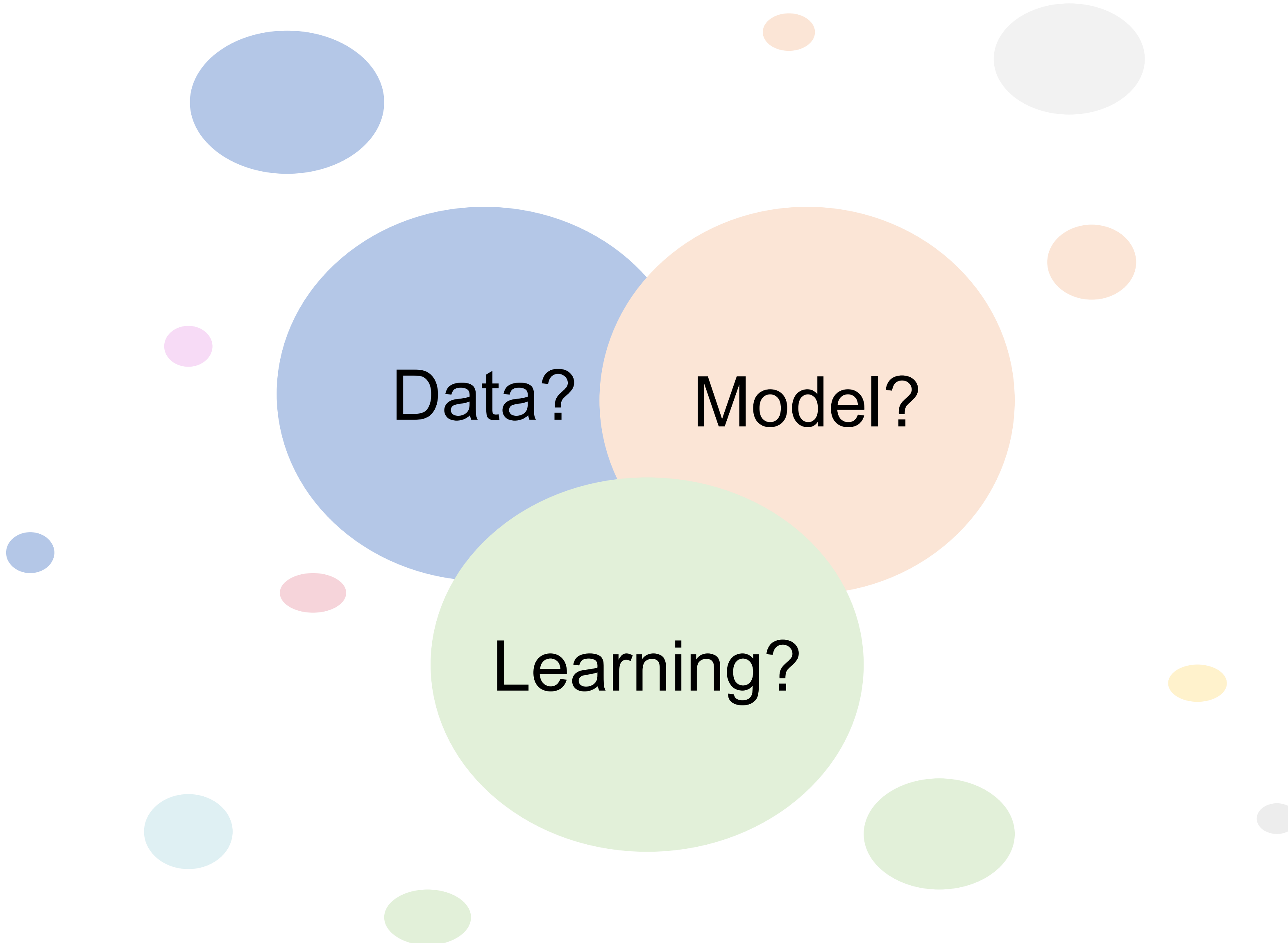
Contrastive Learning:
A Recommendation Perspective

Pengjie Ren (任鹏杰)
IRLab, Shandong University
renpengjie@sdu.edu.cn

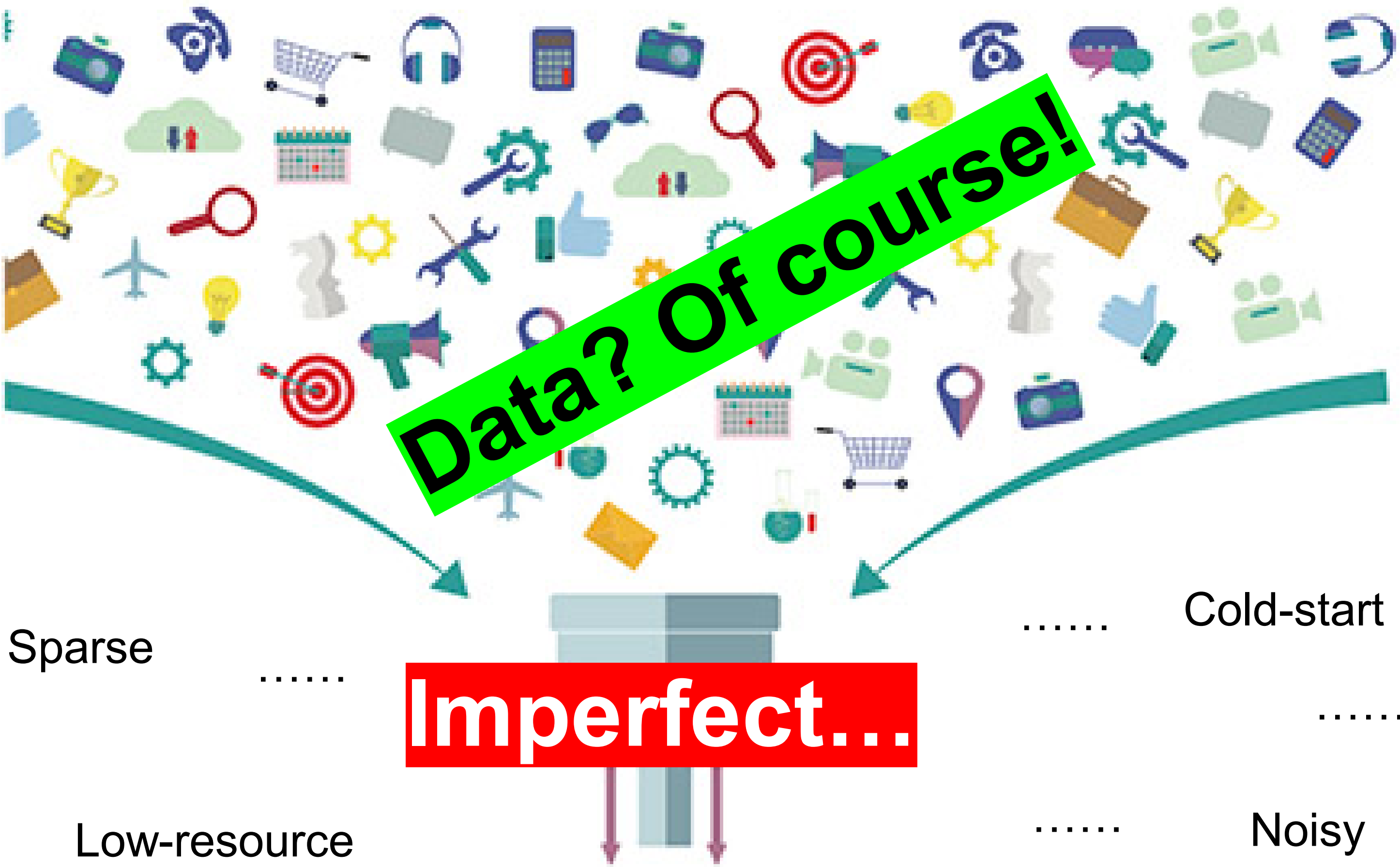
SHANDONG UNIVERSITY

山东大学信息检索实验室
Information Retrieval Lab

What matters for DL?



What matters for DL? Data?



Sparse …… …… Cold-start

……

Low-resource …… Noisy

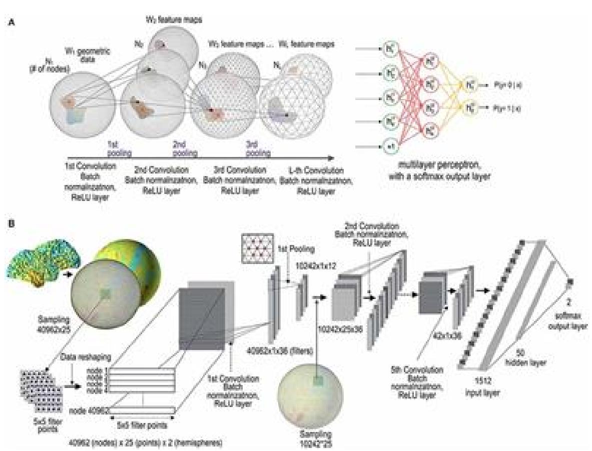
……

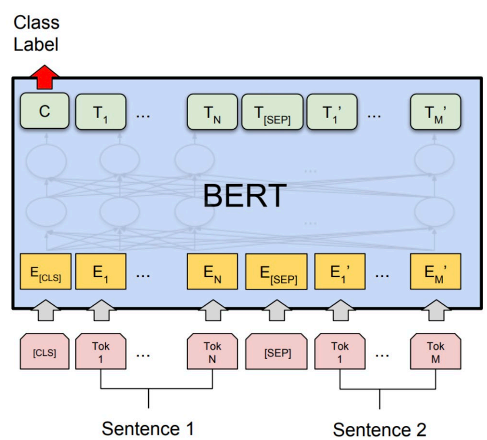
…… Heterogeneous ……

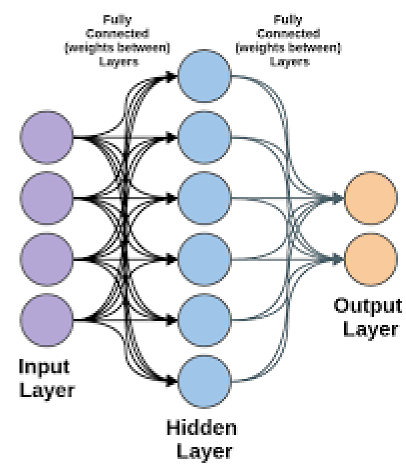
What matters for DL? Model?

Model is getting simpler.

Everything is you need.	Attention is all you need.	MLP is all you need.



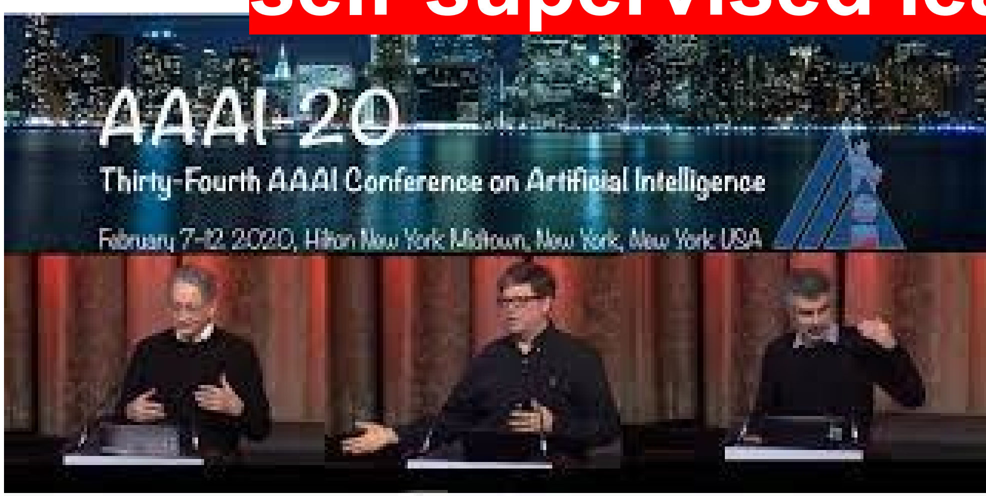




Ashish Vaswani et al. Attention Is All You Need. NeurIPS 2017.
Ilya Tolstikhin et al. MLP-Mixer: An all-MLP Architecture for Vision. NeurIPS 2021.
Luke Melas-Kyriazi. Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet. arXiv 2021.

What matters for DL? Learning!

self-supervised learning (SSL)



Do we have evidence?

MS MARCO Document Ra

date	description
2021/04/25	PROP_step400K base + doc2query top1000(ensemble v0.2)
2021/04/28	Knowledge Retrieval

KeyPhr

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
Jan 27, 2021	**RoR (Single model)** *Anonymous*	**74.9**	**72.2**	**16.4**
2 Sep 3, 2020	EL-QA (Single model) *JD AI Research*	74.6	71.6	16.3

,@5

SQuAD2.0 tests the ability of a system questions, but also abstain when prese based on the provided paragraph.

Rank	
	Huma *Stanf* (Rajpurk
1 Feb 21, 2021	FPN *Ant Service Intel*
2 Feb 24, 2021	IE-Net (ens *RICOH_SRC*
3 Apr 06, 2020	SA-Net on Alber *QIANX*
4 May 05, 2020	SA-Net-V2 (e *QIANX*

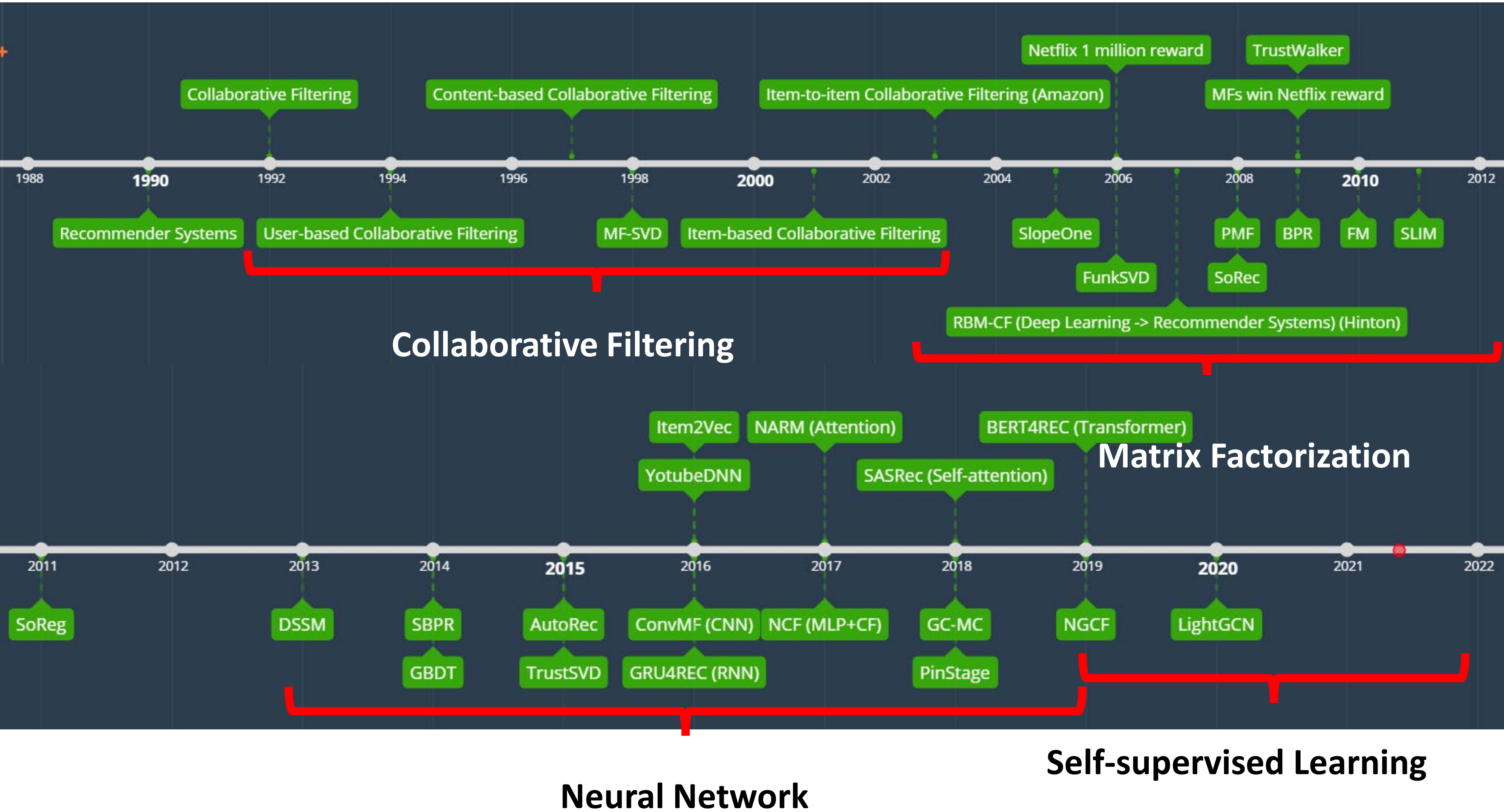
WMT 2014 EN-DE

Models are evaluated on the English-German dataset of the Ninth Workshop on Statistical Machine Translation (WMT 2014) based on BLEU.

Model	BLEU	Paper / Source
Transformer Big + BT (Edunov et al., 2018)	35.0	Understanding Back-Translation at Scale
DeepL	33.3	DeepL Press release
Admin (Liu et al., 2020)	30.1	Very Deep Transformers for Neural Machine Translation
MUSE (Zhao et al., 2019)	29.9	MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning
DynamicConv (Wu et al., 2019)	29.7	Pay Less Attention With Lightweight and Dynamic Convolutions

So where are we with SSL for recommendation?

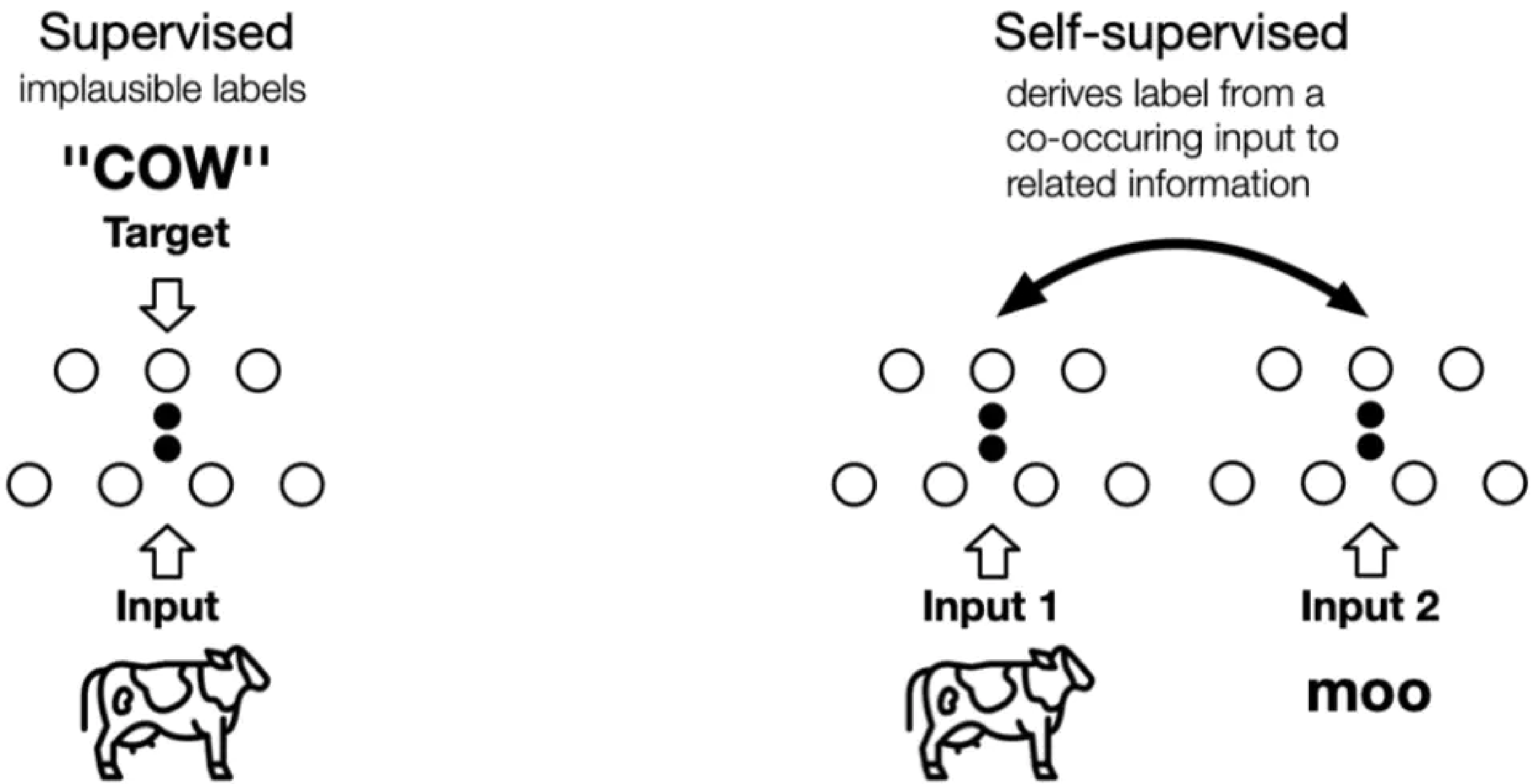
A brief history of recommender systems



Sorry if I miss anything.

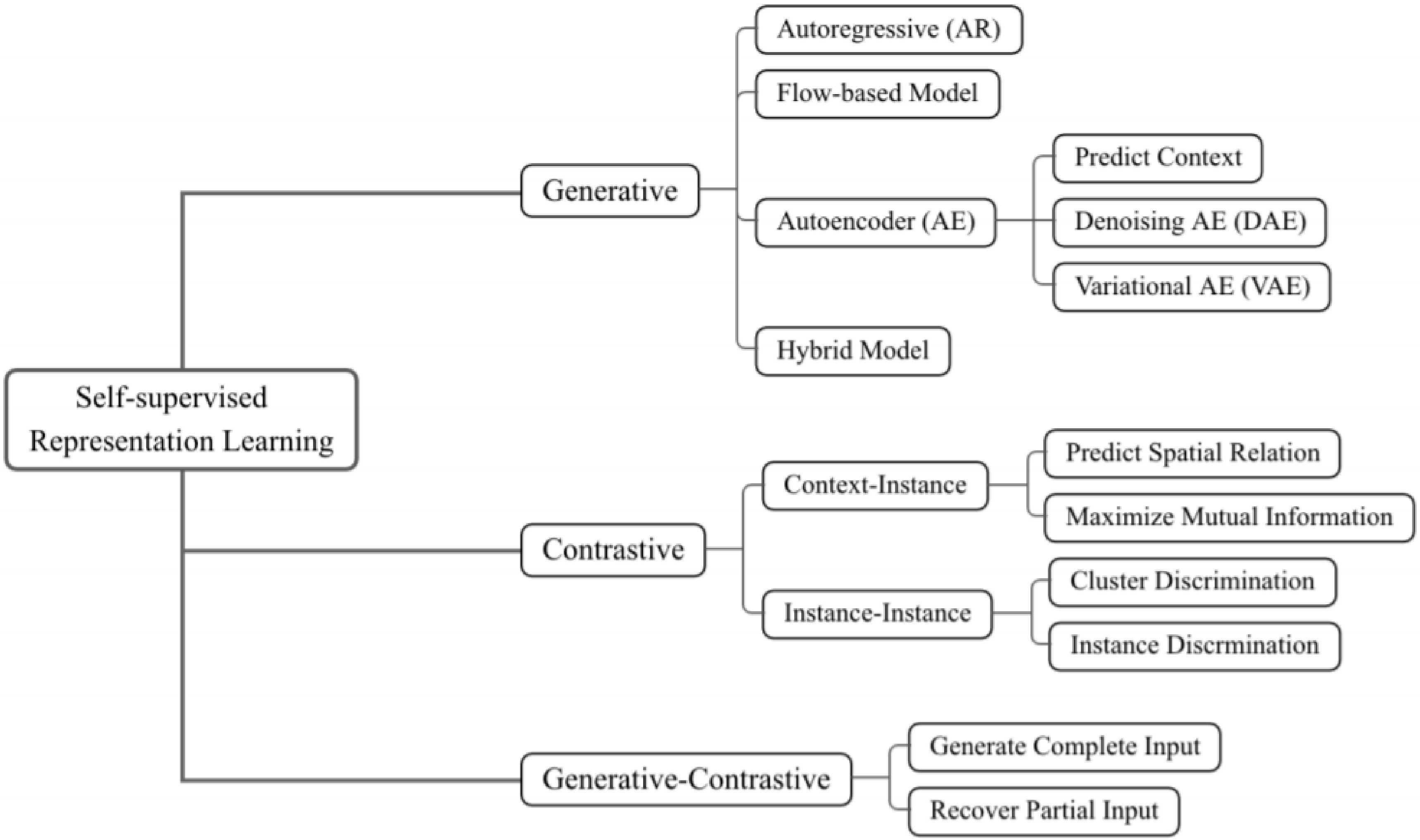
Then let’s talk about SSL a bit…

Supervised vs. unsupervised vs. self-supervised



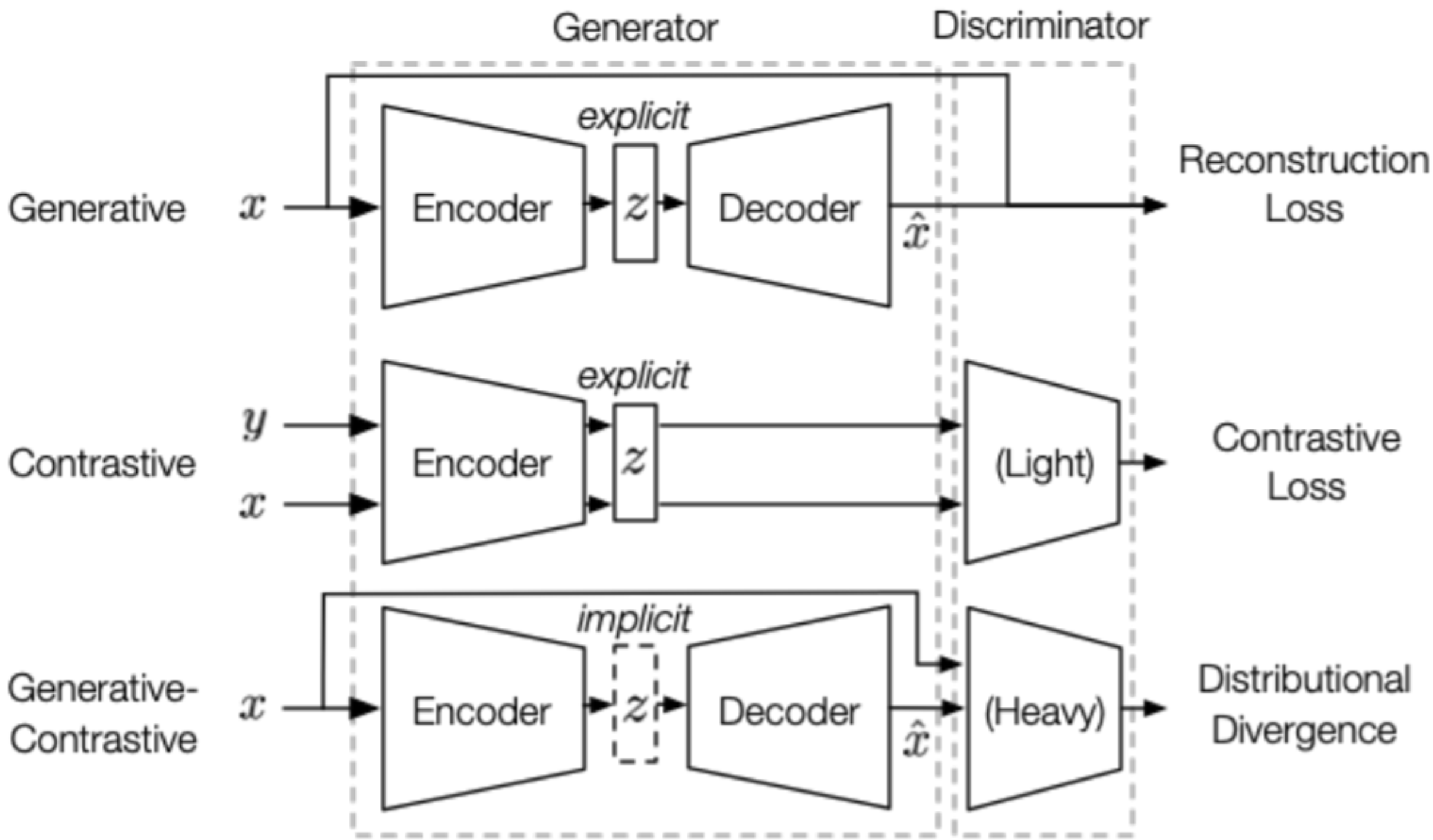
Xiao Liu et al. Self-supervised Learning: Generative or Contrastive. TKDE 2021.

Self-supervised learning in general



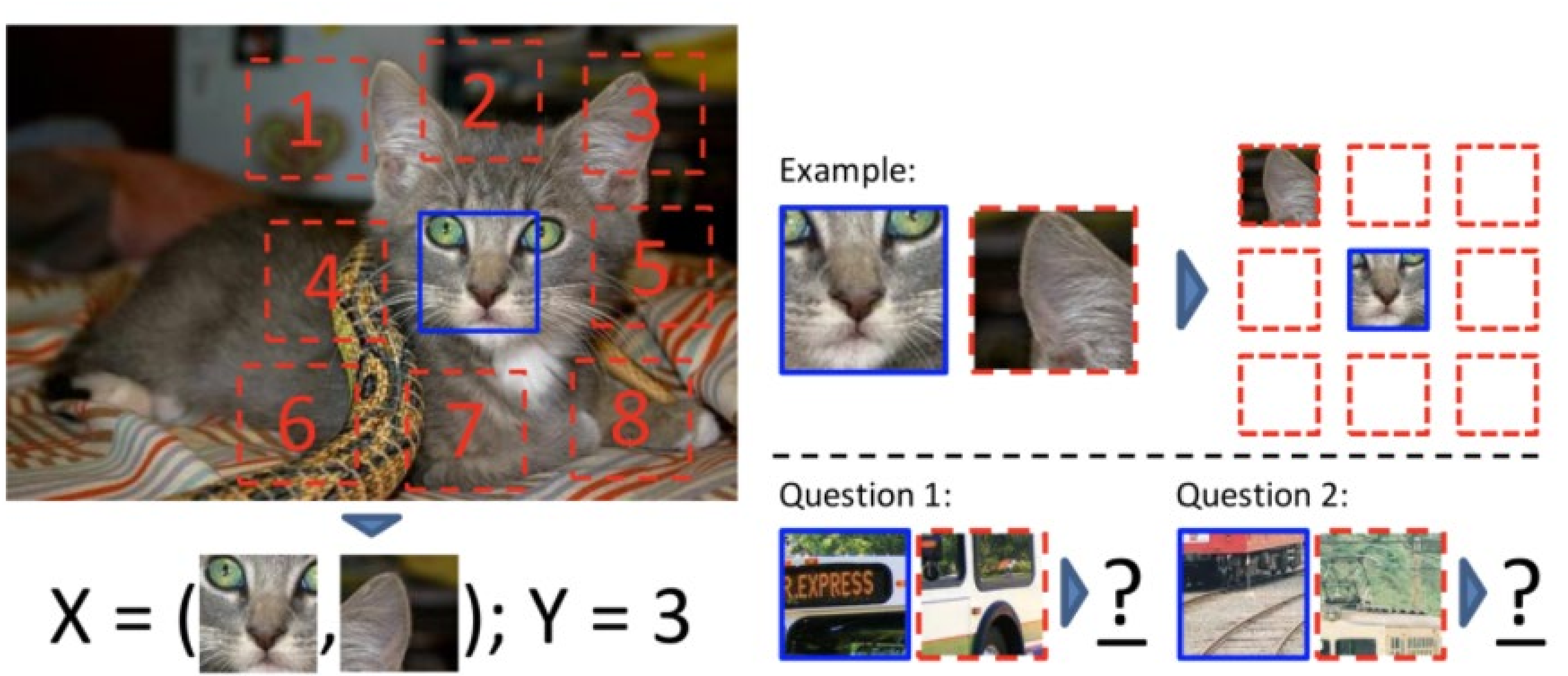
Xiao Liu et al. Self-supervised Learning: Generative or Contrastive. TKDE 2021.

Self-supervised learning in general



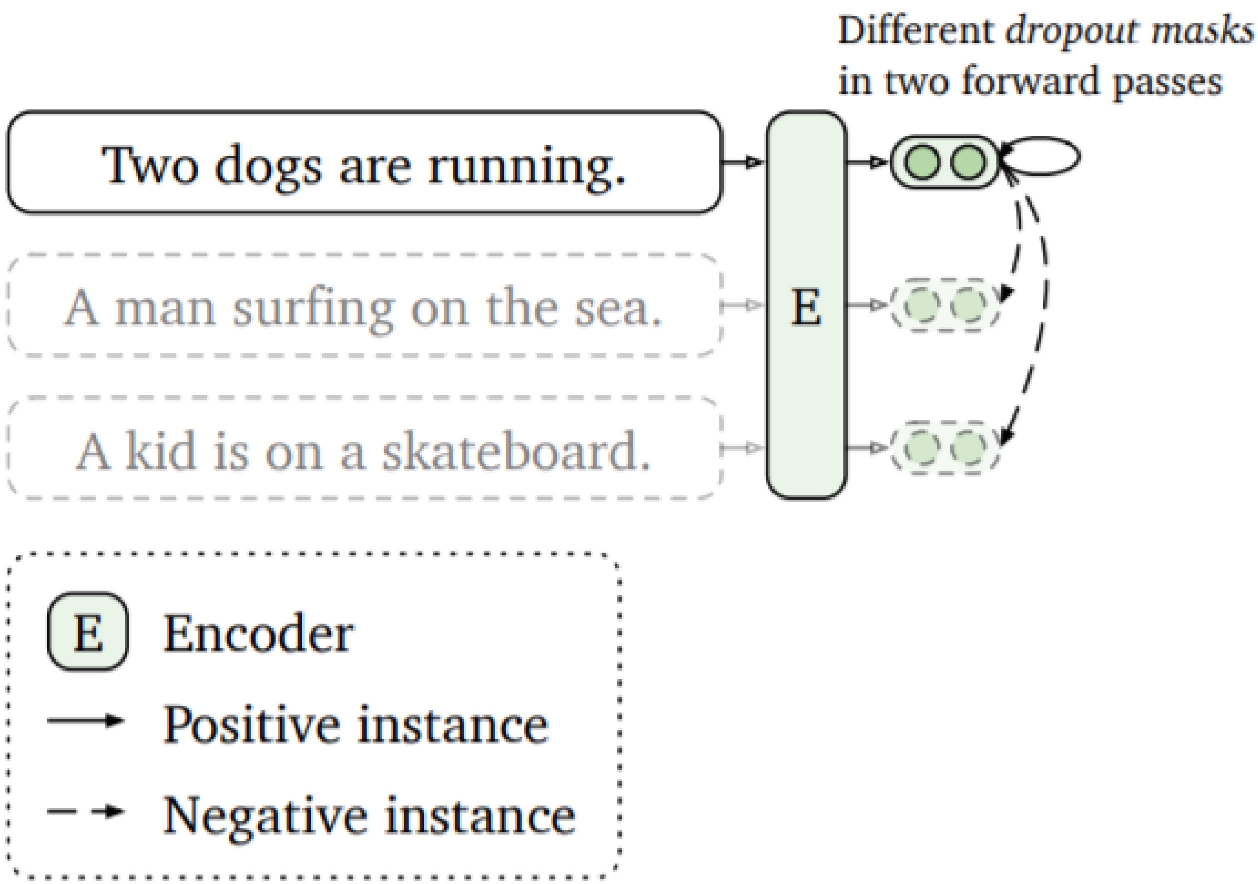
Xiao Liu et al. Self-supervised Learning: Generative or Contrastive. TKDE 2021.

Contrastive self-supervised learning



Carl Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015.

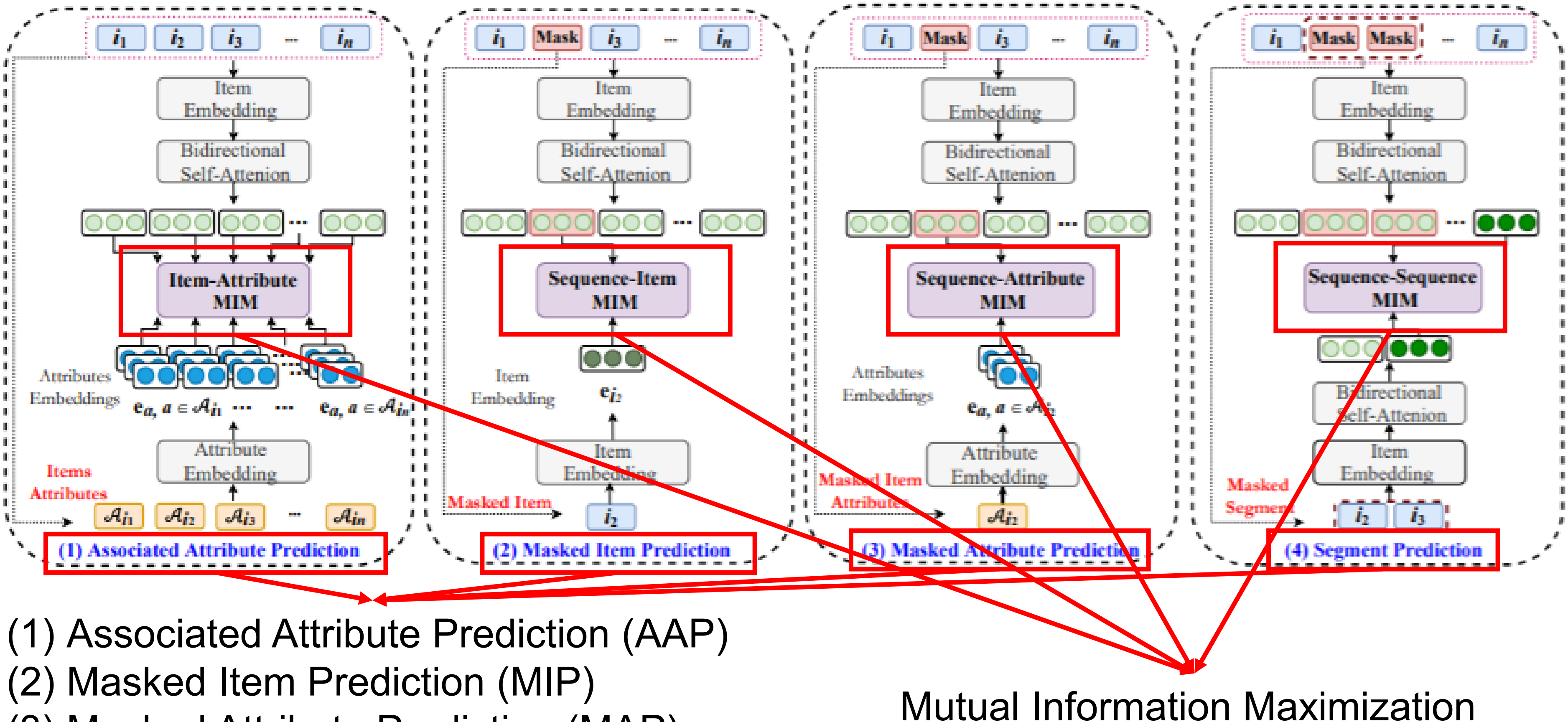
Contrastive self-supervised learning



Tianyu Gao et al. SimCSE: Simple Contrastive Learning of Sentence Embeddings. EMNLP 2021.

S³-Rec

Next Item Prediction overemphasizes the final performance, the association or fusion between context data and sequence data has not been well captured and utilized for sequential recommendation.

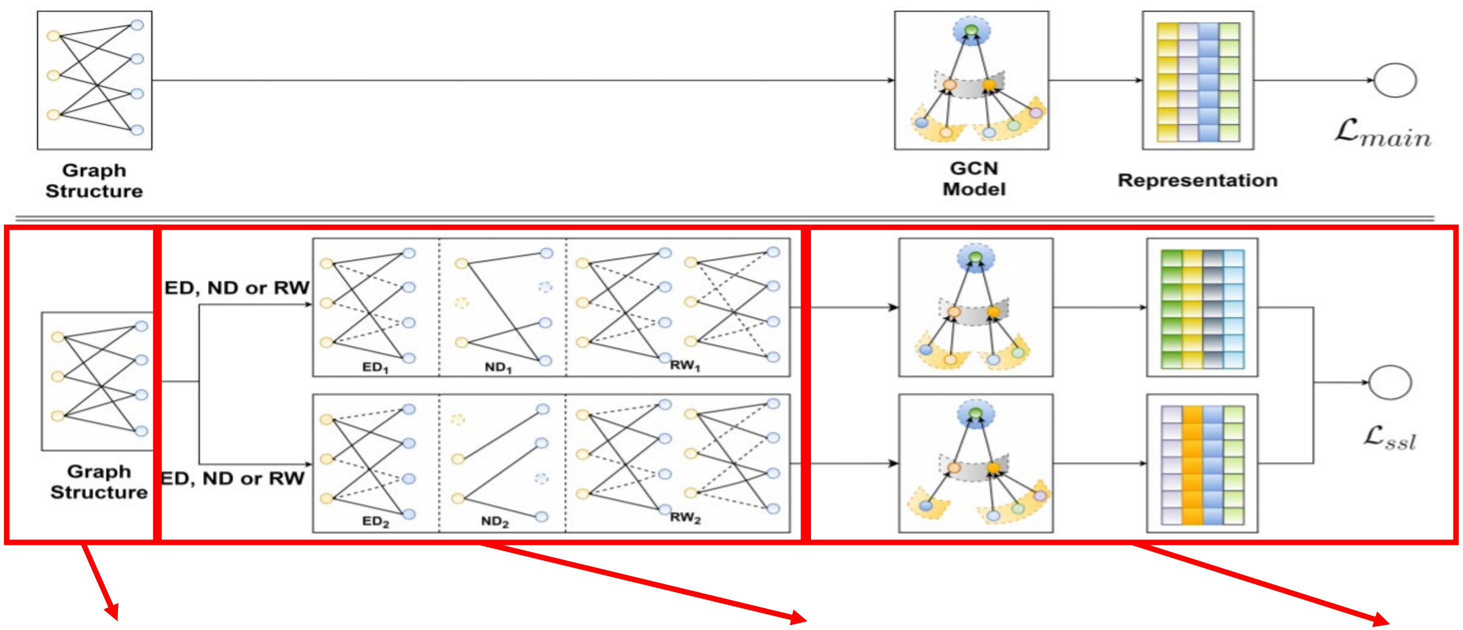


(1) Associated Attribute Prediction (AAP)
(2) Masked Item Prediction (MIP)
(3) Masked Attribute Prediction (MAP)
(4) Segment Prediction (SP)

Kun Zhou et al. S³-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. CIKM 2020.

SGL

(1) High-degree nodes exert larger impact on the representation learning.
(2) Representations are vulnerable to noisy interactions.



Interaction between users and items
→ Bipartite graph
→ Encoder: GCN

Augmentation:
1. Edge Dropout (ED)
2. Node Dropout (ND)
3. Random Walk (RW)

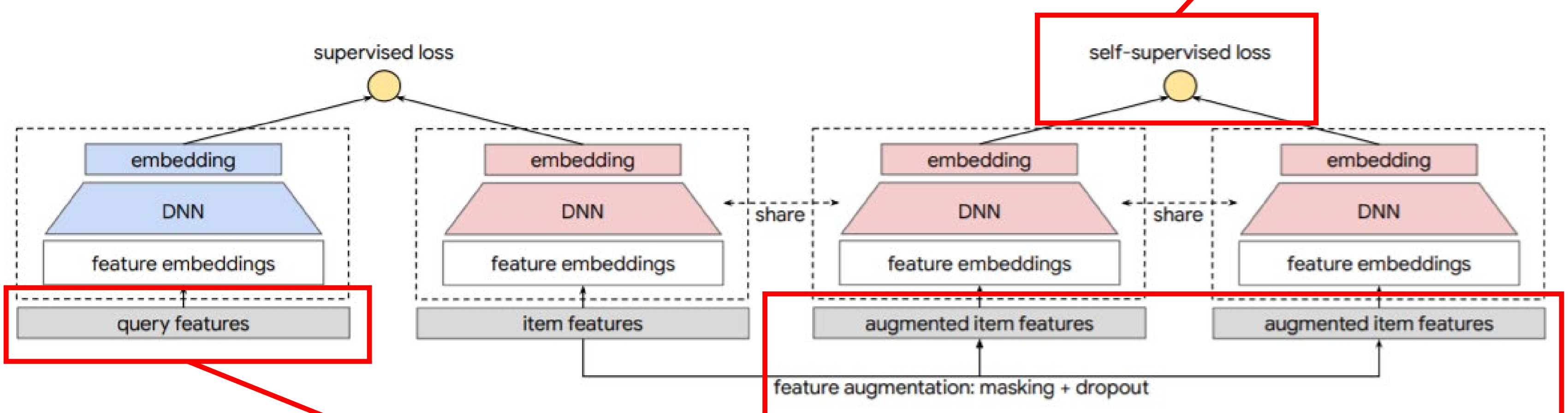
InfoNCE: Maximize the agreement of positive pairs and minimize that of negative pairs

Jiancan Wu et al. Self-supervised Graph Learning for Recommendation. SIGIR 2021.

MSSL

Data sparsity: With millions to billions of items in the corpus, users tend to provide feedback for a very small set of them, causing a power-law distribution.

InfoNCE: Maximize the agreement of positive pairs and minimize that of negative pairs.



(i) Personalized recommendation: when the query is a user;
(ii) Item to item recommendation: when the query is also an item; and
(iii) Search: when the query is a piece of free text.

Augmentation:

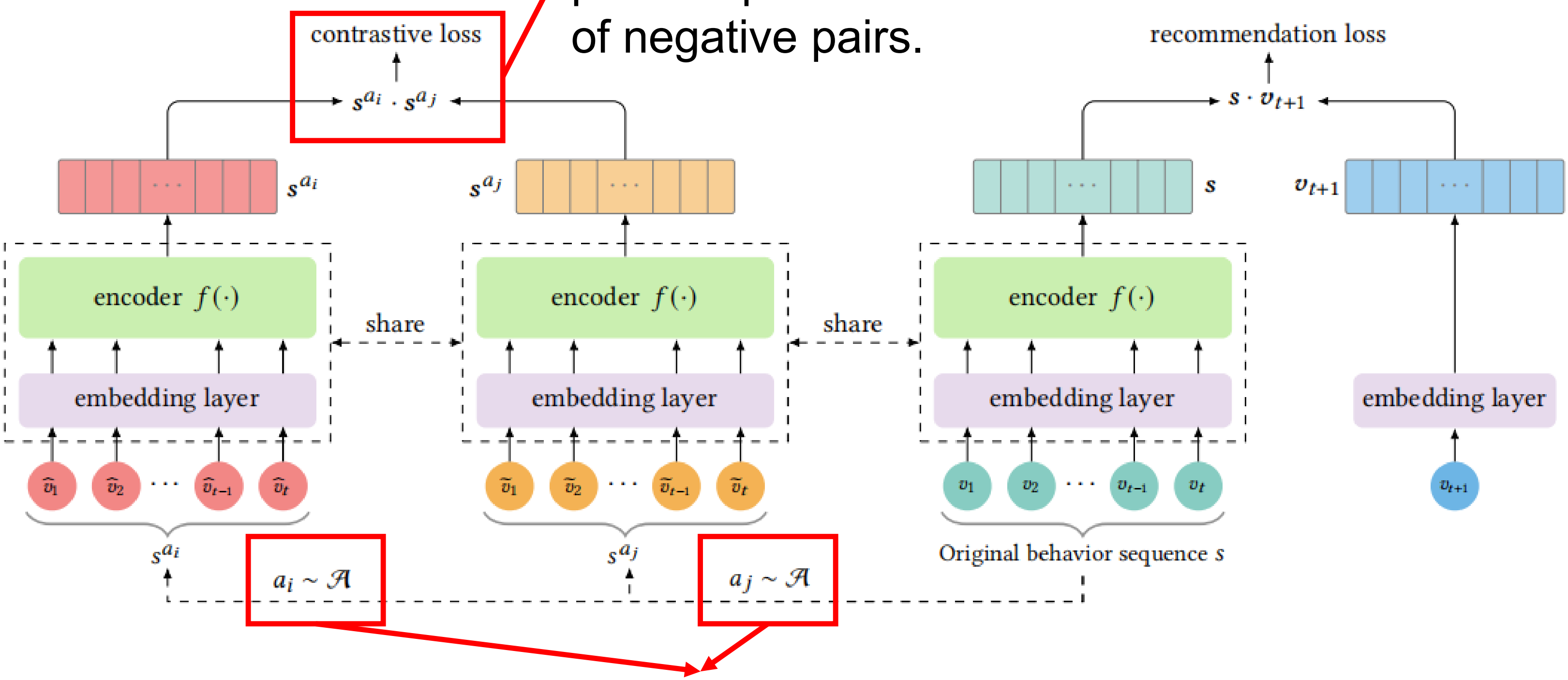
1. Masking: Applying a masking pattern on the set of item features.
2. Dropout: Randomly dropouting categorical feature values.

Tiansheng Yao et al. Self-supervised Learning for Large-scale Item Recommendations. CIKM 2021.

CL4SRec

The data sparsity of SR makes it difficult to get high-quality user representations.

CE: Maximize the agreement of positive pairs and minimize that of negative pairs.



Augmentation:

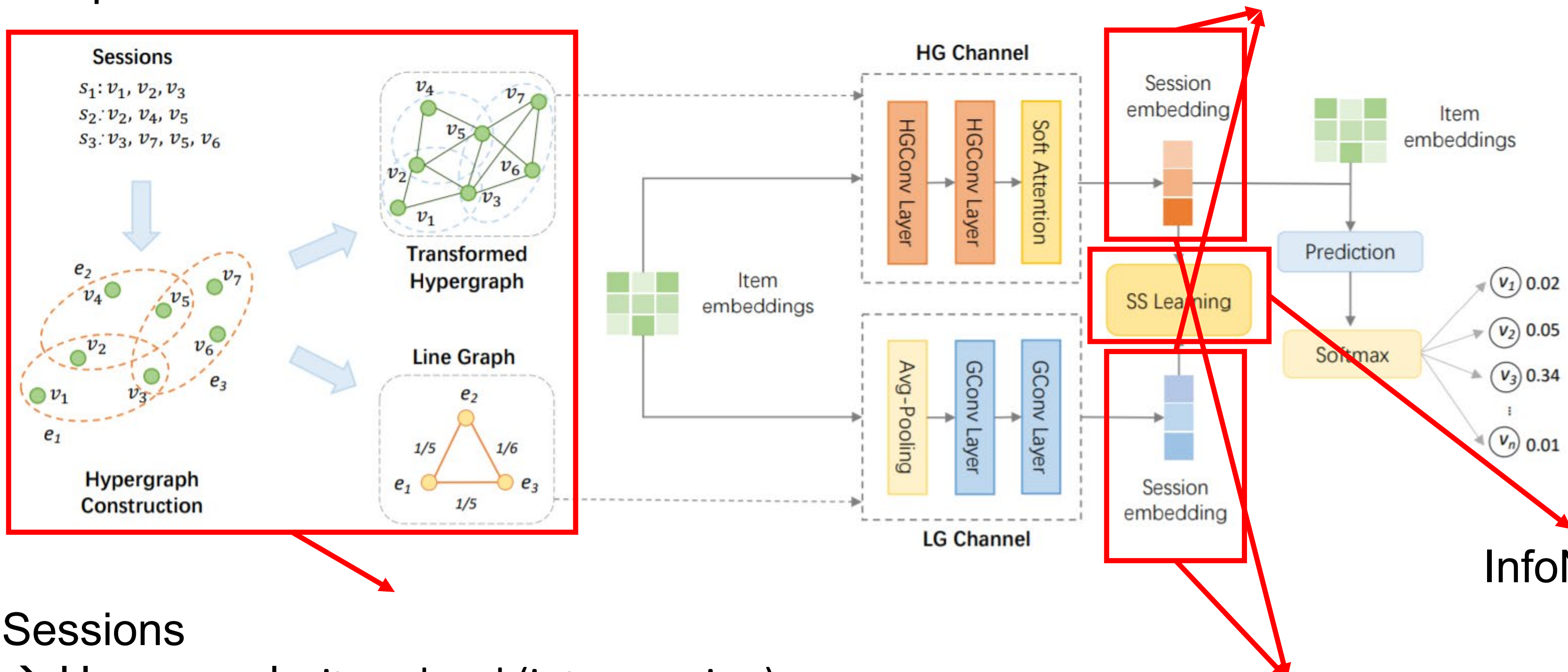
1. Item crop: Randomly selecting a continuous sub-sequence.
2. Item mask: Randomly masking a proportion of items.
3. Item reorder: Randomly shuffling a proportion of items.

Xie Xu et al. Contrastive Learning for Sequential Recommendation. SIGIR 2021.

DHCN

Regarding the two channels as two views characterizing different aspects of sessions.

The two groups of embeddings know little about each other but can mutually complement.



InfoNCE

Sessions

→ Hypergraph: item-level (intra-session) structural information

→ Line graph: session-level (inter-session) structural information

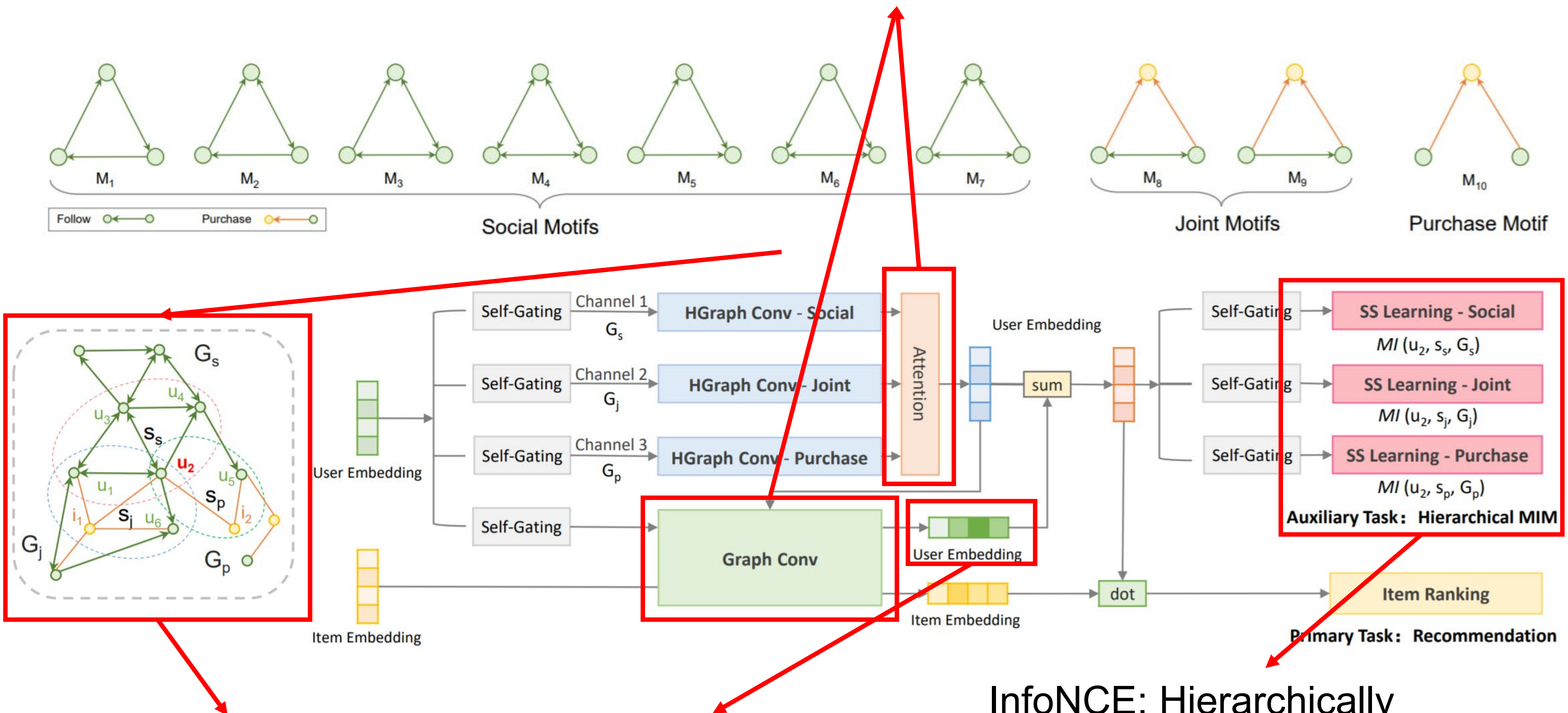
→ Encoder: GNN

Augmentation:
Corrupted session embedding, aka. row-wise and column-wise shuffling.

Xin Xia et al. Self-Supervised Hypergraph Convolutional Networks for Session-based Recommendation. AAAI 2021.

MHCN

The aggregation operations might lead to a loss of high-order information.



social network and user-item interaction
→Motifs→ Hypergraph
→Encoder: GNN

Augmentation:
Corrupted user embedding, aka. row-wise and column-wise shuffling.

InfoNCE: Hierarchically maximizing the mutual information between representations of the user, the user-centered sub-hypergraph, and the hypergraph.

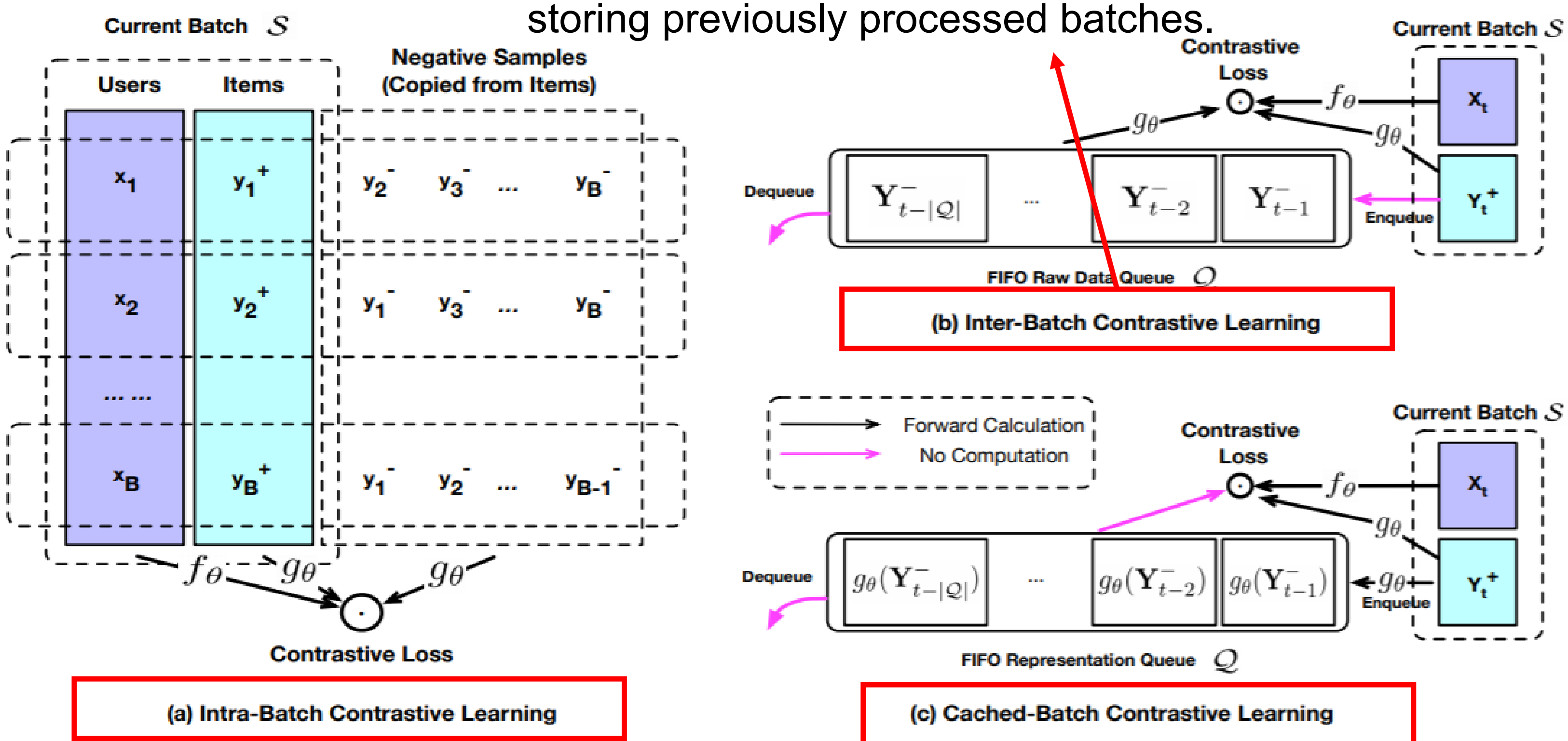
Junliang Yu et al. Self-Supervised Multi-Channel Hypergraph Convolutional Network for Social Recommendation. WWW 2021.

CLRec

Live recommender systems face severe exposure bias.

Reducing the exposure bias by adjusting sampling strategies.

(b) Sampling instances from a fixed-size FIFO queue storing previously processed batches.



(a) Sampling instances in the present batch.

(c) Differing from variant (b) in that the queue caches the representations.

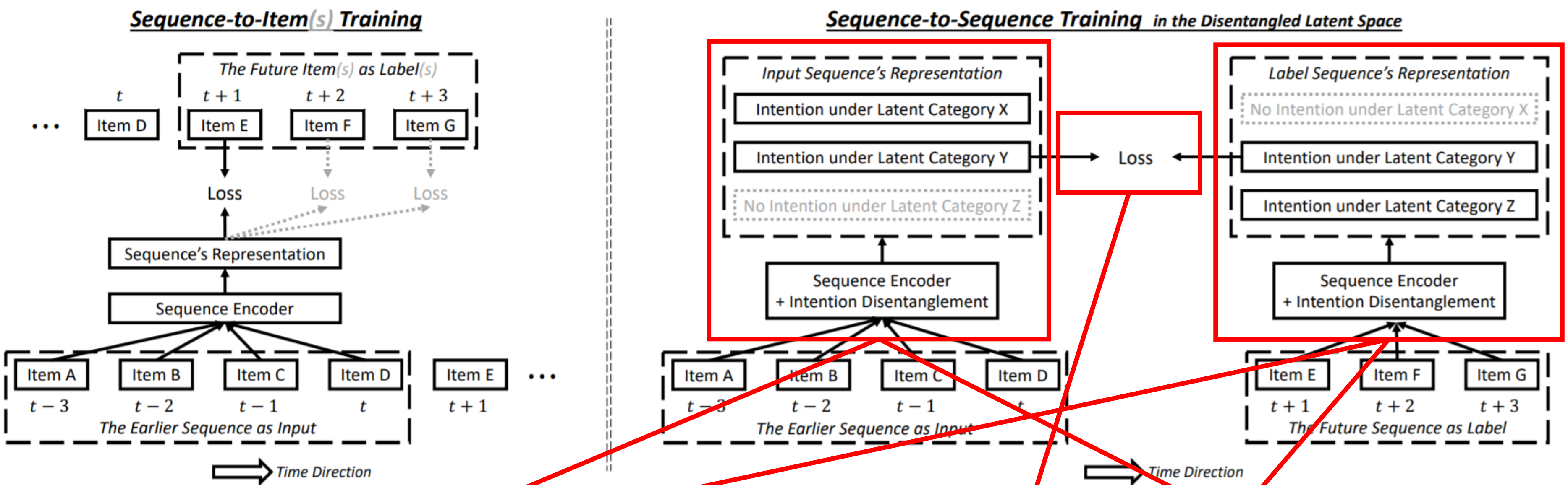
A popular choice of contrastive loss is equivalent to reducing the exposure bias.

Chang Zhou et al. Contrastive Learning for Debiased Candidate Generation in Large-Scale Recommender Systems. KDD 2021.

Seq2seq for recommendation

Seq2item training is myopic and can easily lead to non-diverse recommendation lists.

Mining extra signals for supervision by looking at the longer-term future.



SASRec+Attention
→Intention clustering
→Intention weighting
→Intention aggregation

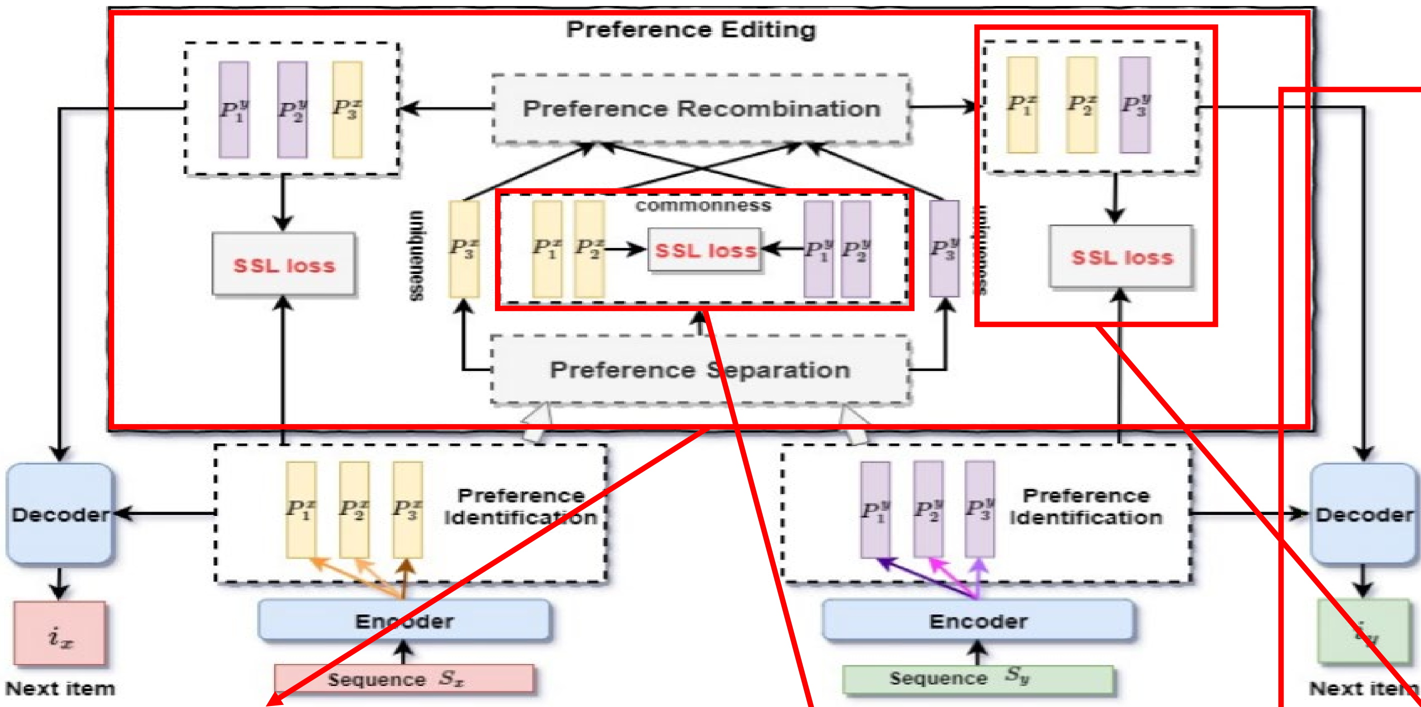
$$\mathcal{L}_{s2s}(\theta, u, t, k) = -\ln p_{\theta}\left(\boxed{\phi_{\theta}^{(k)}(\mathbf{x}_{T_u:t+1}^{(u)})} \mid \boxed{\phi_{\theta}^{(k)}(\mathbf{x}_{1:t}^{(u)})}\right)$$

Jianxin Ma et al. Disentangled Self-Supervision in Sequential Recommenders. KDD 2020.

MrTransformer

The relation between different sequences is under explored.

Discriminating the common and unique preference representations between a pair of sequences.



CE: Next item prediction based on the recombined representation.

Transformer+Coverage
→Preference editing
→Preference identification
→Preference separation
→Preference recombination

MSE: The common representation is close enough to each other.

MSE: The recombined representation is close enough to the original preference representation.

Muyang Ma et al. Improving Transformer-based Sequential Recommenders through Preference Editing. Under review at TOIS.

How pretraining helps recommendation?

Learning from Imperfection

Make full use of available data.

Make use of more data.

Sparse

……

Noisy

Low-resource

……

Heterogeneous

……

Cold-start

……

……

Multimodal pretraining

Computational efficiency

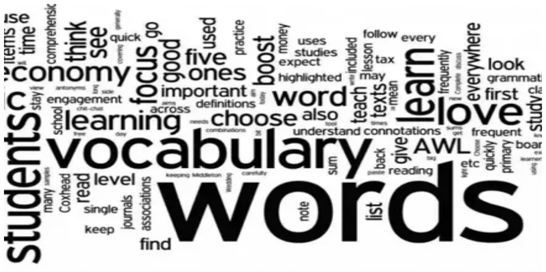
Model architecture

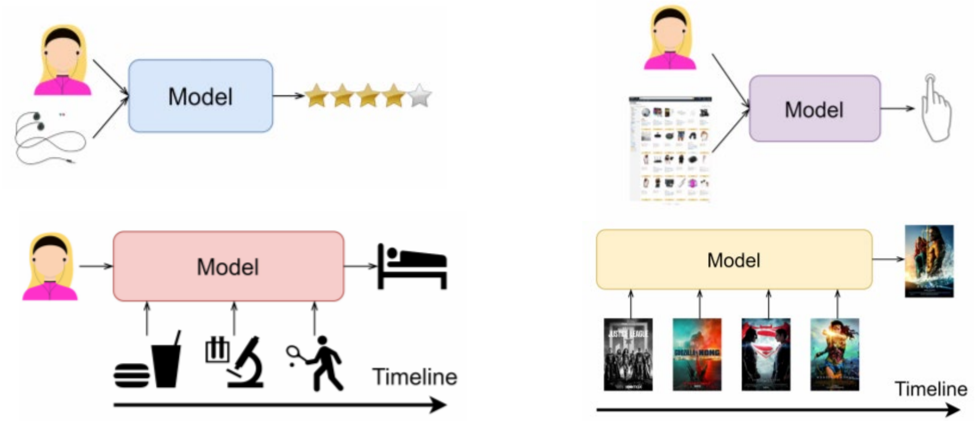
……

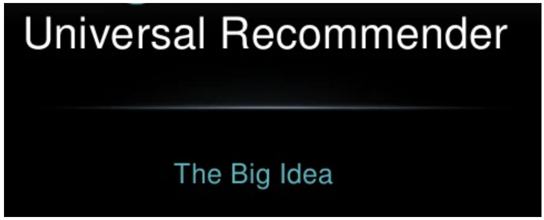
Theoretical foundations

……

What is unique for recommendation pretraining?







Item

User

Thank you for your attention!

Pengjie Ren
renpengjie@sdu.edu.cn
https://pengjieren.github.io/
https://ir.sdu.edu.cn/

山东大学
SHANDONG UNIVERSITY

山东大学信息检索实验室
Information Retrieval Lab