Attention gets you the right size and fit in fashion

Karl Hajjar*, Julia Lasserre, Alex Zhao*, Reza Shirvany

Zalando SE

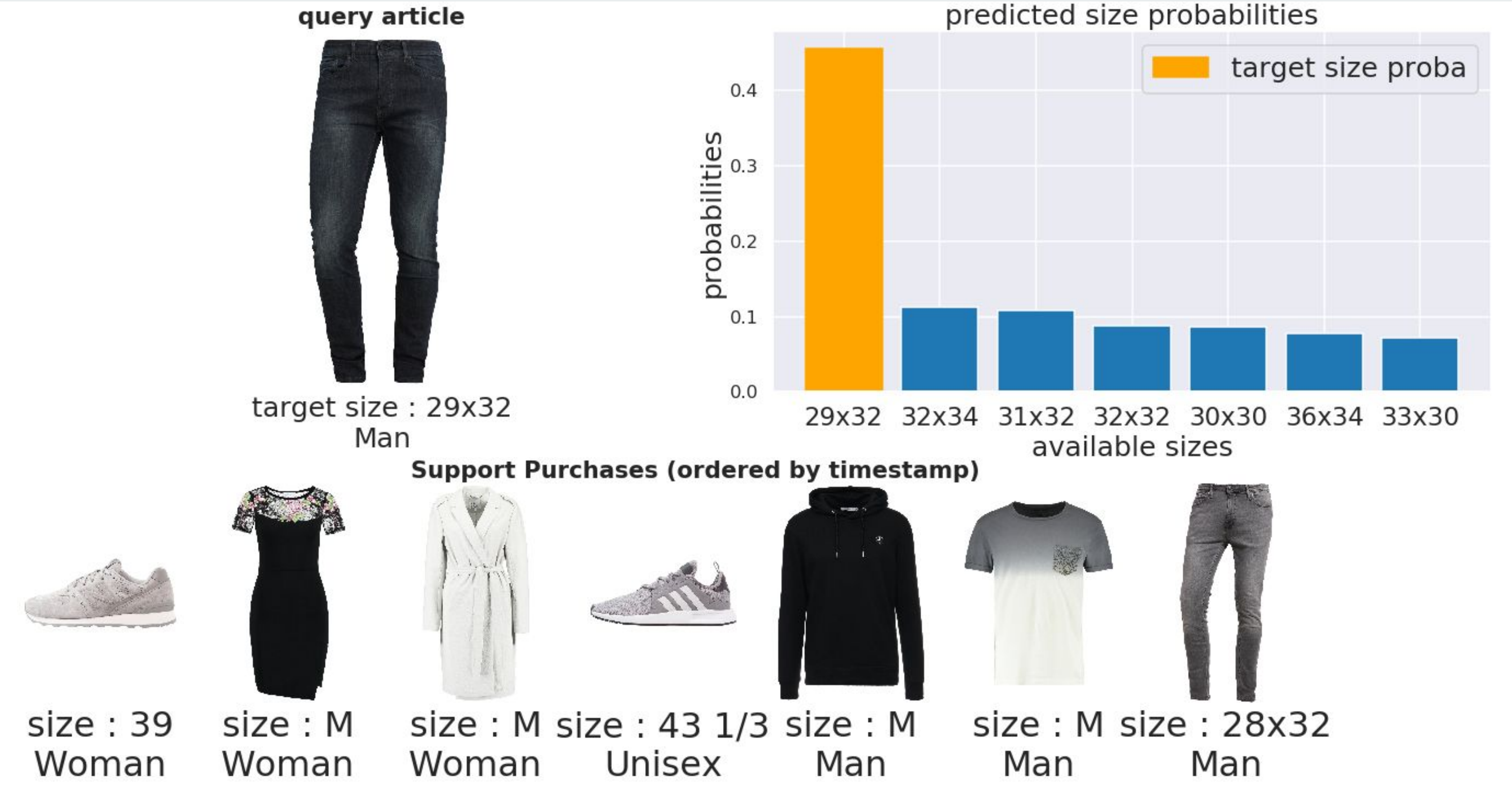
*Speakers

Outline

- The size recommendation problem
- Recent approaches
- A transformer architecture for size recommendation
- Thorough evaluation
  - Cross-category recommendation
  - Multi-user accounts
  - Online scenario

The size recommendation problem

Size Reco as a ML problem



Why is size reco difficult ?

- Sparsity in the article-size pairs encountered.

- Noise :
  - The “right size” is subjective.
  - Multiple users behind a single account.
  - Size systems/standards vary from brand to brand.

- Emotionally engaging topic : what if the recommended size differs from the customer’s expectation ?
  → see ***Vecchi et. al (2015)***

Approaches to size recommendation

- Size recommendation has gained attention only in the last 3-4 years.

- Different methods :
  - Size tables and aggregated article measurements (old-fashioned).
  - Using customer metadata : images / scans, questionnaires (personal info).
  - Using the **history of past purchases** of a customer.

- Emerging body of work published on the last type of approaches since 2017 as it does **not** require any personal data.

What would an ideal size recommender do?

1. Perform well on metrics of interest (*e.g.* accuracy)
2. Naturally handle the various existing size systems
3. Adapt to new customers / information without retraining or fine-tuning (**online scenario**)
4. Leverage cross-category information
5. Handle multi-user accounts
6. Be transparent when making a size prediction → ***interpretability***

Recent approaches

What would an ideal recommender do?

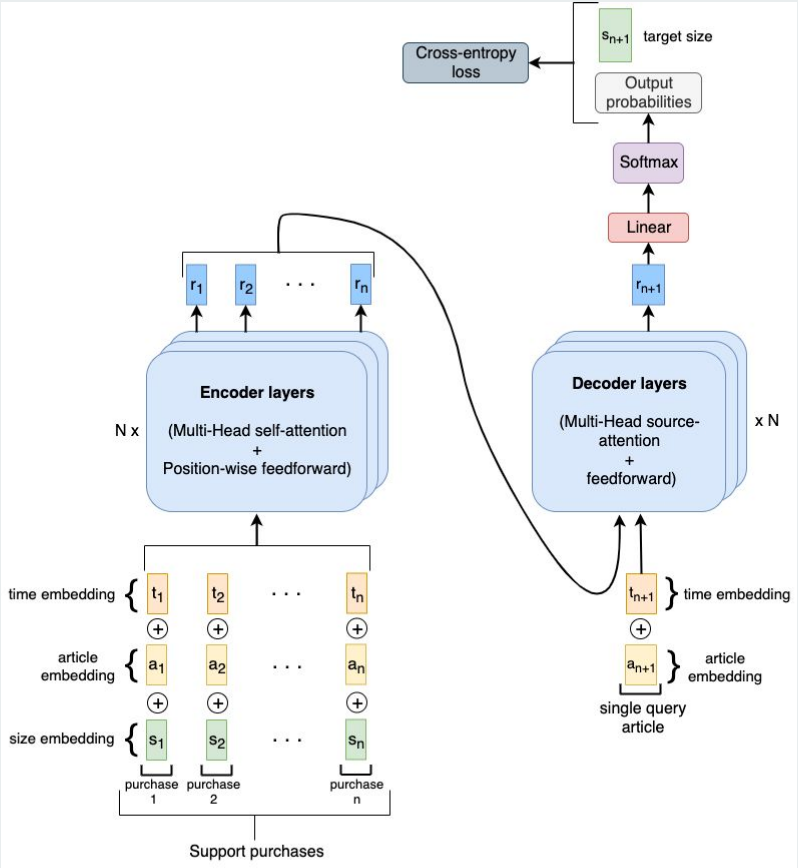
	Per-category[1]	SFNet[2]	MetalSF[3]
1. Perform well on metrics of interest	✅	✅	✅
2. Naturally handle the various existing size systems	❌	✅	✅
3. Adapt to new customers / information without retraining or fine-tuning	❌	❌	✅
4. Leverage cross-category information	❌	✅	✅
5. Handle multi-user accounts	😐	😐	😐
6. Be transparent when making a size prediction → ***interpretability***	❌	❌	😐

1. ***Sembium et. al (2017, 2018), Abdulla et. al (2017), Guigourès et. al (2018), Dogani et. al (2019)***
2. ***Sheik et. al (2019)***
3. ***Lasserre et. al (2020)***

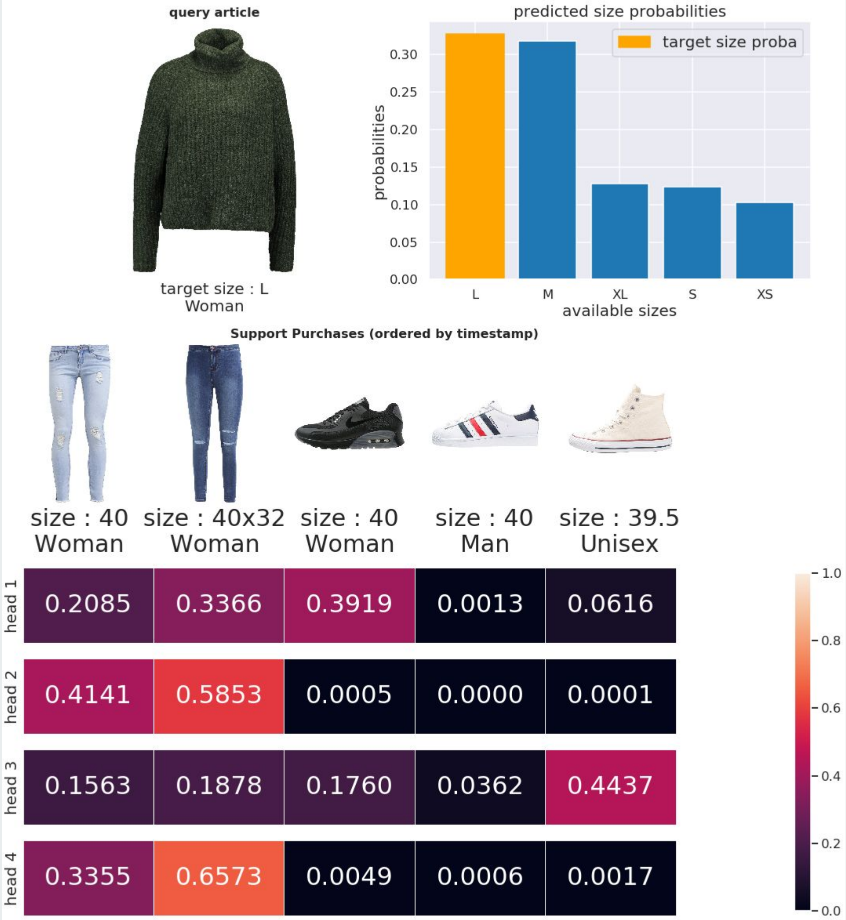
Our approach using Transformers

A Transformer architecture using attention

- Standard Transformer architecture: flexible inputs
- *"Translating"* from an article to a size
- Source sentence = previous purchases of customer
- Target sentence = new article whose size has to be decoded



Attention weights interpretability



Weights computed by dot-product attention ***Vaswani et. al (2017)***

Previous purchases linearly combined using those weights

These weights

- are positive and sum to 1
- are easy to interpret
- can be combined when multi-head attention is used

What would an ideal recommender do?

		Per-category	SFNet	MetalSF	**Attention**
1.	Perform well on metrics of interest	✅	✅	✅	✅
2.	Naturally handle the various existing size systems	❌	✅	✅	✅
3.	Adapt to new customers / information without retraining or fine-tuning	❌	❌	✅	✅
4.	Leverage cross-category information	❌	✅	✅	✅
5.	Handle multi-user accounts	😐	😐	😐	😐
6.	Be transparent when making a size prediction → ***interpretability***	❌	❌	😐	✅

Model evaluation

Model evaluation

1. On past purchases in an offline scenario
2. On cross-category samples
3. On multi-user accounts
4. On past purchases in an online scenario

Performance - *Offline*

- Test samples with customers belonging to the training set
- Purchases of customer in the training set are considered as "previous" purchases when testing

	log lik.	top-1	top-2	top-3	mAUC
Bayesian (*Guigourès et. al*)	-1.46	0.47	0.72	0.84	0.79
PSE (*Dogani et. al*)	-1.47	0.53	0.77	0.87	0.82
SFnet (*Sheikh et. al*)	-1.20	0.55	0.79	0.89	0.85
MetalSF (*Lasserre et. al*)	**-1.04**	0.60	0.83	0.92	**0.89**
attention-based (ours)	-1.11	**0.61**	**0.84**	**0.93**	0.88

Performance - *Cross-category*

Test purchases in a specific category C from training customers that have not shopped in C before

(a) Upper garments (13*k* test samples).

	log lik.	top-1	top-2	top-3	mAUC
popularity	-1.82	0.31	0.60	0.75	0.64
SFnet (*Sheikh et. al*)	-1.43	0.37	0.62	0.77	0.67
MetalSF (*Lasserre et. al*)	**-1.30**	0.41	0.69	0.86	0.73
attention-based (ours)	-1.60	**0.45**	**0.73**	**0.89**	**0.75**

(b) Lower garments (15*k* test samples).

	log lik.	top-1	top-2	top-3	mAUC
popularity	-2.54	0.24	0.45	0.60	0.71
SFnet (*Sheikh et. al*)	-1.79	0.35	0.57	0.71	0.75
MetalSF (*Lasserre et. al*)	-1.60	0.38	0.61	0.76	0.80
attention-based (ours)	**-1.30**	**0.40**	**0.64**	**0.78**	**0.81**

Performance - *Multi-user accounts*

Test purchases in a specific target gender G from training customers

These customers are grouped by type of history

- cold-start: no prior purchases of G articles
- consistent: only prior purchases of G articles
- mixed: various target genders in prior purchases

	Bayesian [12]		PSE [15]		SFnet [16]		MetalSF [17]		**attention-based**	
target gender	men	women	men	women	men	women	men	women	men	women
cold-start (no related history)	0.28	0.28	0.28	0.28	0.32	0.30	0.34	0.30	**0.37**	**0.34**
consistent (always same gender)	0.44	0.46	0.50	0.51	0.52	0.54	0.55	0.58	**0.59**	**0.60**
mixed (various genders in history)	0.44	0.47	0.49	0.54	0.48	0.55	**0.55**	**0.61**	0.54	**0.61**

Performance - *New customers: the online scenario*

Test purchases from test customers added one by one: online scenario

The first article's size is predicted using popularity
The second article's size is predicted based on the first purchased size
The nth article's size is predicted based on the first (n-1) purchased sizes

	log lik.	top-1	top-2	top-3	mAUC
MetalSF (*Lasserre et. al*)	**-1.23**	0.59	0.79	0.88	0.87
attention-based (ours)	-1.34	**0.60**	**0.81**	**0.90**	0.87

2. Attention model Advantages

- Great flexibility
  → can adapt to new customers and articles

- Trained once on all categories and leverages cross-category information

- Goes towards **interpretability**

Future work

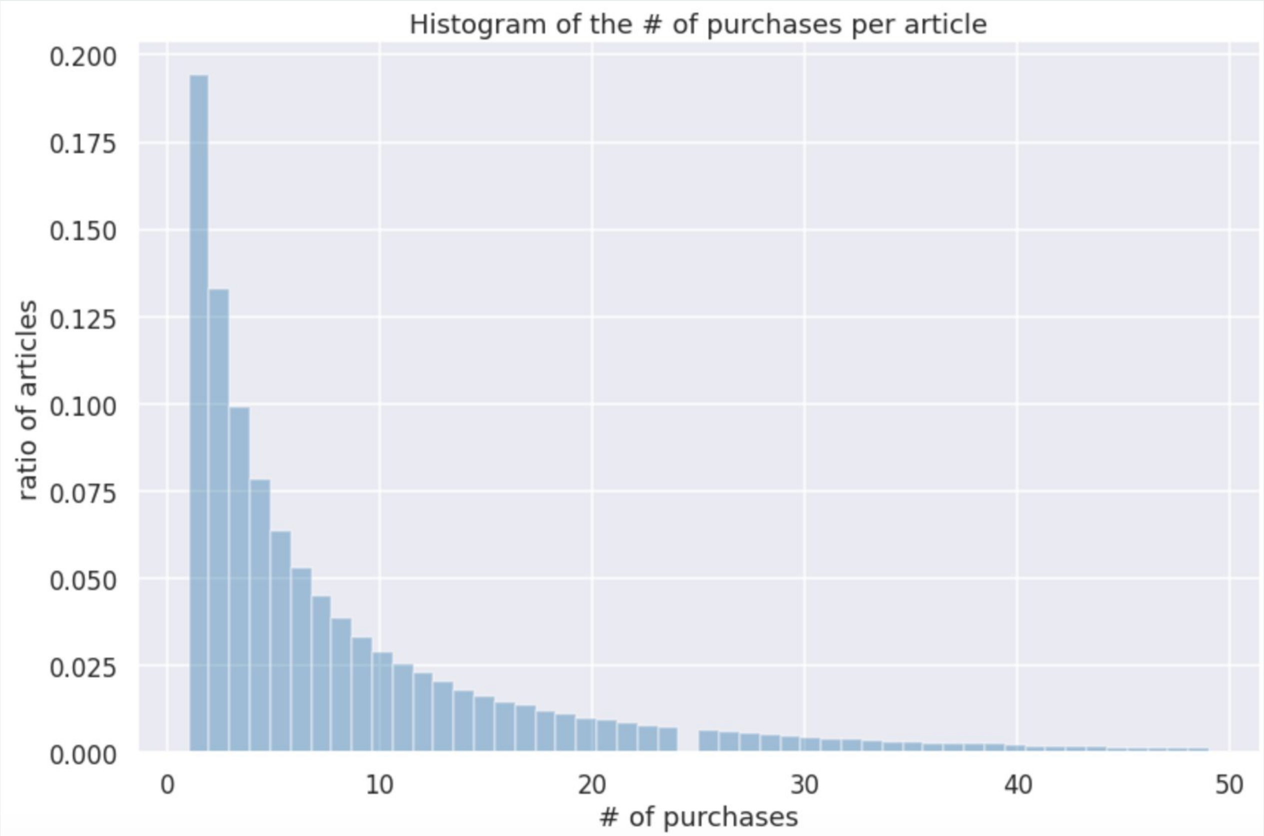
Future work

- Look at the embeddings
  - Customers that purchase similarly
  - Articles that size similarly
  - Sizes that are similar (conversion from brand to brand)
- Integrate more article meta-data such as fit, shape and material
- Translate weights into meaningful explanations for the customer

Thank you for your attention !

Backup slides

Sparsity in the articles purchased

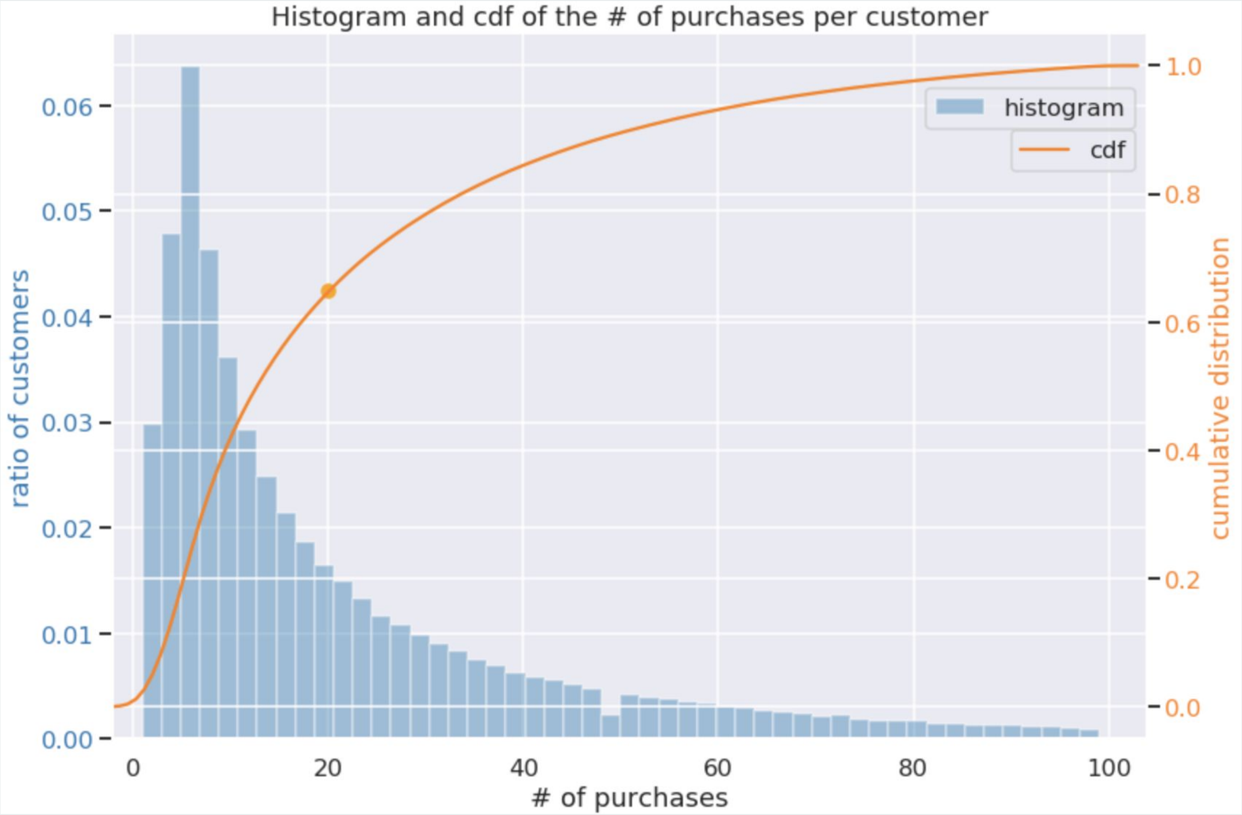


+ ~10 sizes per article

=

even sparser

Sparsity in purchases per customer

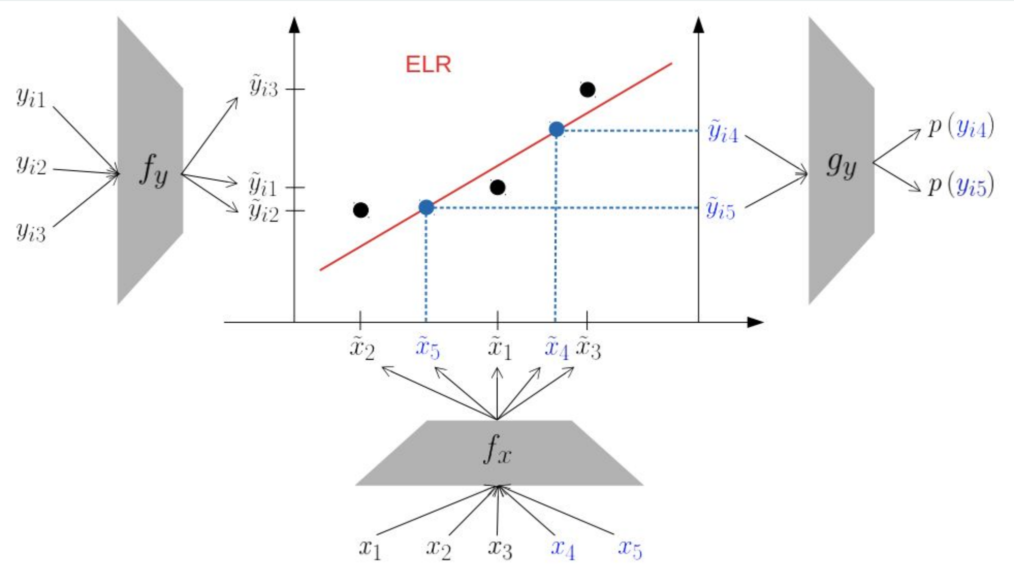


> 60% of customers have 20 purchases or less

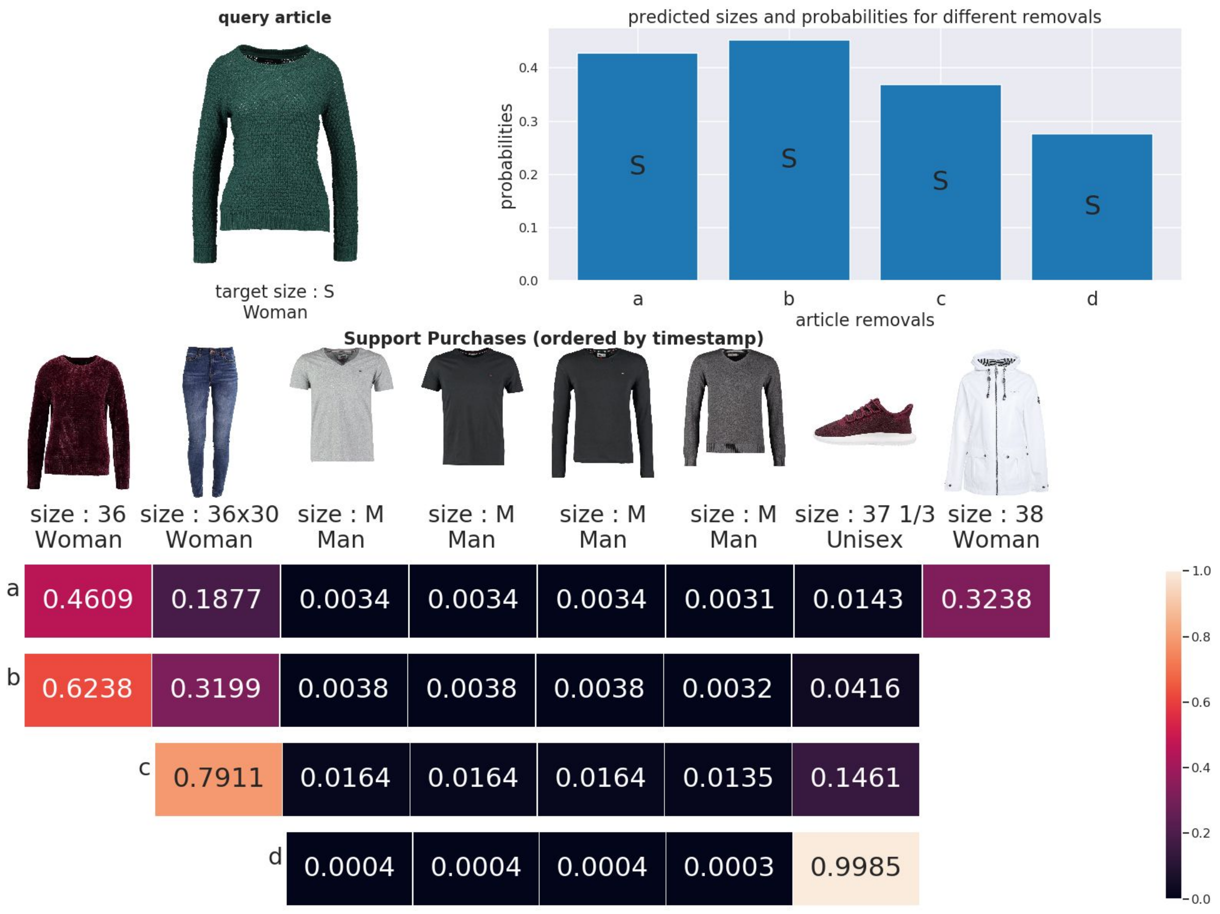
Very recent Meta-learning approach : MetalSF

- Each customer is a new task
- Article embeddings + size embeddings + Embedded Linear Regression
- At test time, ELR trained on previous purchases
- Size is decoded from the output of ELR

MetalSF ***Lasserre et. al (2020)***



Attention adapts to the purchase history



Changing the purchase history
⇒
attention adapts its focus to get the right amount of information for prediction

Learning the “true” size of articles & customers

- Articles and Customers have “true” but unknown latent sizes.

- Use history of orders and fitness feedback (‘too big’, ‘fit’, ‘too small’).

- Learn latent sizes by matching the customer size to the article size corrected by the fitness feedback.

  → using a latent factor model ***Sembium et. al (2017)***
  → using a hierarchical Bayesian model ***Guigourès et. al (2018)***

Matching customer & article embeddings

- Learn article embeddings
    - Pre-training ***Abdulla et. al (2017)***
    - As part of the model ***Dogani et. al (2019), Sheik et. al (2019)***

- Learn customer embeddings
    - Averaging article embeddings ***Abdulla et. al (2017), Dogani et. al (2019)***
    - Learning them separately ***Sheik et. al (2019)***

- Predict a size by combining the learned customer and article embeddings
    - XGBoost ***Abdulla et. al (2017)***
    - Neural network ***Sheik et. al (2019)***
    - Inner products ***Dogani et. al (2019)***

Recent approaches (2017-2019)

- Series of recent work “matching” customer information to article information: ***Sembium et. al (2017,2018), Abdulla et. al (2017), Guigourès et. al (2018), Dogani et. al (2019), Sheik et. al (2019)***

- ***Sembium et. al (2017-2018), Guigourès et. al (2018)*** apply to numerical size systems

- Only **SFNet *Sheik et. al (2019)*** trains a single model for all fashion categories

- In all those works, customer information is summarized in a single vector : direct access to past purchases is lost at prediction time