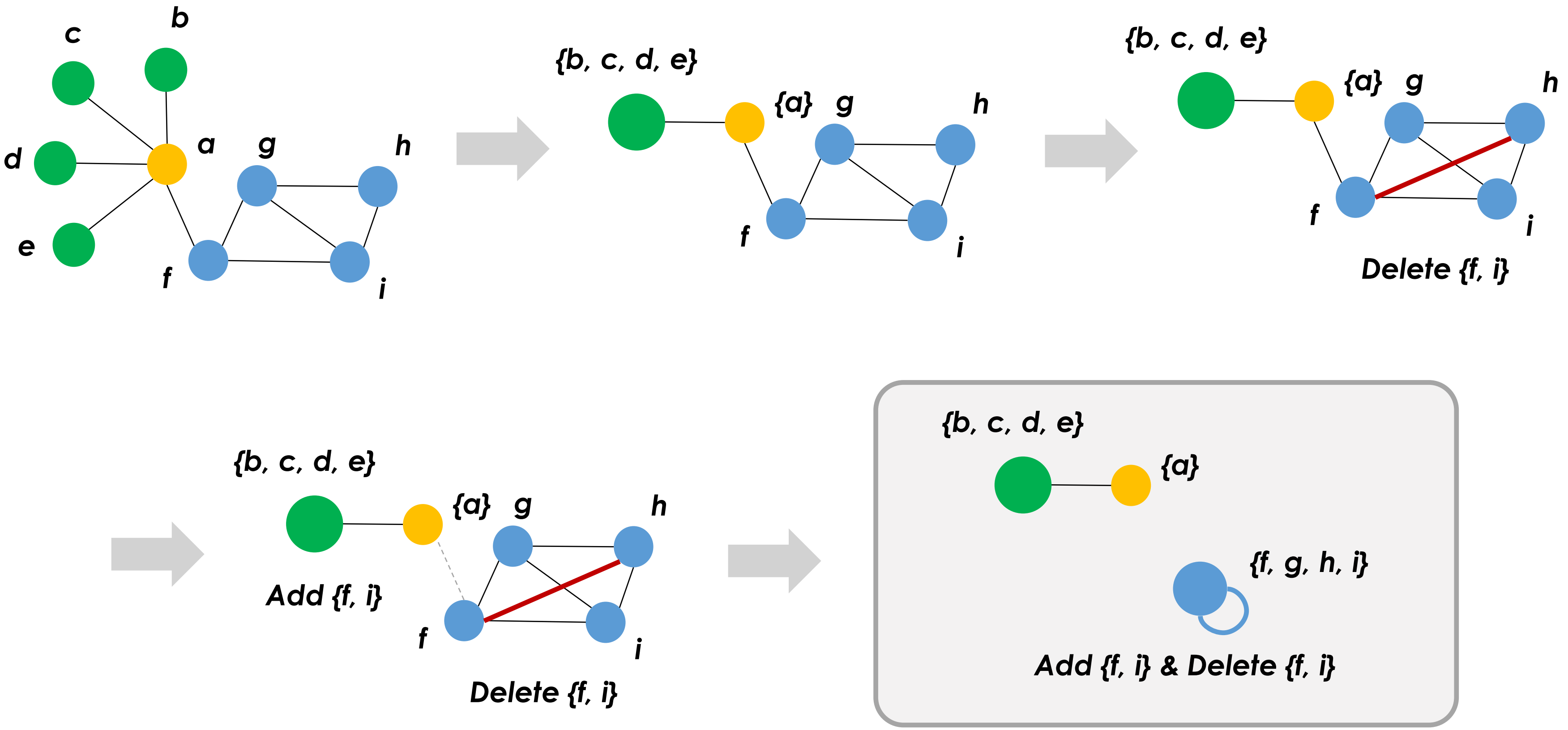
- **8 papers in WWW, KDD, SIGIR, ICDM, and VLDB**
  - How Much and When Do We Need Higher-order Information in Hypergraphs? A Case Study on Hyperedge Prediction (**WWW** 2020, **KAIST**)
  - Structural Patterns and Generative Models of Real-world Hypergraphs (**KDD** 2020, **KAIST**)
  - SSumM: Sparse Summarization of Massive Graphs (**KDD** 2020, **KAIST**)
  - Incremental Lossless Graph Summarization (**KDD** 2020, **KAIST**)
  - Hypergraph Motifs: Concepts, Algorithms, and Discoveries (**VLDB** 2020, **KAIST**)
  - Evolution of Real-world Hypergraphs: Patterns and Models without Oracles (**ICDM** 2020, **KAIST**)
  - Unsupervised Differentiable Multi-aspect Network Embedding (**KDD** 2020, **POSTECH**)
  - ASiNE: Adversarial Signed Network Embedding (**SIGIR** 2020, **Hanyang Univ.**)



Lossless Graph Summarization

- **Generating a lossless graph summarization**

The initial input has 9 nodes with 10 edges.



The final output has 3 nodes with 4 edges.

Lossless Graph Summarization

- **How to effectively generate a lossless graph summarization for a dynamic environment?**
  - When the graphs are dynamically changed, minimizes the updates without reconstruction.

- **How to efficiently summarize the graph with a given bit constraint?**
  - **Given**: a graph $\boldsymbol{G}$
  - **Find**: a summary graph $\overline{\boldsymbol{G}}$
  - **Objective**: Minimize the difference between $\boldsymbol{G}$ and $\overline{\boldsymbol{G}}$.
  - **Subject to**: the size of $\overline{\boldsymbol{G}}$ in bits $\leq \boldsymbol{k}$.

Incremental Lossless Graph Summarization, KDD 2020 (KAIST)

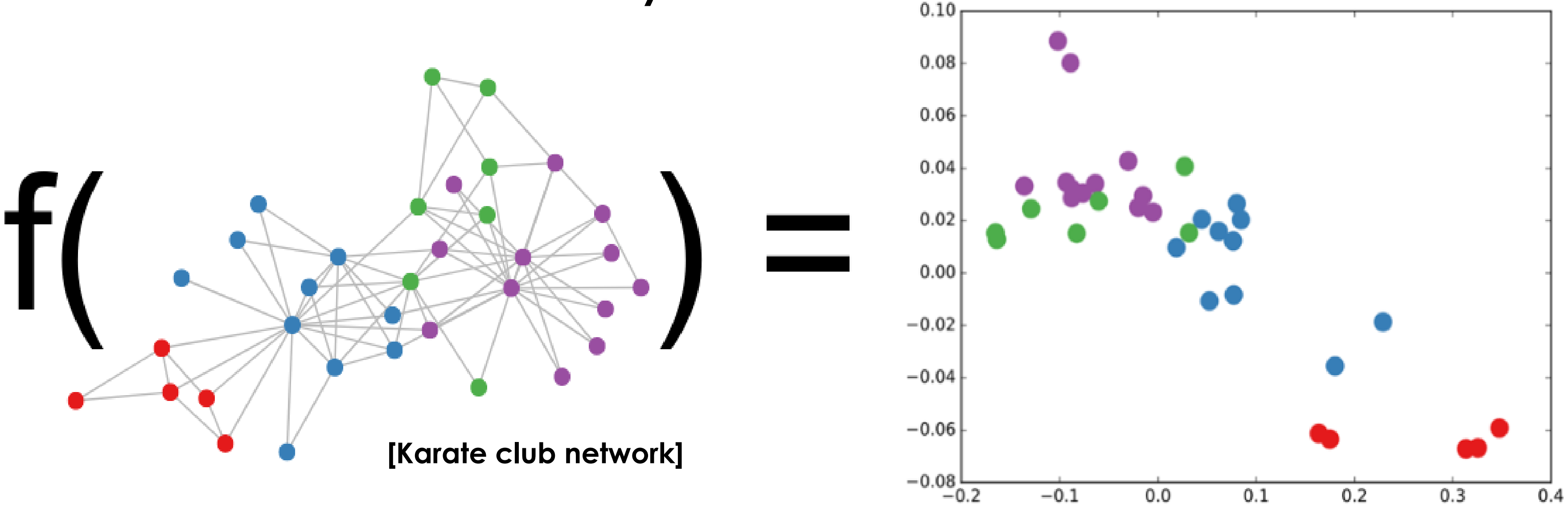
SSumM: Sparse Summarization of Massive Graphs, KDD 2020 (KAIST)

Graph Embedding

- **The similar nodes in a network have similar vector representation.**
  - The similarity in the embedding space **approximates** the similarity in the original network.

Parameters initialized randomly

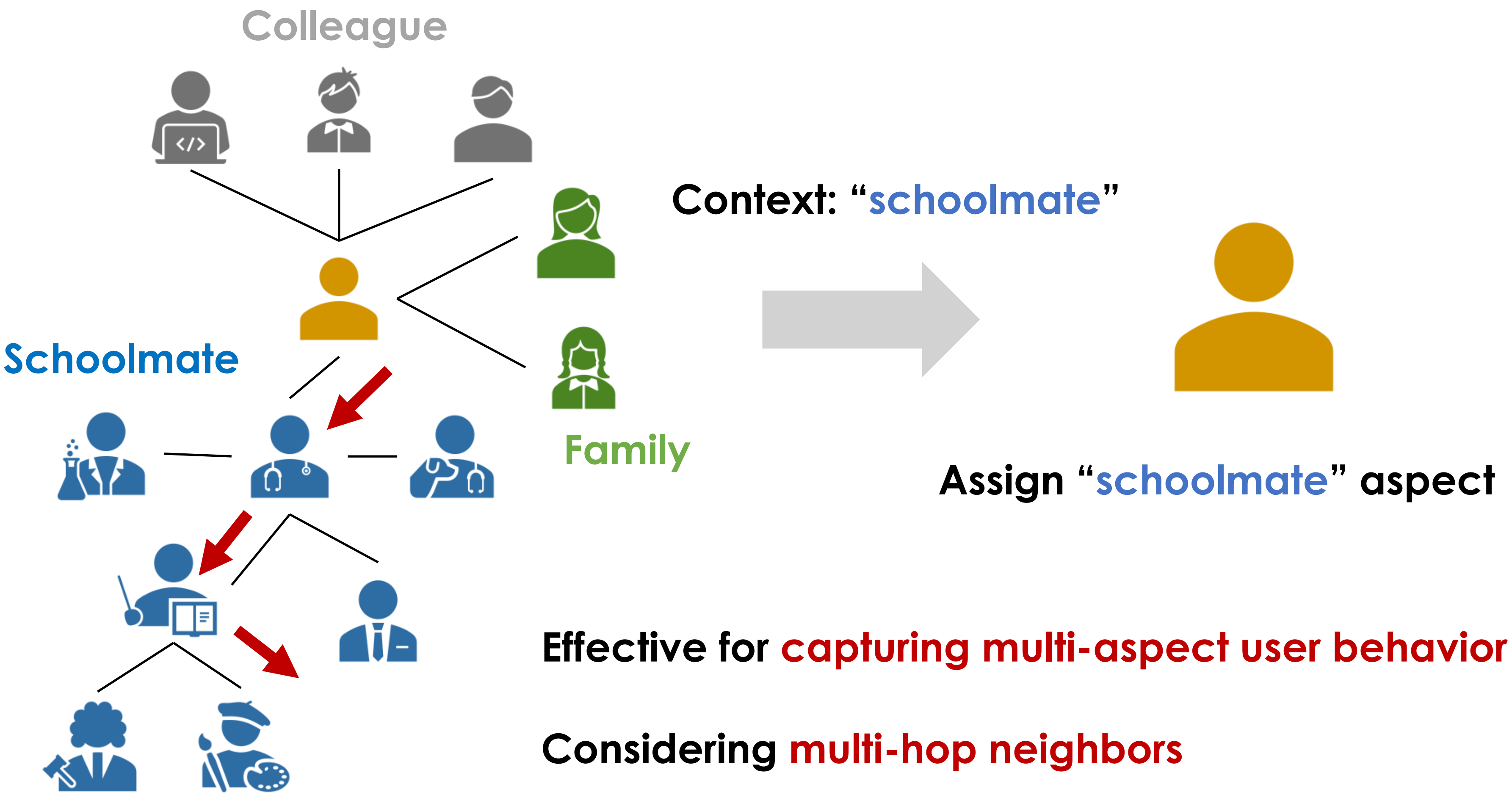
2-dim output per node



[Karate club network]

Source: https://iamsiva11.github.io/graph-embeddings-2017-part1/

Multi-aspect Graph Embedding



Context: "schoolmate"

Assign "schoolmate" aspect

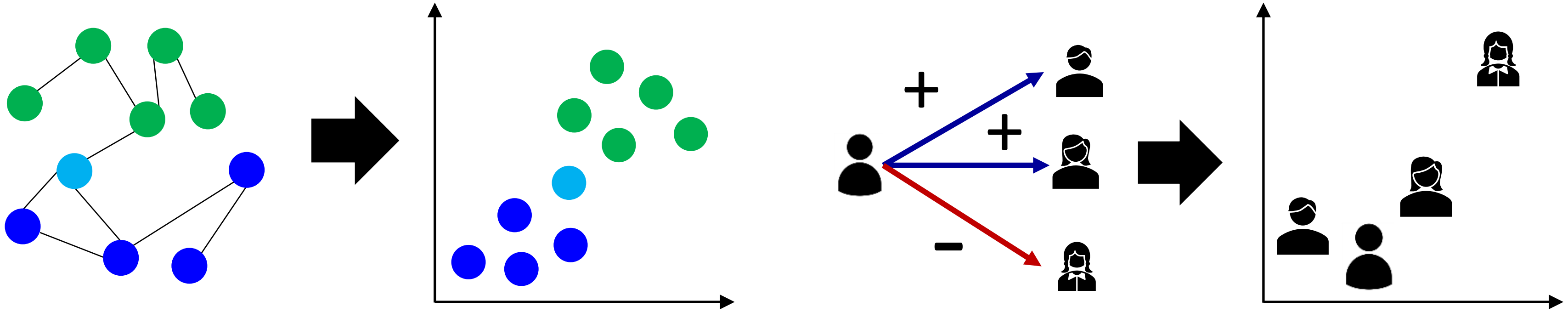
Effective for capturing multi-aspect user behavior

Considering multi-hop neighbors

Unsupervised Differentiable Multi-aspect Network Embedding, KDD 2020 (POSTECH)

Signed Graph Embedding

- **How to represent nodes in the embedding space?**
  - Nodes with the positive edges to be close
  - Nodes with the negative edges to be distant



- **A theory for signed graphs**

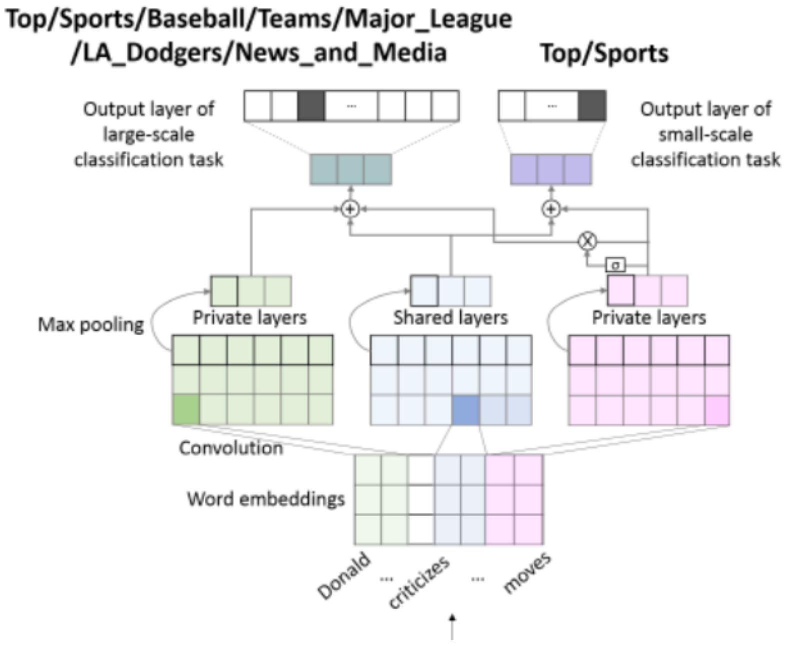
"A friend (+) of my friend (+) is my friend (+)"
"A friend (+) of my enemy (-) is my enemy (-)"
"An enemy (-) of my friend (+) is my enemy (-)"
"An enemy (-) of my enemy (-) is my friend (+)"

ASiNE: Adversarial Signed Network Embedding, SIGIR 2020 (Hanyang Univ.)

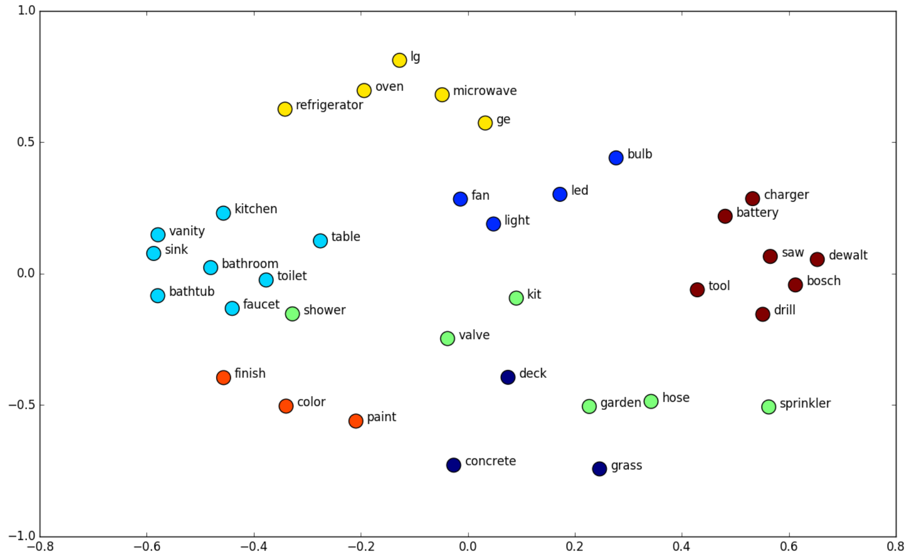
Text Mining and Understanding

Case 3: Text Mining and Understanding

- **Various NLP applications by understanding text**
  - Lage-scale text classification
  - Reading comprehension for question answering
  - Translating human language to machine language


Top/Sports/Baseball/Teams/Major_League
/LA_Dodgers/News_and_Media
Top/Sports
Output layer of large-scale classification task
Output layer of small-scale classification task
σ
Max pooling
Private layers
Shared layers
Private layers
Convolution
Word embeddings
Donald
...
criticizes
...
moves

Donald Trump criticizes Dodgers manager for bullpen moves.



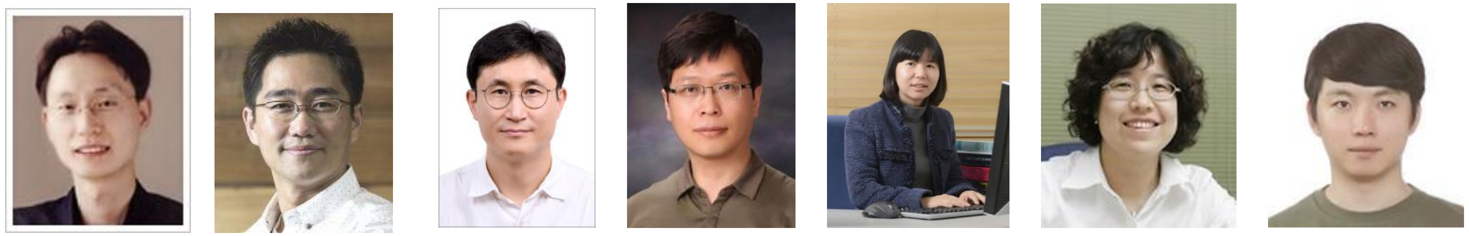
Case 3: Text Mining and Understanding

- **Natural language processing (NLP) lets computers to process and analyze large accounts of natural language data.**
  - Human-computer interaction
  - Includes the automation of linguistic forms, activities, and methods of communication.

AI + **Linguistics** = **NLP**

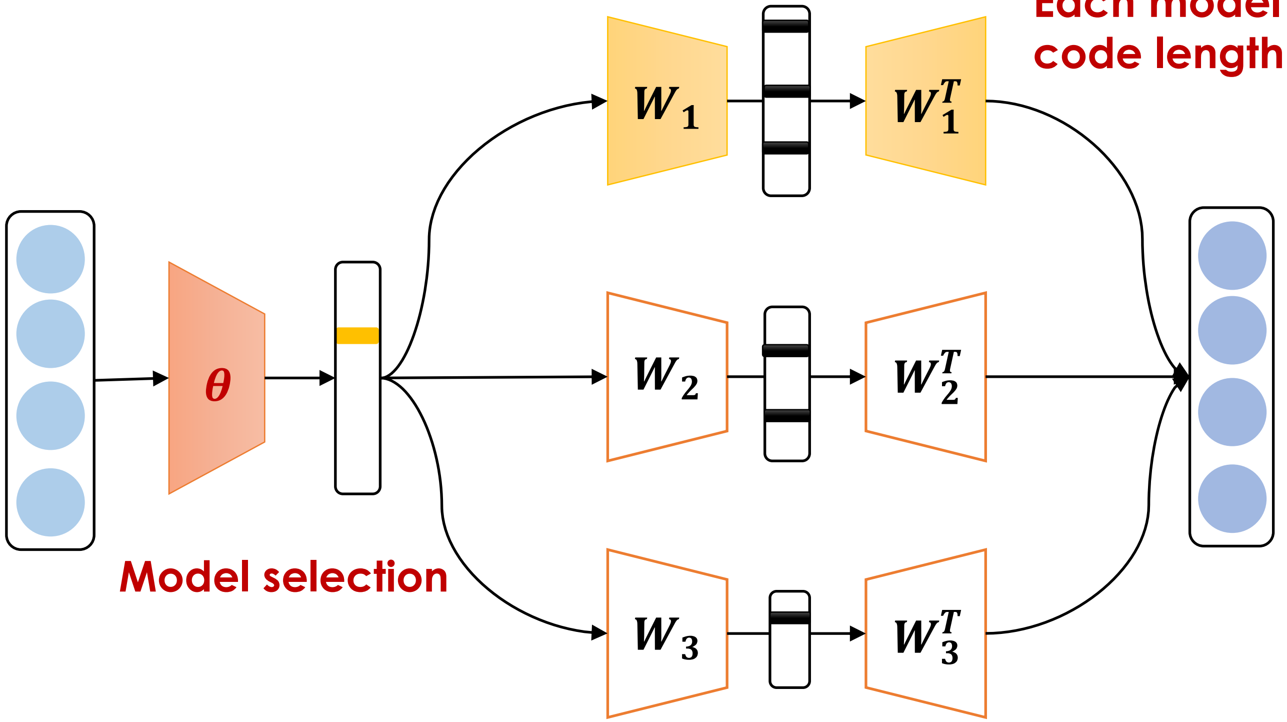
Our Achievements in Text Mining

- **9 papers in ACL, EMNLP, COLING, CIKM, and VLDB**
  - Adaptive Compression of Word Embeddings, **ACL** 2020 (**Korea Univ.**)
  - Multi-pretraining for Large-scale Text Classification, **EMNLP** 2020 (**Korea Univ.**)
  - ST-GRAT: A Novel Spatio-temporal Graph Attention Network for Accurately Forecasting Dynamically Changing Road Speed, CIKM 2020 (**KAIST**)
  - Multi-Task Learning for Knowledge Graph Completion with Pre-trained Language Models, **COLING** 2020 (**SKKU**)
  - Natural language to SQL: Where are we today? **VLDB** 2020, (**POSTECH**)
  - Extracting Chemical–Protein Interactions via Calibrated Deep Neural Network and Self-training, **EMNLP** 2020 (**GIST**)
  - Learning with Limited data for Multilingual Reading Comprehension, **EMNLP** 2019 (**Yonsei Univ.**)
  - Less is More: Attention Supervision with Counterfactuals for Text Classification, **EMNLP** 2020 (**Yonsei Univ.**)
  - Retriever-Augmented and Controllable Review Generation, **COLING** 2020 (**Yonsei Univ.**)



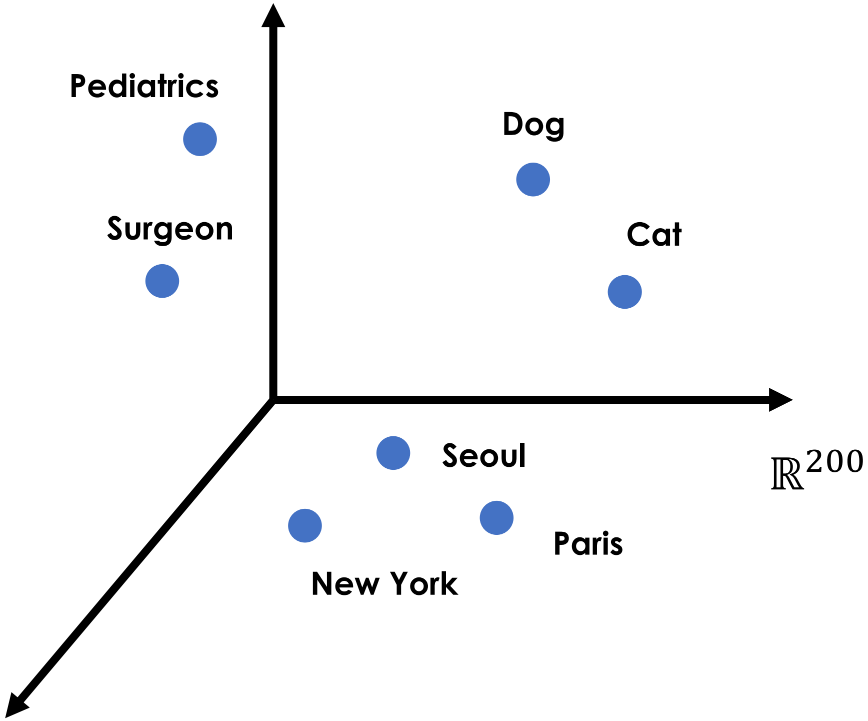
Word Embeddings Compression

- **Word embedding compression by adaptively assigning different lengths of discrete codes**
  - Using codes with a longer length for task-sensitive words



Each model has a different code length.

Model selection

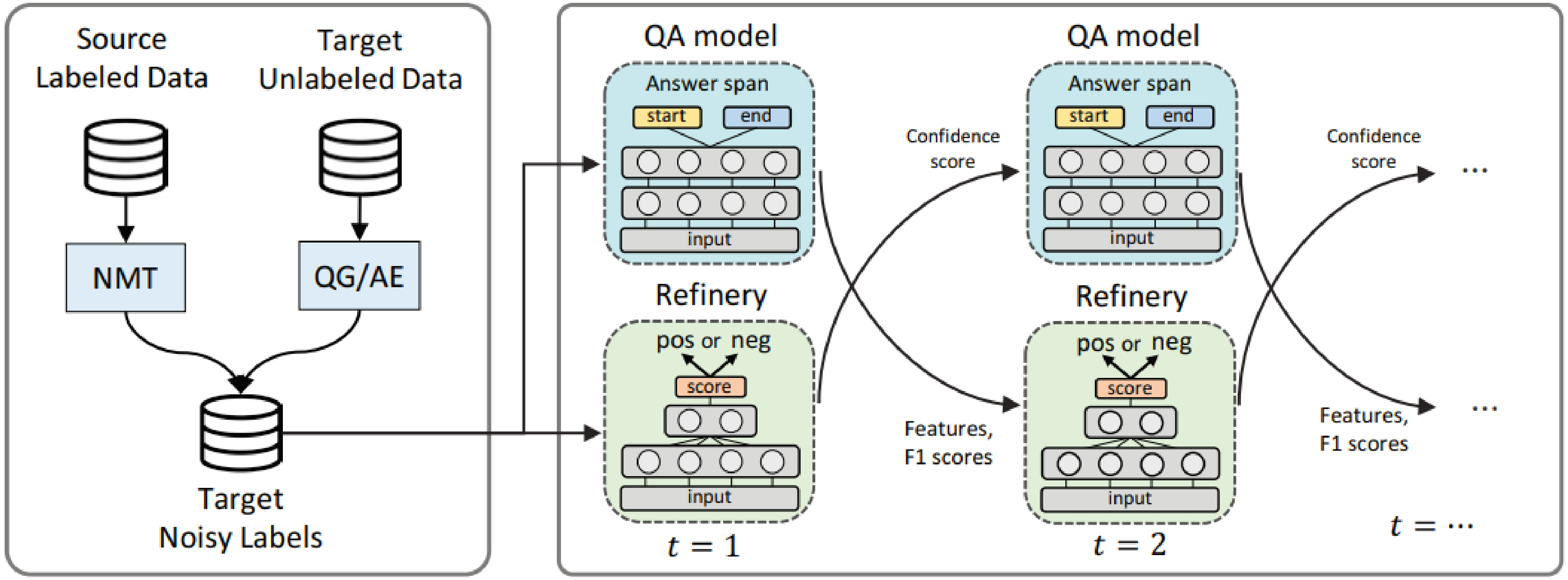


Example of word embeddings

Adaptive Compression of Word Embeddings, ACL 2020 (Korea Univ.)

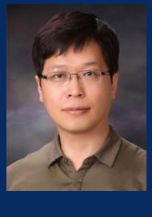
Multilingual Reading Comprehension

- **Supporting question answering in a new language with limited training resources**
  - **Transfer labels** from another language.
  - Generate labels from unlabeled data using a **translator** and an **automatic labeling function**.

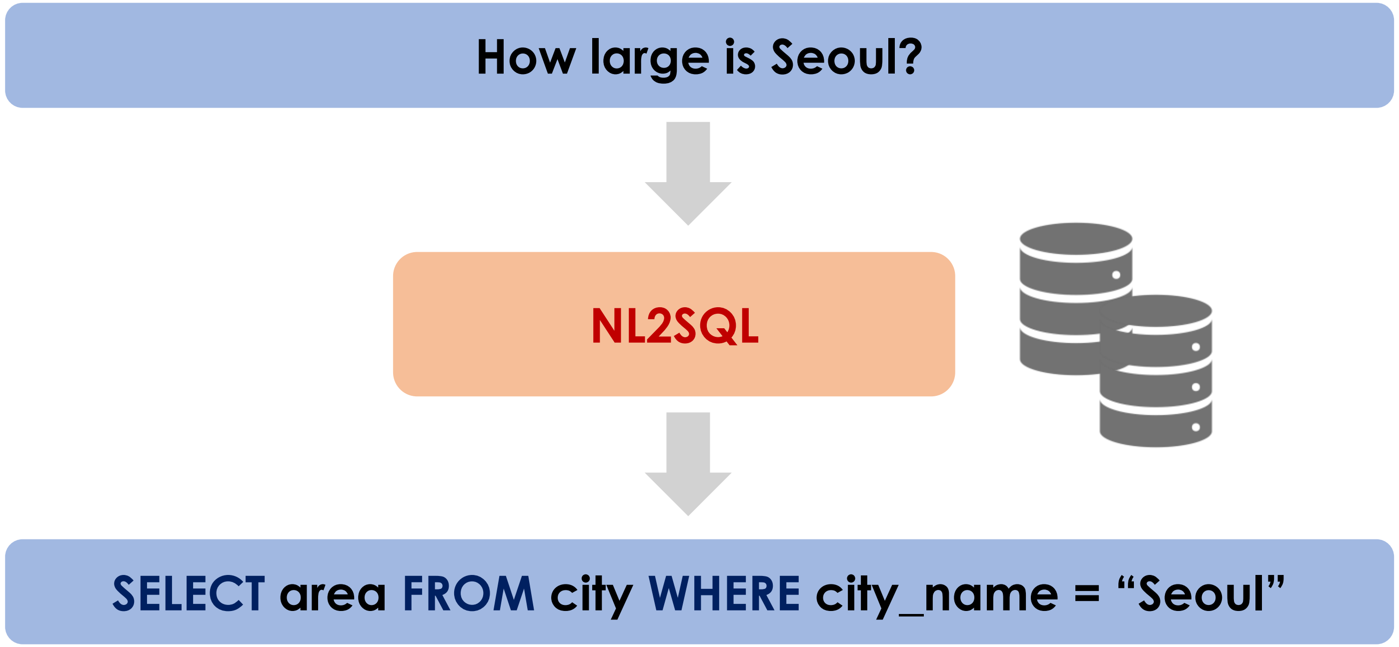


Learning with Limited data for Multilingual Reading Comprehension, EMNLP 2019 (Yonsei Univ.)

NLP-to-SQL Translation



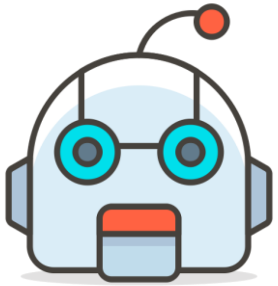
- Generating an SQL statement to answer a natural language question on a relational database



How large is Seoul?

NL2SQL

SELECT area FROM city WHERE city_name = “Seoul”

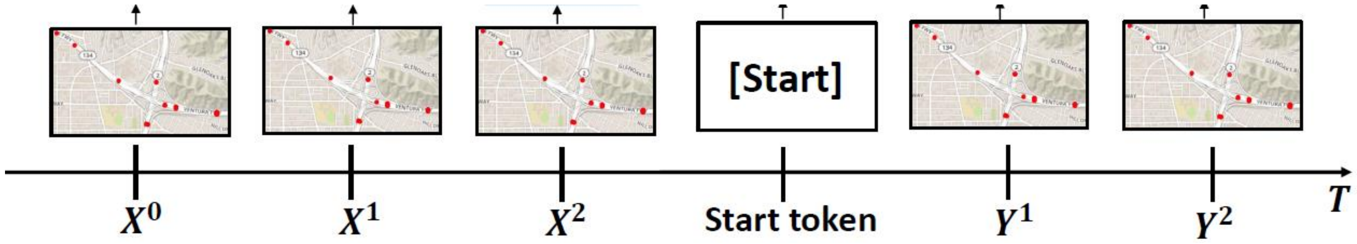


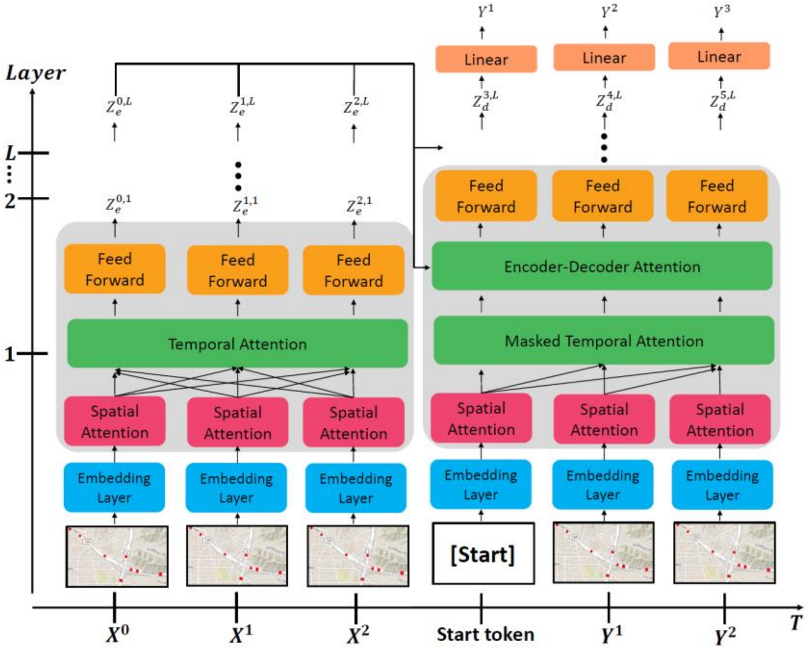
- Provide a comprehensive survey for NL2SQL.

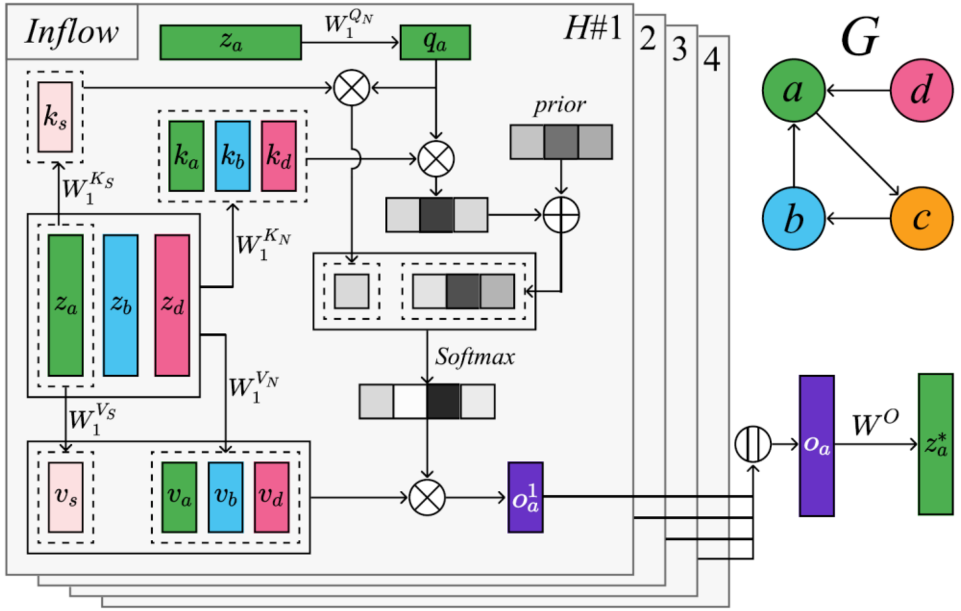
Natural language to SQL: Where are we today? VLDB 2020 (POSTECH)

Road Traffic Speed Prediction

- **Self-attentional model for forecasting spatio-temporal data**







ST-GRAT: A Novel Spatio-temporal Graph Attention Network for Accurately Forecasting Dynamically Changing Road Speed, CIKM 2020 (KAIST)

Q&A