Lecture at Waseda University

Extreme Multi-label Classification

July 1, 2022

Chikara Hashimoto
Rakuten Institute of Technology

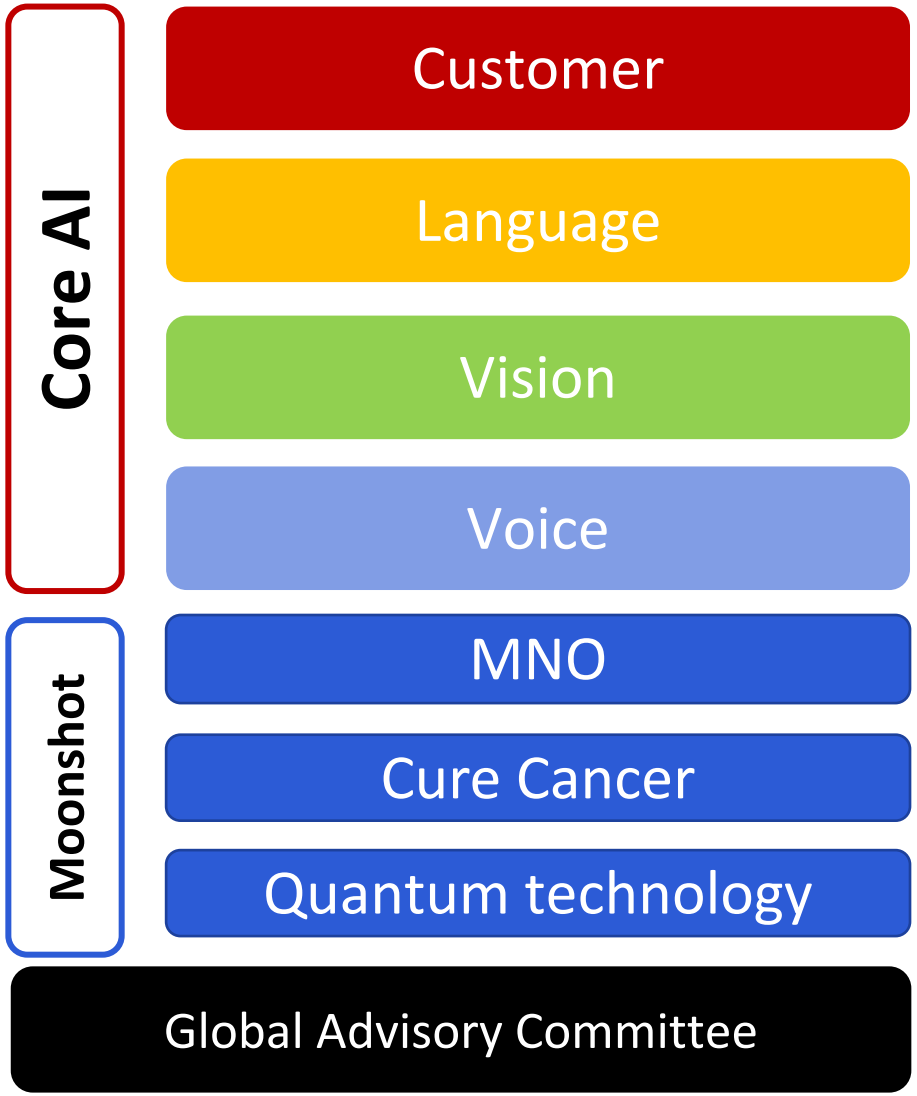
Rakuten

Rakuten Institute of Technology (RIT)
楽天技術研究所

Rakuten Institute of Technology (RIT)

- R&D division of Rakuten Group, established in Dec. 2015.
- Currently, 100+ researchers are working in 6 locations globally.

Research Areas



Deep customer understanding for Rakuten membership

Machine Translation

Fully automated product catalog for Rakuten services

Frictionless image-based experience design

Innovate customer-smart phone interface with voice

Autonomous network with software defined network

Design of novel 5G experiences

AI diagnosis with higher accuracy and scale

Realistic use case of quantum computer and network

Raise the bar to match worlds' top tier research institute

NLP

	Conference	Paper Title
2020 (7)	CHI	EMI: An Expressive Mobile Interactive Robot
	ACL	Lexically Constrained Neural Machine Translation with Levenshtein Transformer
	CIKM	Learning to Profile: User Meta-Profile Network for Few-Shot Learning
	EMNLP	Can Automatic Post-Editing Improve NMT?
	NeurIPS	Steady-State Analysis of Episodic Reinforcement Learning
		A General Large Neighborhood Search Framework for Solving Integer Linear Programs
	AAAI	Simpson's Bias in NLP Training
2021 (8)	ACL	Simple and Effective Query Expansion for QA-Based Product Attribute Extraction
	CVPR	MRAN: Multi-Resolution Attention Network for Facial Action Unit Recognition
	SIGIR	Neural Representations in Hybrid Recommender Systems: Prediction versus Regularization
	RecSys	Shared Neural Item Representations for Completely Cold Start Problem
		Towards Source-Aligned Variational Models for Cross-Domain Recommendation
	EMBC	Dual Encoder Attention U-net for nuclei segmentation
		Blur-Robust Nuclei Segmentation for Immunofluorescence Images
		Spatial Context-aware RNA-data prediction from microscopy H&E images

We are Hiring!

https://rit.rakuten.com/careers/

Vast and exciting data

- The Rakuten Group offers a full range of services related to daily life, from **shopping** and **online payment** to **telecommunications**, **travel**, and **finance**. The vast amount of diverse data we are able to generate and access makes research at RIT dynamic and exciting.

Real business-related research

- At RIT, your research will be put to **real use** and exposed to millions of people around the world every day.

One global team

- Rakuten has **six global offices**, but no matter where we are in the world, we share the same challenges and successes as one team.

Choose the best field for your personal growth

- We provide an environment to **learn**, **grow** and **engage** with excellent researchers around the world. Receive guidance from renowned scientists and AI researchers and start exciting conversations with your peers.

We are Hiring!

RIT Tokyo Language Program

- (Internship) Research Scientist
- (Entry Level) Research Scientist
- Senior Research Scientist

Position Details

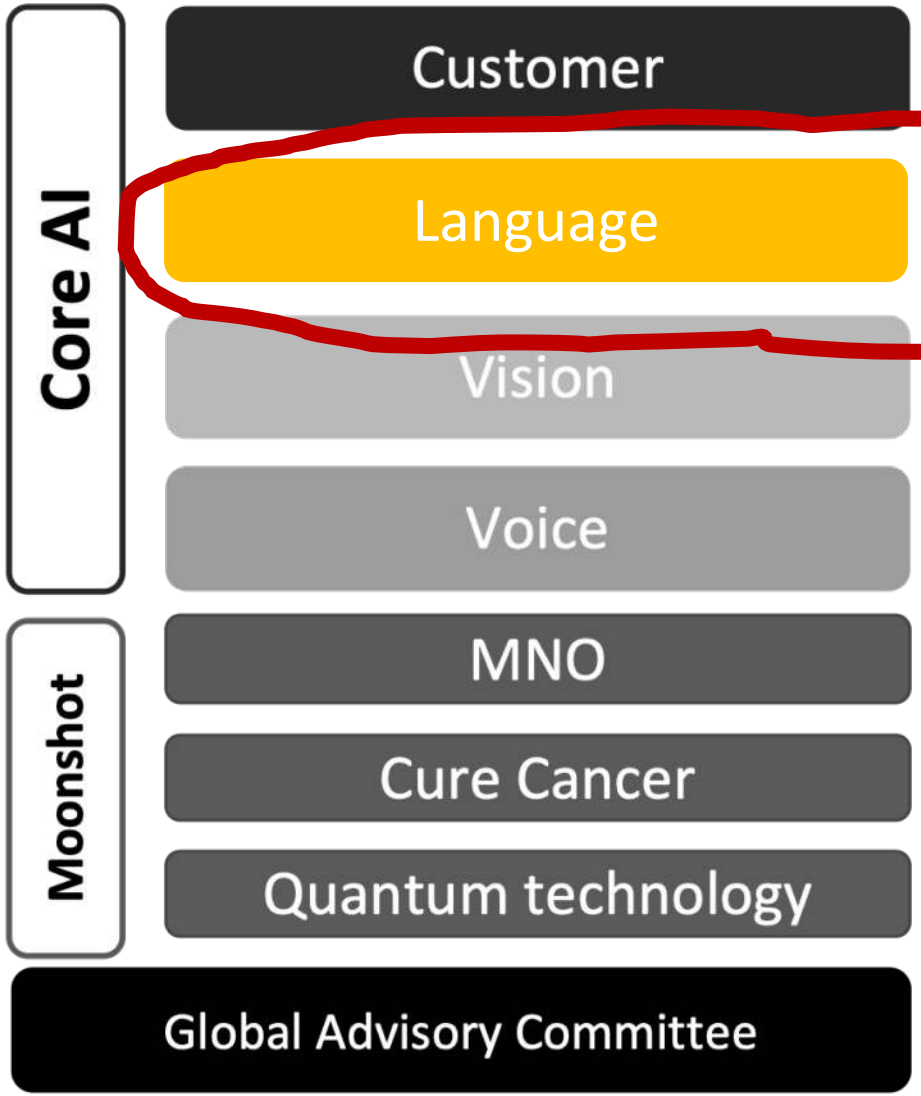
- Design algorithms and build proto-type systems to solve defined scientific problems
- Demonstrate algorithms / proto-type systems to business stakeholders
- Collaborate with engineers to deploy algorithms / proto-type systems into production services
- Collaborate with other scientists on research and paper publications
- Participate in scientific conferences and contribute to the scientific community with paper publications

Mandatory Qualifications

- Bachelor's degree in computer science, related research field or equivalent experience
- English fluency for communicating with researchers, engineers, and business stakeholders
- Proficiency in reading and processing Japanese text (our product catalog is written in Japanese)
- Coding skills: fluency in Python
- Experience with Linux environment

RIT’s Language Program

Research Areas



Deep customer understanding for Rakuten membership

Machine Translation

Fully automated product catalog for Rakuten services

Frictionless image-based experience design

Innovate customer-smart phone interface with voice

Autonomous network with software define network

Design of novel 5G experiences

AI diagnosis with higher accuracy and scale

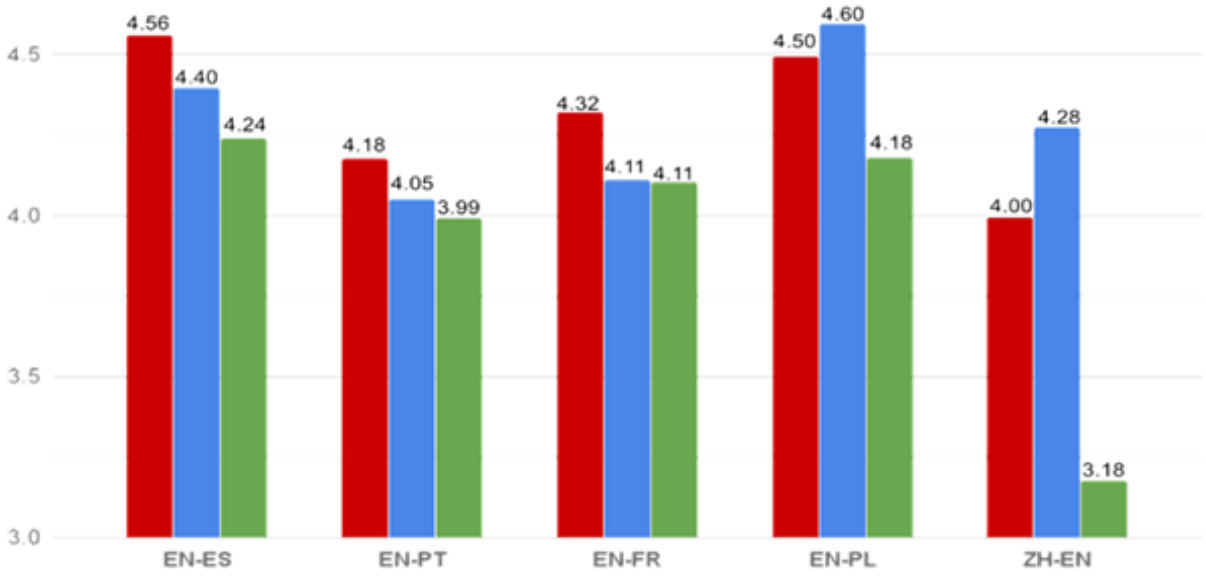
Realistic use case of quantum computer and network

Raise the bar to match worlds' top tier research institute

Machine Translation

Human-level machine translation for TV dramas

- Learned from translated subtitles from Rakuten Viki



RIT Translate

Human

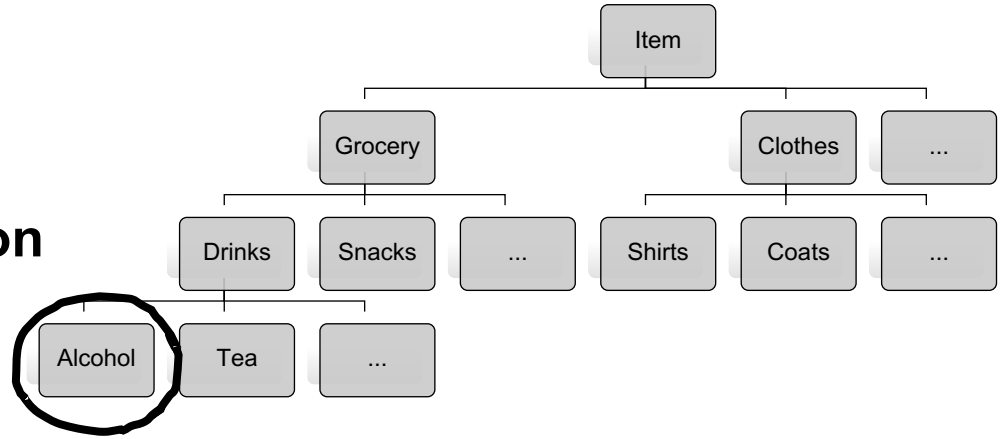
Other leading tech provider

Rated by bi-lingual speakers on a 5-point scale for adequacy and fluency

Currently developing MT in 35 languages

Automated Catalog Organization

Item Classification



Attribute Value Extraction

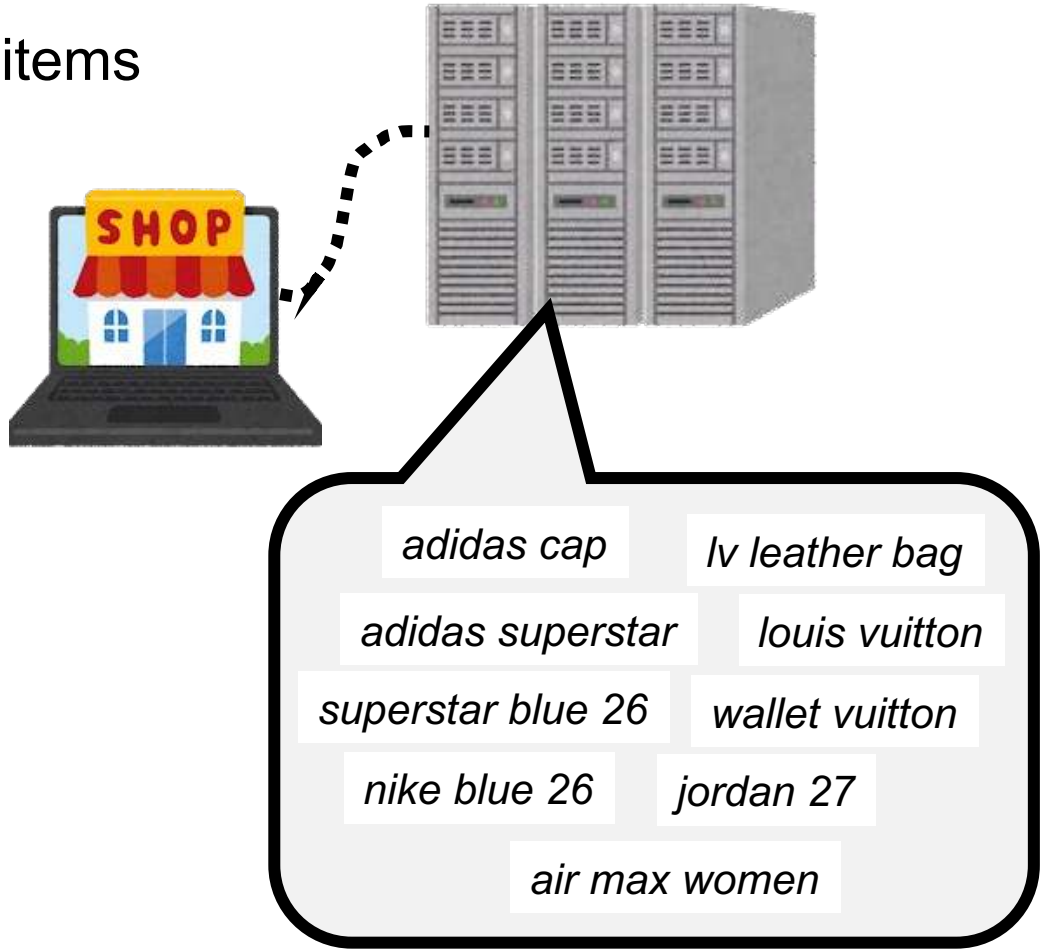
Attribute	Value
Wine Name	Bourgogne Chardonnay
Wine Type	White Wine
Taste	Spicy
Producer	Bouchard Pere & Fils
Origin	France　Bourgogne
Capacity	375 ml / Half Bottle

[【楽天市場】ブルゴーニュ　Bourgogne > いつでも安心、信頼のブルゴーニュ > ブシャール2　Bouchard P&F：ヒグチワイン Higuchi Wine]
Accessed on Wed. March 23 2022

User Demand Understanding

Shopping Query Intent Prediction

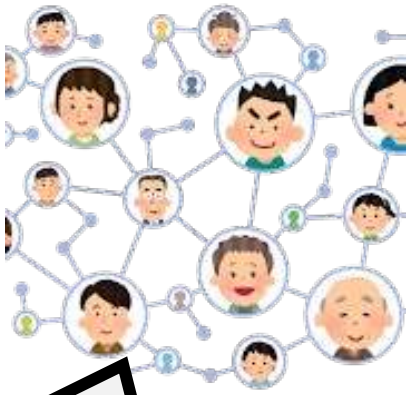
- Mapping a query to attribute values of items



eXtreme Multi-label Classification (XMC) Introduction

Example

Hashtag Prediction



#cocktailparty #datingevent #vacation #children #school #party #sunset #boysandgirls #junglegym #park #summer #beach #palmtree #matchmaking #festival

Can be millions

#children
#junglegym
#park
#boysandgirls

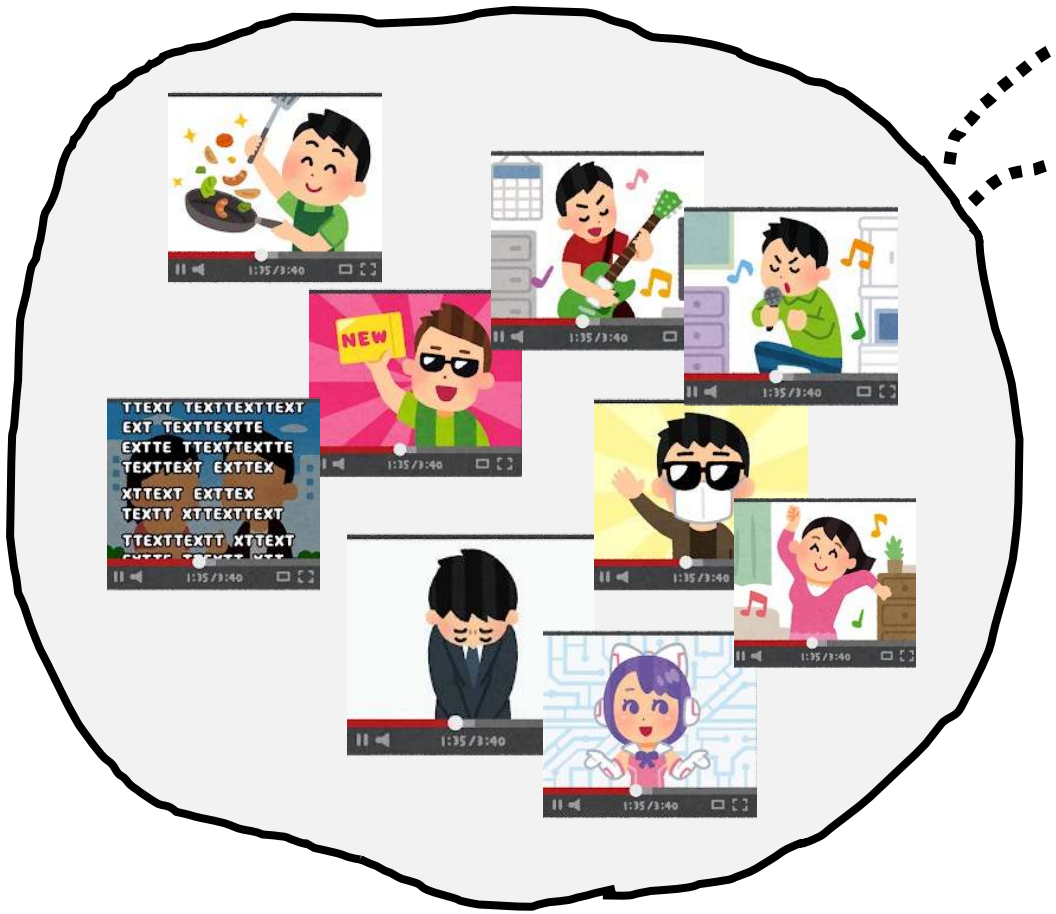
#sunset
#beach
#vacation
#palmtree

#party
#cocktailparty
#matchmaking
#datingevent

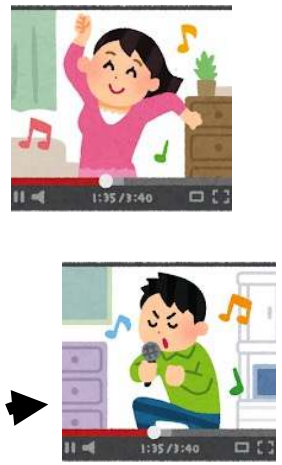
Relevant hashtags

Example (cont.)

Recommendation



Can be millions



Relevant videos

Example (cont.)

Shopping Query Intent Prediction

Can be millions

Task

Input: x

Labels:
y1
y2
y3
...

Can be millions

(x, y1) → ✓
(x, y2) → ✓
(x, y3) → ✗
...

Brand: adidas ✓
Brand: Nike ✗
...
Color: Blue ✓
Color: Red ✗
...
Size: 26.0cm ✓

"adidas blue 26"

Brand: adidas
Brand: Nike
Color: Red
Size: 26.0cm
...

(*"adidas blue 26"*, Brand: adidas) → ✓
(*"adidas blue 26"*, Brand: Nike) → ✗
(*"adidas blue 26"*, Color: Red) → ✗
(*"adidas blue 26"*, Size: 26.0cm) → ✓
...

Task (cont.)

Input	Feature 1	Feature 2	Feature 3	...	Label 1	Label 2	Label 3	Label 4	...
“nike white 26”	0.1	0.9	0.2	...	0	1	0	1	...
“adidas red”	0.3	0.4	0.7	...	1	0	1	0	...
“nike air jordan”	0.1	0.2	0.8	...	0	1	0	0	...

Brand: adidas
Brand: Nike
Color: Red
Size: 26cm

“adidas blue 26”

Brand: adidas
Brand: Nike
Color: Red
Size: 26.0cm
...

(*“adidas blue 26”*, Brand: adidas) → ✓
(*“adidas blue 26”*, Brand: Nike) → ✗
(*“adidas blue 26”*, Color: Red) → ✗
(*“adidas blue 26”*, Size: 26.0cm) → ✓
...

Conferences on Extreme Classification

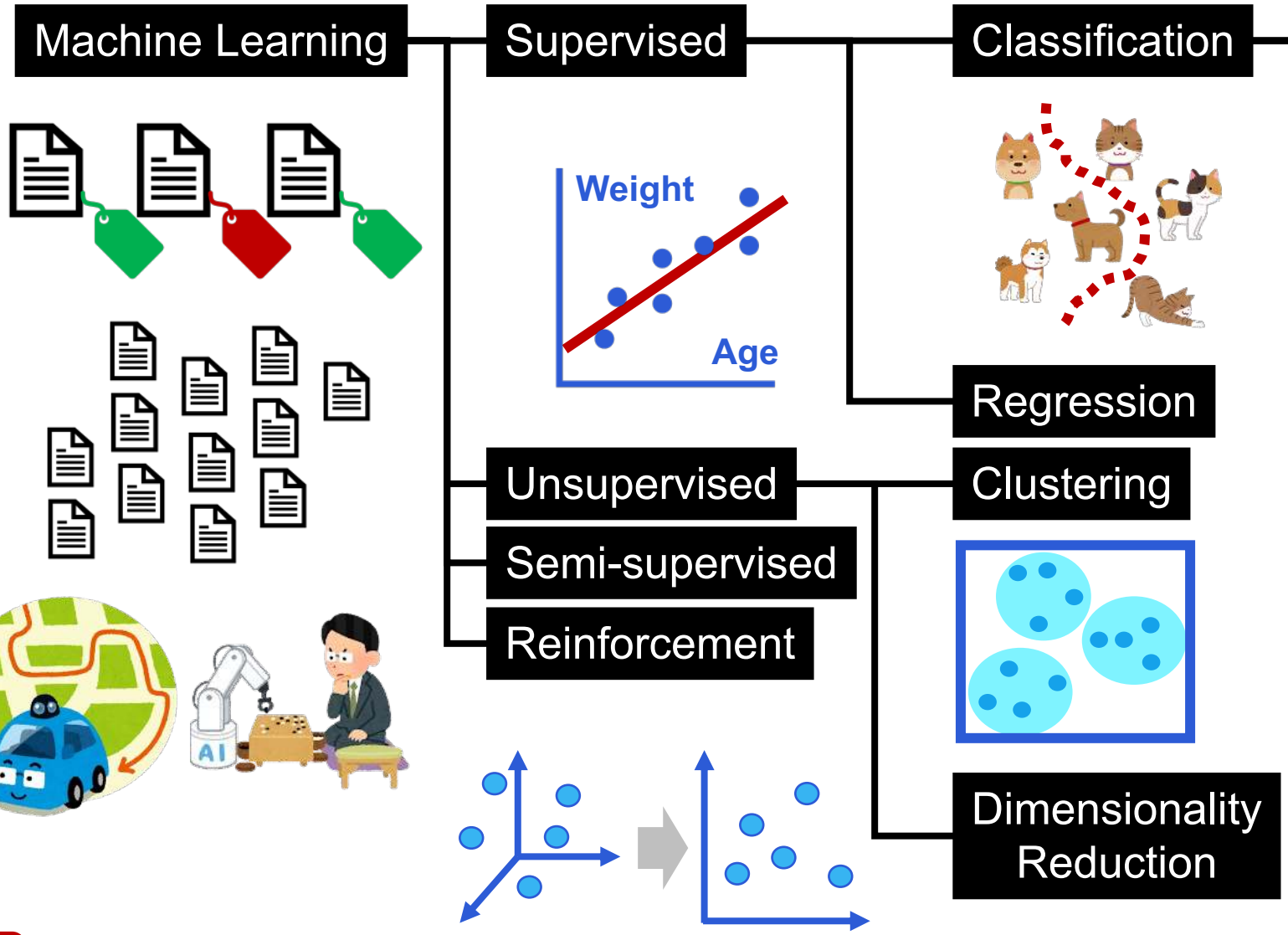
- Extreme Classification 2020, The ICML Workshop on Extreme Classification: Theory and Applications
- The CVPR 2020 Extreme Classification in Computer Vision Workshop
- The 2018 Dagstuhl Seminar on Extreme Classification
- The WWW 2018 Workshop on Extreme Multilabel Classification for Social Media
- The NIPS 2017 Extreme Classification Workshop
- The NIPS 2016 Extreme Classification Workshop
- The NIPS 2015 Extreme Classification Workshop
- The ICML 2015 Extreme Classification Workshop
- The NIPS 2013 Extreme Classification Workshop

Extreme Classification Repository

http://manikvarma.org/downloads/XC/XMLRepository.html

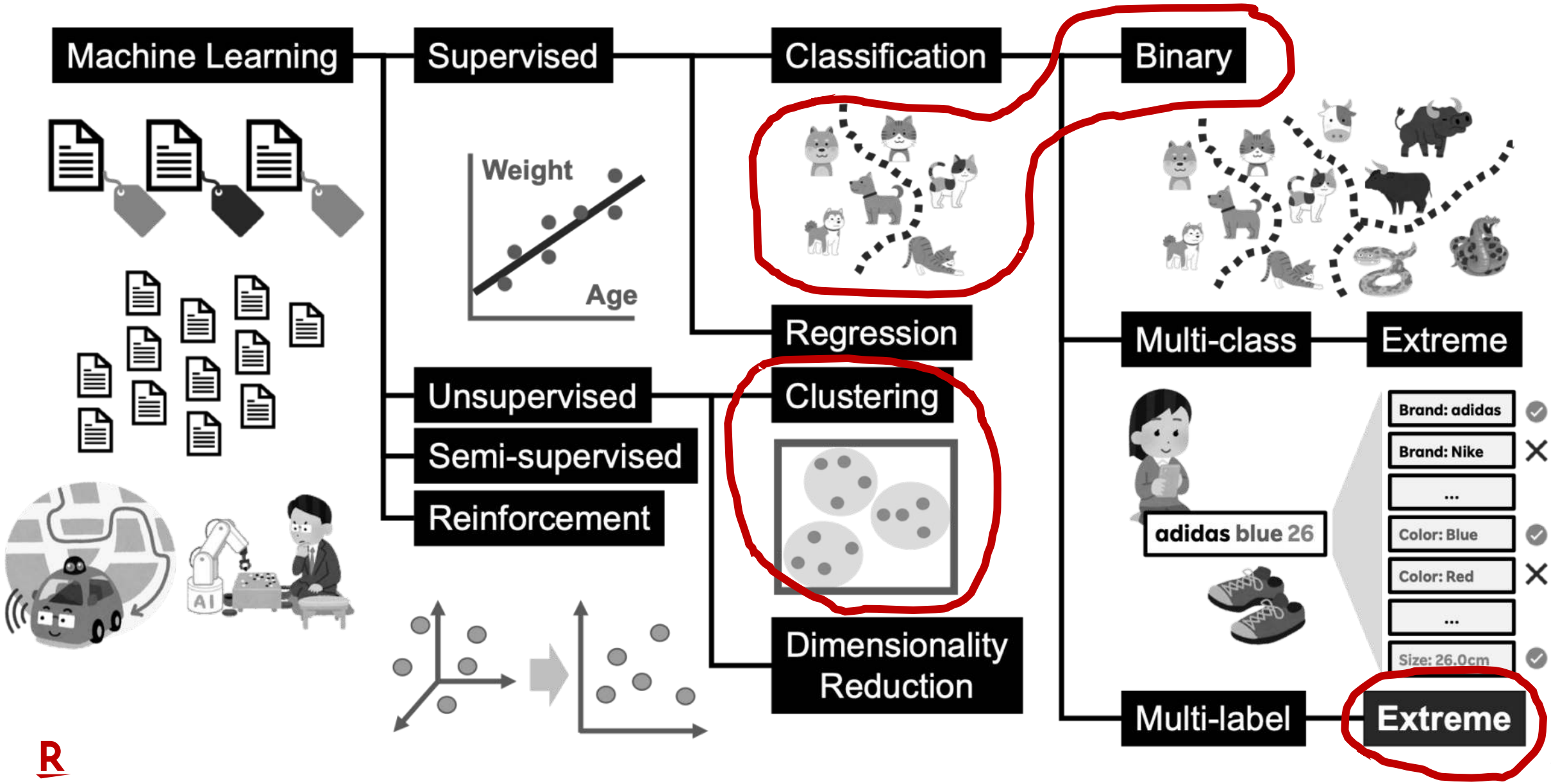
- Datasets
- Code
- Evaluation metrics
- Benchmarked results
- References

XMC in Machine Learning Typology



eXtreme Multi-label Classification (XMC) Preliminary

Underlying Technologies for XMC



Binary Classification

adidas or not

adidas blue 26

Brand: adidas

Something else

adidas superstar

superstar blue 26

nike blue 26

louis vuitton

Training Data

Query	x_1 *adidas*	x_2 *superstar*	x_3 *blue*	...	y adidas?
adidas superstar	1	1	0	...	1
superstar blue 26	0	1	1	...	1
nike blue 26	0	0	1	...	0
louis vuitton	0	0	0	...	0

Labels

Training

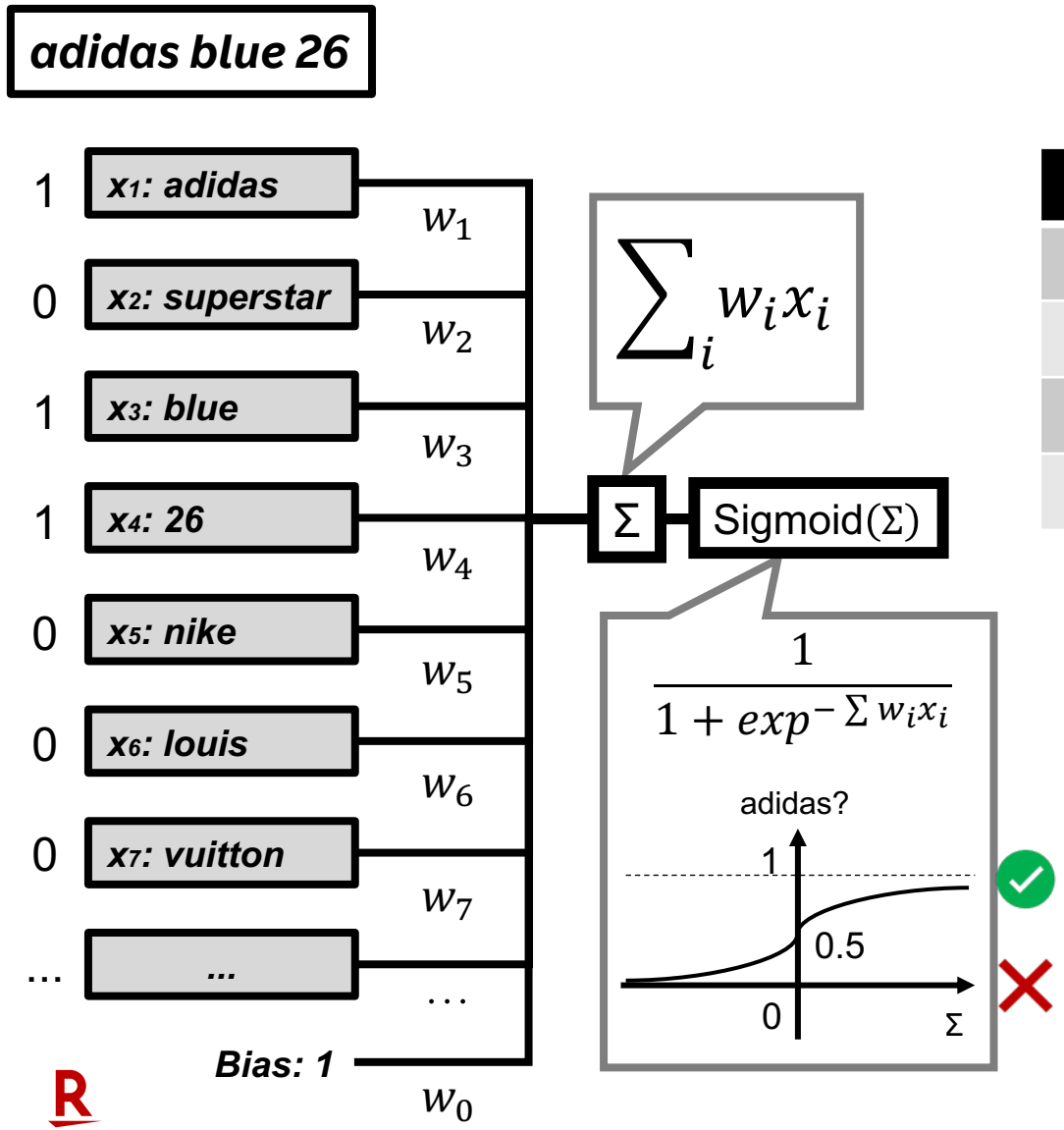
Prediction

adidas blue 26

vuitton blue

adidas-or-not model

Binary Classification: Logistic Regression



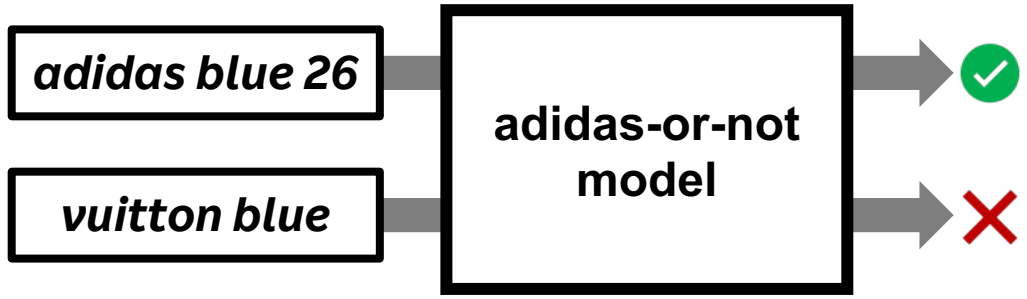
Training Data

	x1	x2	x3	...	y
Query	***adidas***	***superstar***	***blue***	***...***	**adidas?**
adidas superstar	1	1	0	...	1
superstar blue 26	0	1	1	...	1
nike blue 26	0	0	1	...	0
louis vuitton	0	0	0	...	0

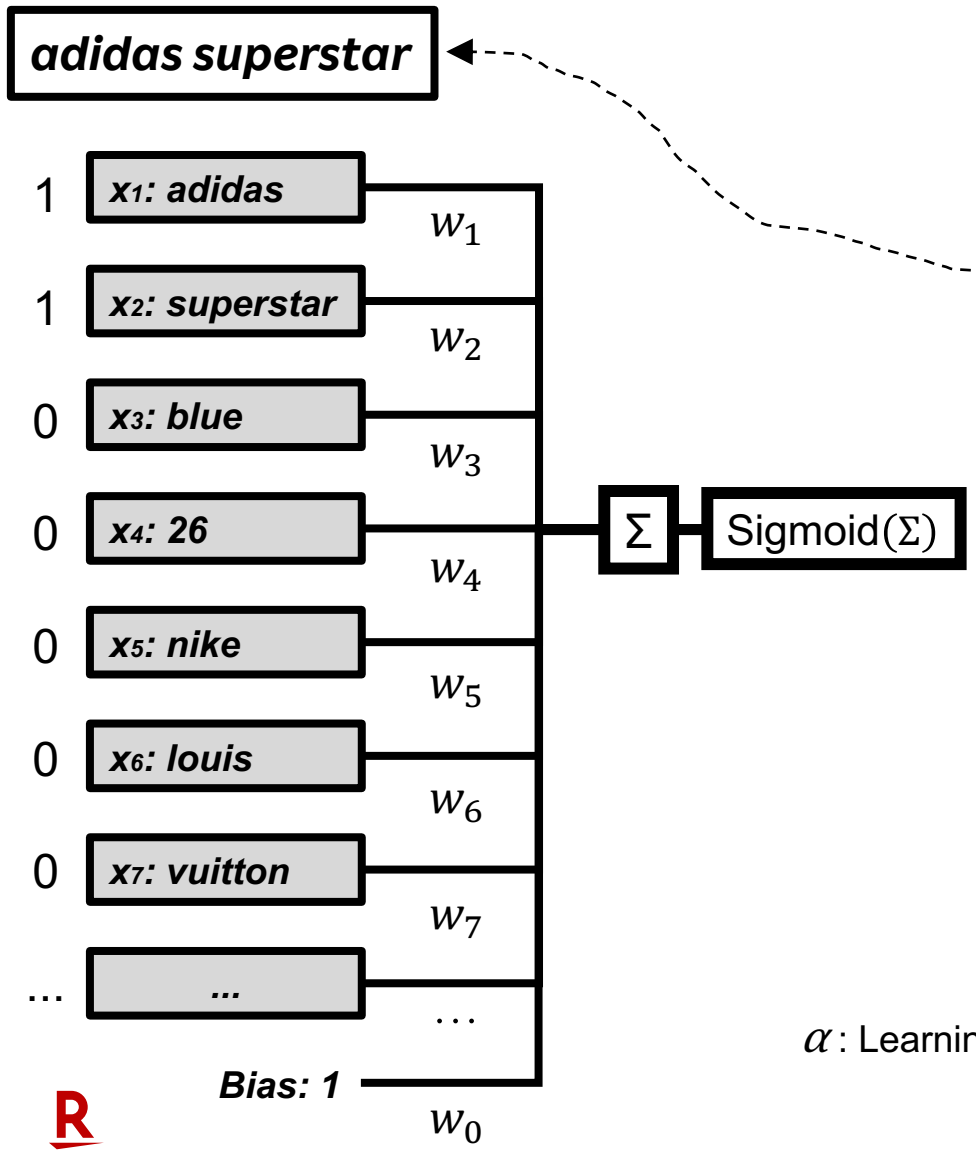
Training

$w_1, w_2, w_3, \ldots, w_7, w_0$

Prediction



Binary Classification: Training with Stochastic Gradient Decent



Training Data

	x1	x2	x3	...	y
Query	adidas	superstar	blue	...	adidas?
adidas superstar	1	1	0	...	1
superstar blue 26	0	1	1	...	1
nike blue 26	0	0	1	...	0
louis vuitton	0	0	0	...	0

① Initialize $w_1, w_2, w_3, \dots, w_7, w_0$ with random values.

② Take a random data point from training data.

③ Compute Sigmoid(Σ).

④ Update each w_i with:

$$w_i := w_i + \alpha\big(y - Sigmoid(\Sigma)\big)x_i$$

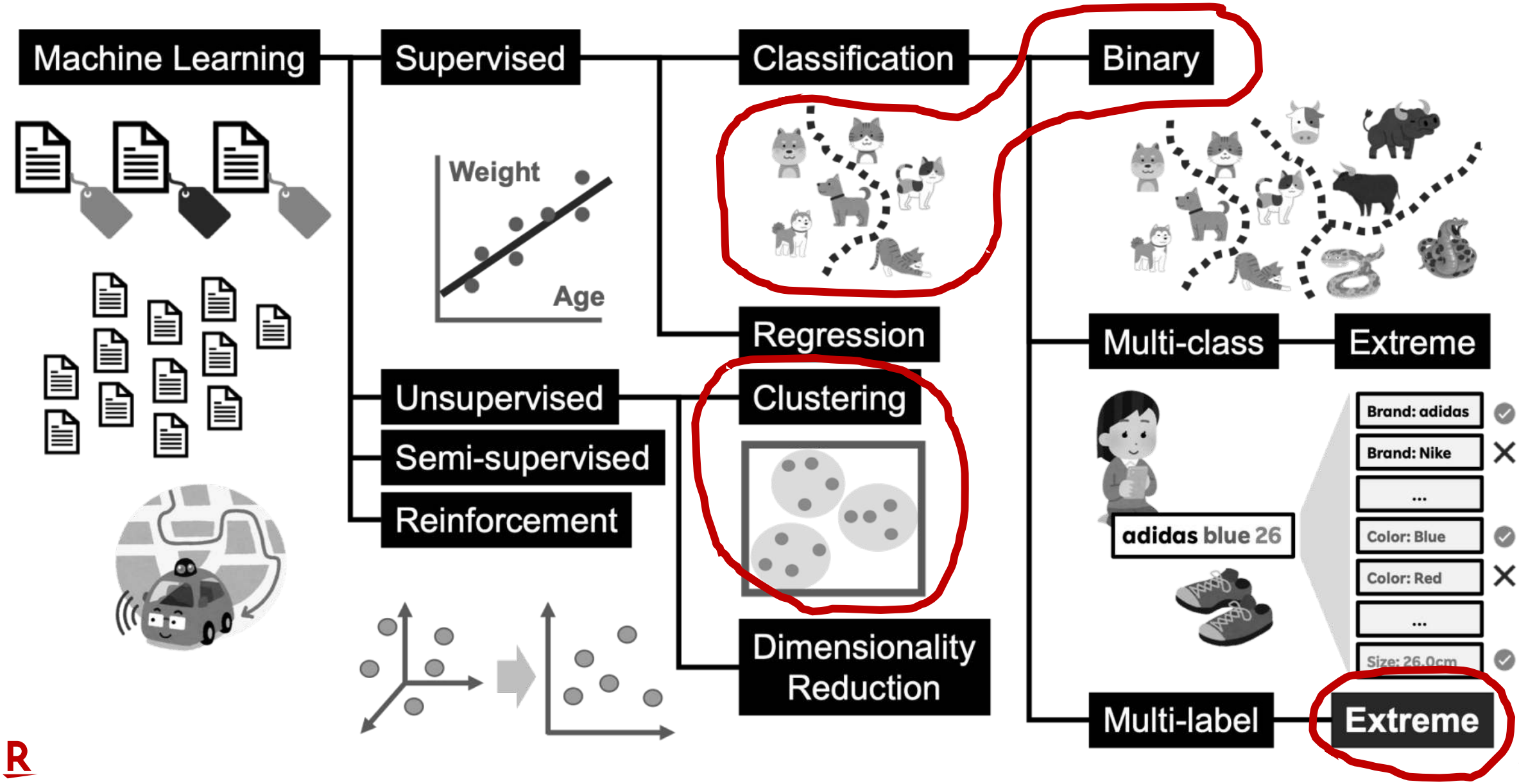
x : Feature value

y : True value

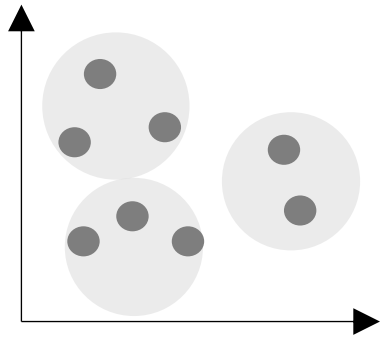
α : Learning rate (e.g. 0.1)

youtube logistic regression stanford cs229

Underlying Technologies for XMC



Clustering



Data points in
a vector space

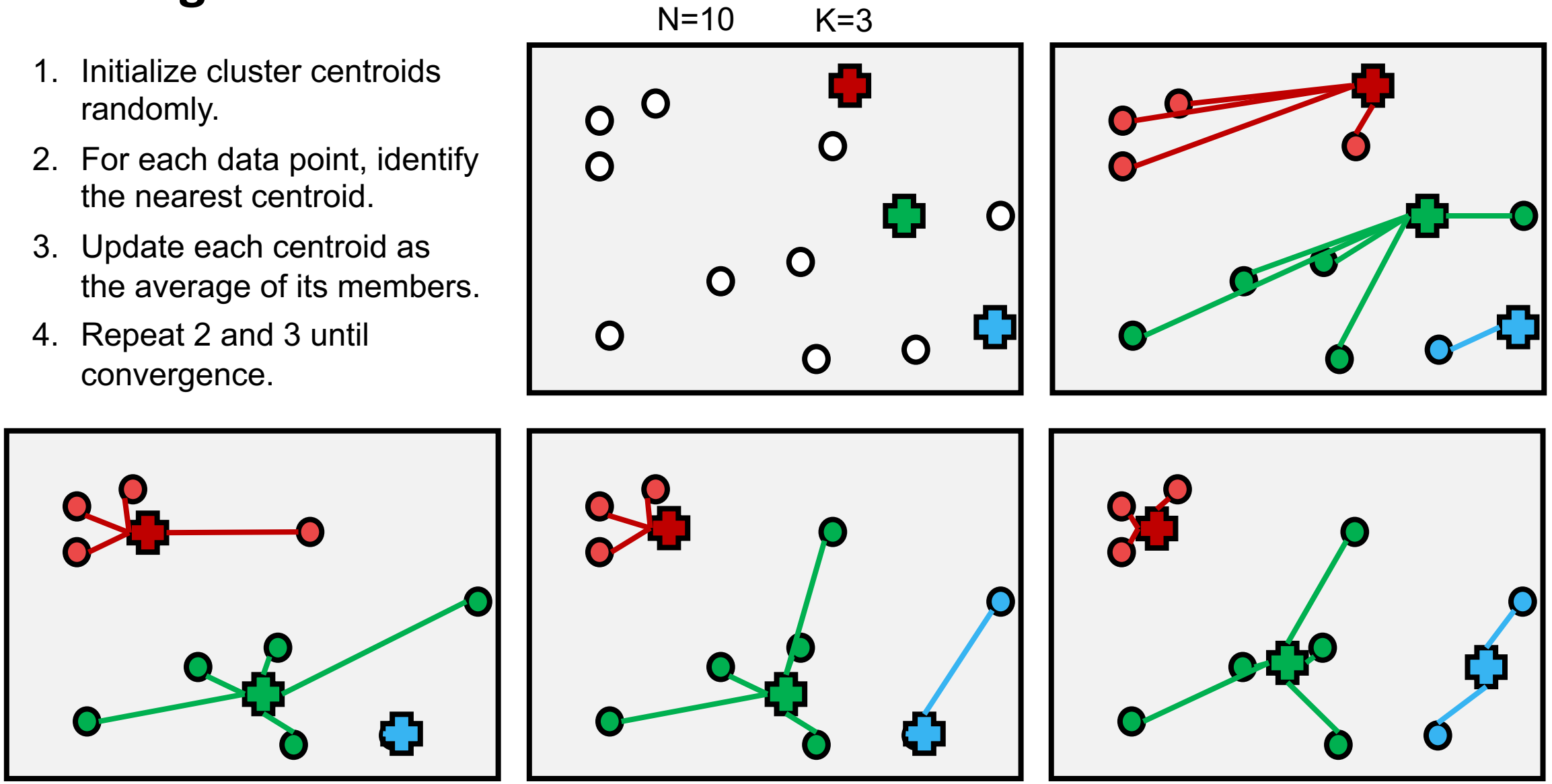
Query
adidas superstar
superstar blue 26
nike blue 26
louis vuitton

x_1	x_2	x_3	...
adidas	***superstar***	***blue***	***...***
1	**1**	0	...
0	**1**	**1**	...
0	0	**1**	...
0	0	0	...

Vector

Clustering: K-Means

1. Initialize cluster centroids randomly.
2. For each data point, identify the nearest centroid.
3. Update each centroid as the average of its members.
4. Repeat 2 and 3 until convergence.



http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Clustering Labels

Clustering helps.

Can be millions.

Instance	Feature 1	Feature 2	Feature 3	...	Label 1	Label 2	Label 3	Label 4	...
"nike white 26"	0.1	0.9	0.2	...	0	1	0	1	...
"adidas red"	0.3	0.4	0.7	...	1	0	1	0	...
"nike air jordan"	0.1	0.2	0.8	...	0	1	0	0	...

Brand: adidas

Brand: Nike

Color: Red

Size: 26cm

"nike white 26"

Brand: adidas
Brand: Nike
Color: Red
Size: 26cm
...

(*"nike white 26"*, Brand: adidas) → ✗
(*"nike white 26"*, Brand: Nike) → ✓
(*"nike white 26"*, Color: Red) → ✗
(*"nike white 26"*, Size: 26cm) → ✓
...

Clustering Labels (cont.)

Features for label clustering: Positive Instance Feature Aggregation (PIFA)

Query
“nike white 26”
“adidas red”
“nike air jordan”

Feature 1	Feature 2	Feature 3	...
0.1	0.9	0.2	...
0.3	0.4	0.7	...
0.1	0.2	0.8	...

Label 1	Label 2	Label 3	Label 4	...
0	1	0	1	...
1	0	1	0	...
0	1	0	0	...

Brand: adidas

Brand: Nike

Color: Red

Size: 26cm

Features for
Brand: Nike

Feature 1 = (0.1 + 0.1) / 2
Feature 2 = (0.9 + 0.2) / 2
Feature 3 = (0.2 + 0.8) / 2
...

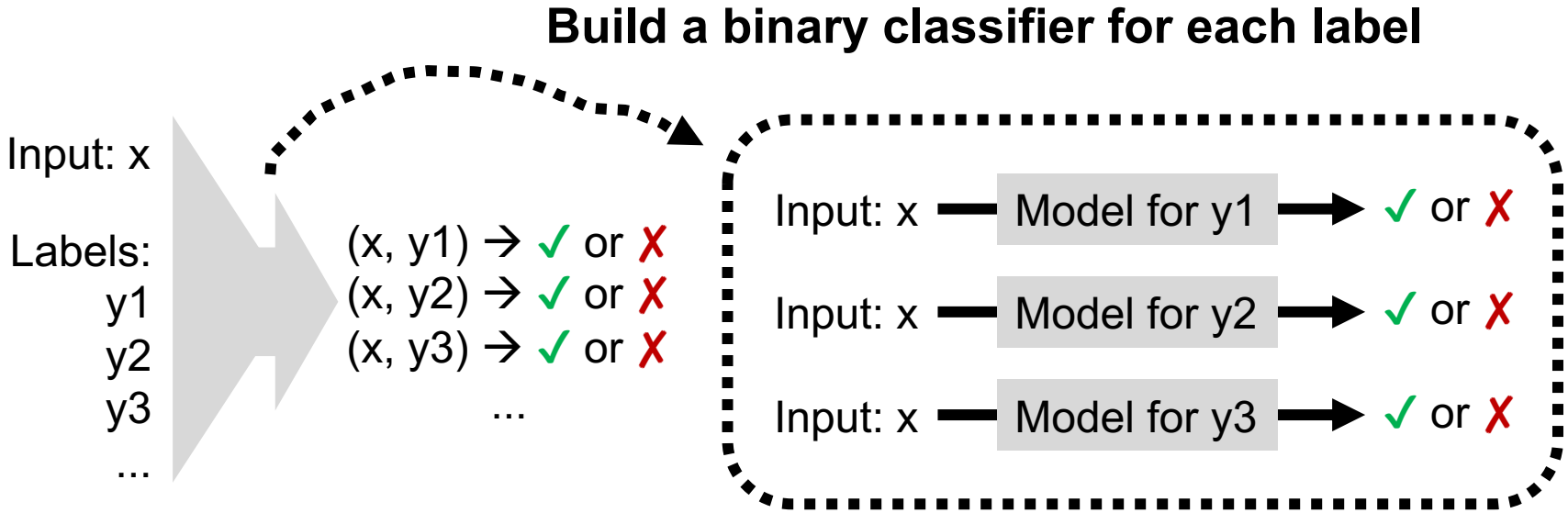
Feature 1	Feature 2	Feature 3	...
0.1	0.55	0.5	...

eXtreme Multi-label Classification (XMC) Simple Solution

Simple Solution

One-Versus-Rest (OVR)

a.k.a One-Versus-All (OVA)
Binary Relevance (BR)



Query	Feature 1	Feature 2	Feature 3	...	Label	Label 2	Label 3	Label 4	...
"nike white 26"	0.1	0.9	0.2	..	0	1	0	1	...
"adidas red"	0.3	0.4	0.7	..	1	0	1	0	...
"nike air jordan"	0.1	0.2	0.8	...	0	1	0	0	...

Brand: adidas
Brand: Nike
Color: Red
Size: 26cm

Model for Brand: adidas

Model for Brand: Nike

Issues for OVR

Requires long time for prediction

Can be millions

Can be millions

Query	Feature 1	Feature 2	Feature 3	..	Label	Label 2	Label 3	Label 4	...
"nike white 26"	0.1	0.9	0.2	..	0	1	0	1	...
"adidas red"	0.3	0.4	0.7	..	1	0	1	0	...
"nike air jordan"	0.1	0.2	0.8	...	0	1	0	0	...

Brand: adidas

Brand: Nike

Color: Red

Size: 26cm

Model for Brand: adidas

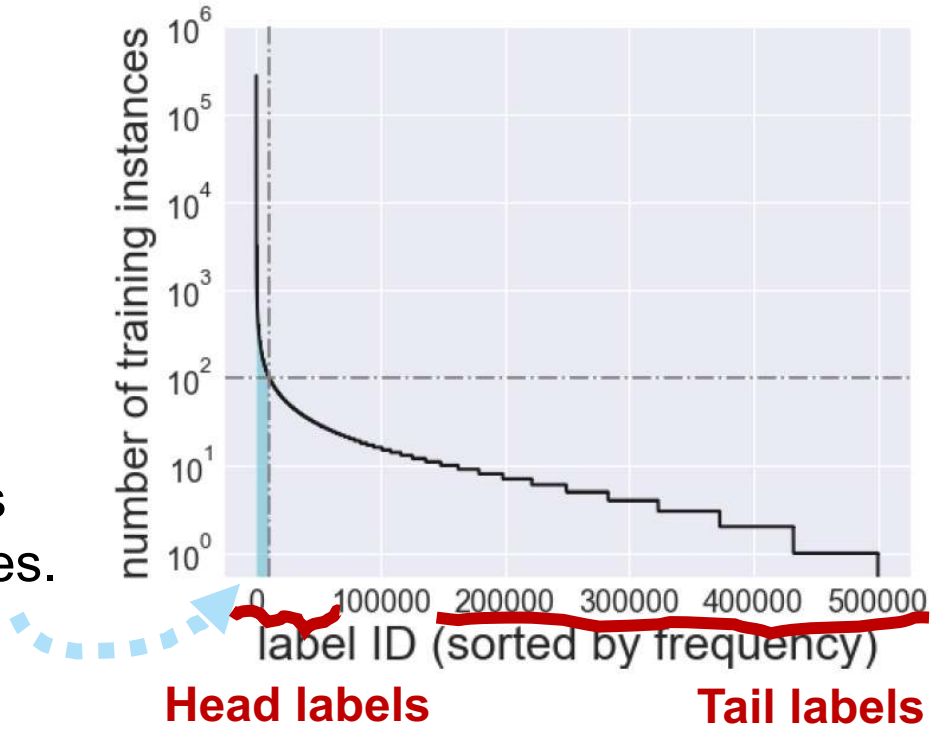
Model for Brand: Nike

Issues for OVR (cont.)

Some labels may have only a few positive training instances

- *Long-tail distribution*

In Wiki-500K, only 2.1% of the labels have more than 100 training instances.



Query	Feature 1	Feature 2	Feature 3	...	Label	Label 2	Label 3	Label 4	...
"nike white 26"	0.1	0.9	0.2	..	0	1	0	1	...
"adidas red"	0.3	0.4	0.7	..	1	0	1	0	...
"nike air jordan"	0.1	0.2	0.8	...	0	1	0	0	...

Brand: adidas

Brand: Nike

Color: Red

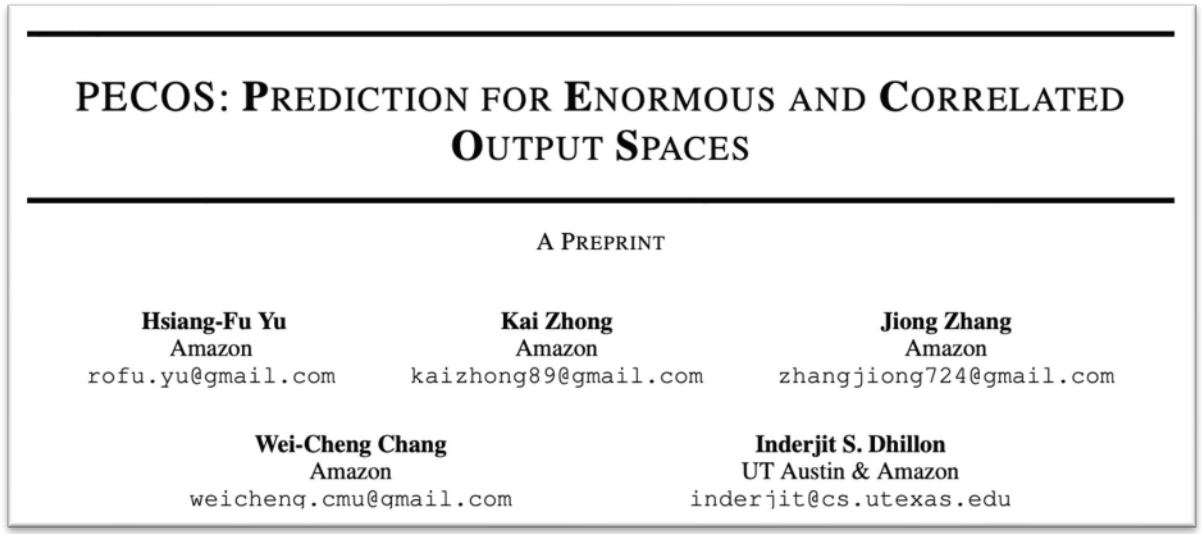
Size: 26cm

Model for Brand: adidas

Model for Brand: Nike

eXtreme Multi-label Classification (XMC) State-of-the-Art

State-of-the-Art Method



PECOS: PREDICTION FOR ENORMOUS AND CORRELATED OUTPUT SPACES

A PREPRINT

Hsiang-Fu Yu
Amazon
rofu.yu@gmail.com

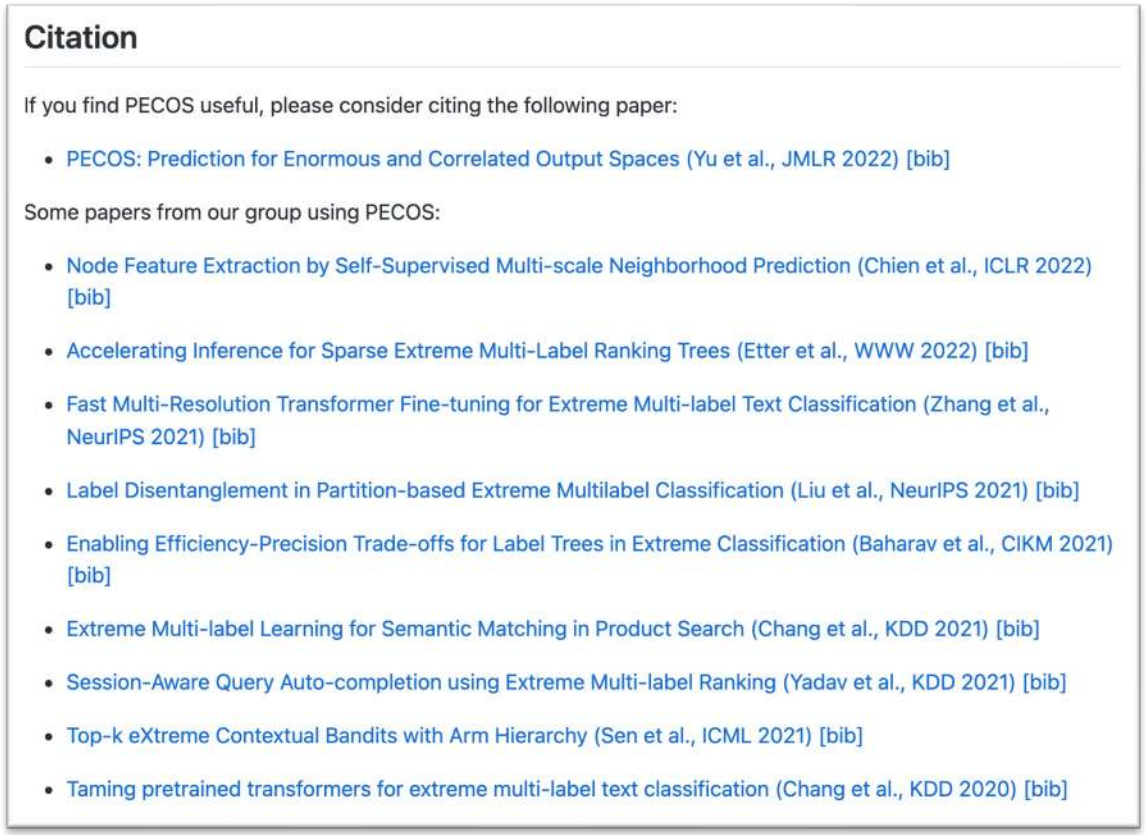
Kai Zhong
Amazon
kaizhong89@gmail.com

Jiong Zhang
Amazon
zhangjiong724@gmail.com

Wei-Cheng Chang
Amazon
weicheng.cmu@gmail.com

Inderjit S. Dhillon
UT Austin & Amazon
inderjit@cs.utexas.edu

https://arxiv.org/pdf/2010.05878.pdf ***JMLR 2022***



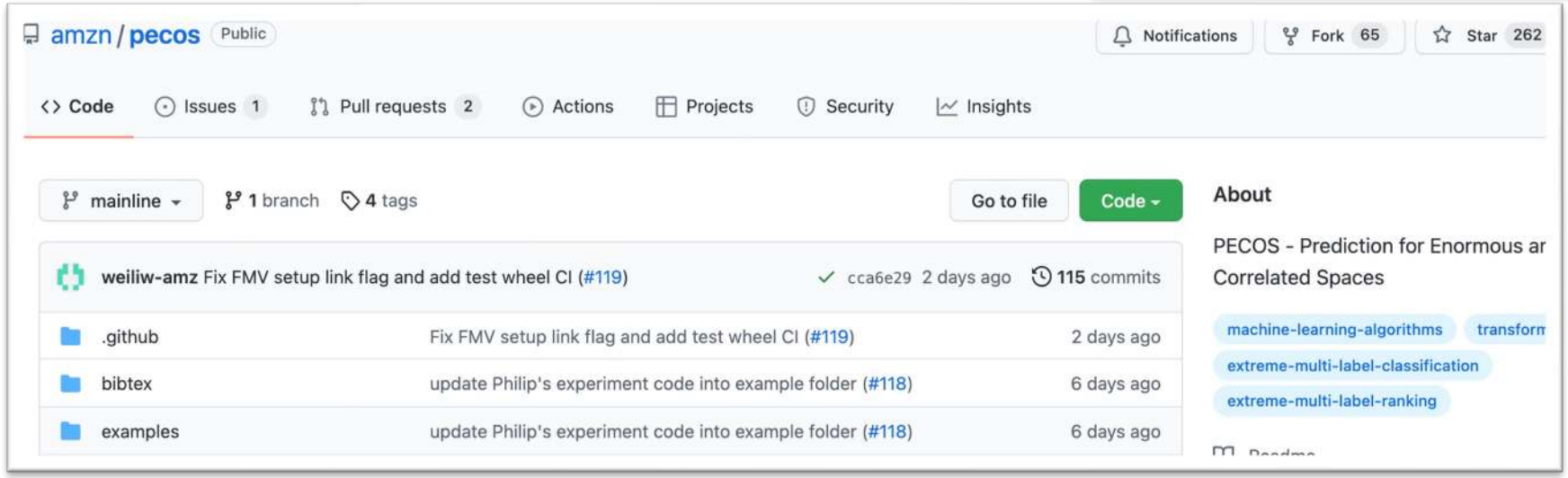
Citation

If you find PECOS useful, please consider citing the following paper:

- PECOS: Prediction for Enormous and Correlated Output Spaces (Yu et al., JMLR 2022) [bib]

Some papers from our group using PECOS:

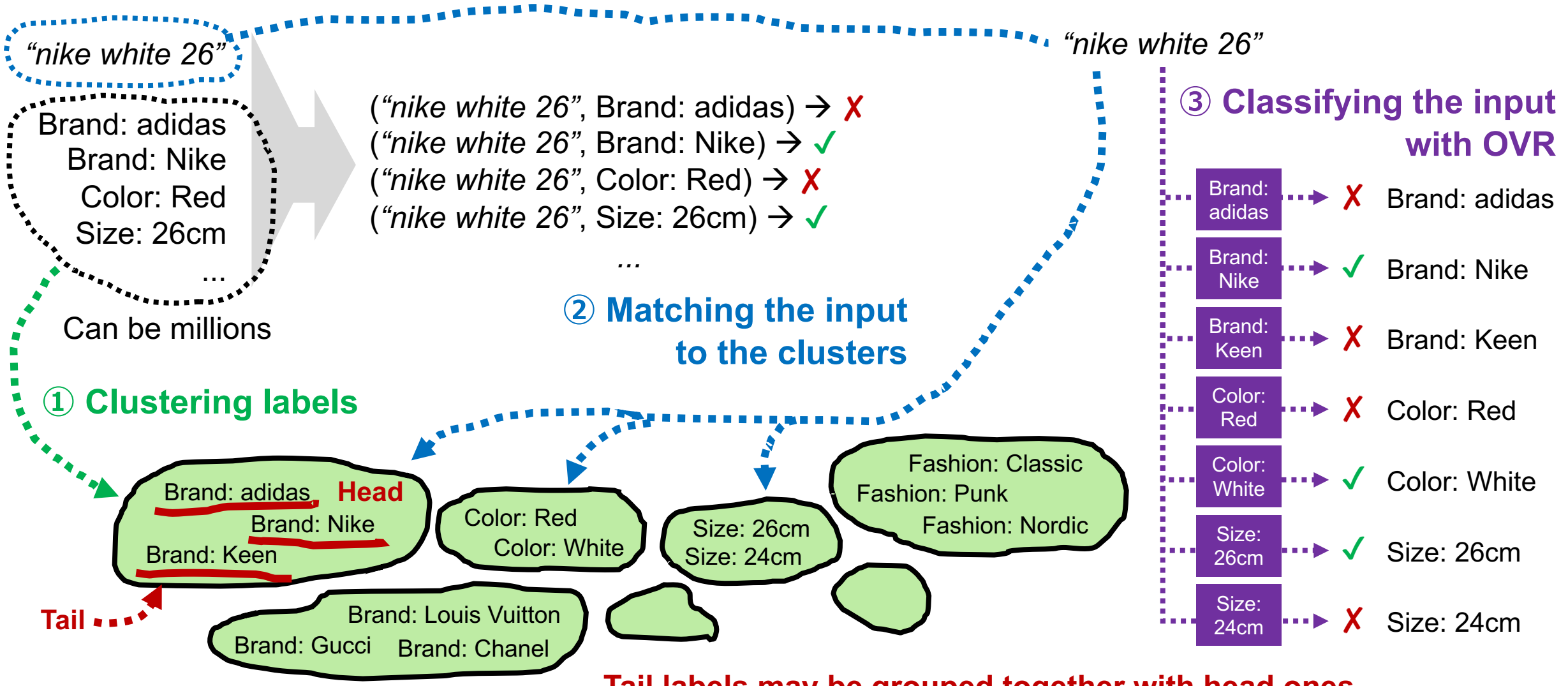
- Node Feature Extraction by Self-Supervised Multi-scale Neighborhood Prediction (Chien et al., ICLR 2022) [bib]
- Accelerating Inference for Sparse Extreme Multi-Label Ranking Trees (Etter et al., WWW 2022) [bib]
- Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification (Zhang et al., NeurIPS 2021) [bib]
- Label Disentanglement in Partition-based Extreme Multilabel Classification (Liu et al., NeurIPS 2021) [bib]
- Enabling Efficiency-Precision Trade-offs for Label Trees in Extreme Classification (Baharav et al., CIKM 2021) [bib]
- Extreme Multi-label Learning for Semantic Matching in Product Search (Chang et al., KDD 2021) [bib]
- Session-Aware Query Auto-completion using Extreme Multi-label Ranking (Yadav et al., KDD 2021) [bib]
- Top-k eXtreme Contextual Bandits with Arm Hierarchy (Sen et al., ICML 2021) [bib]
- Taming pretrained transformers for extreme multi-label text classification (Chang et al., KDD 2020) [bib]



https://github.com/amzn/pecos

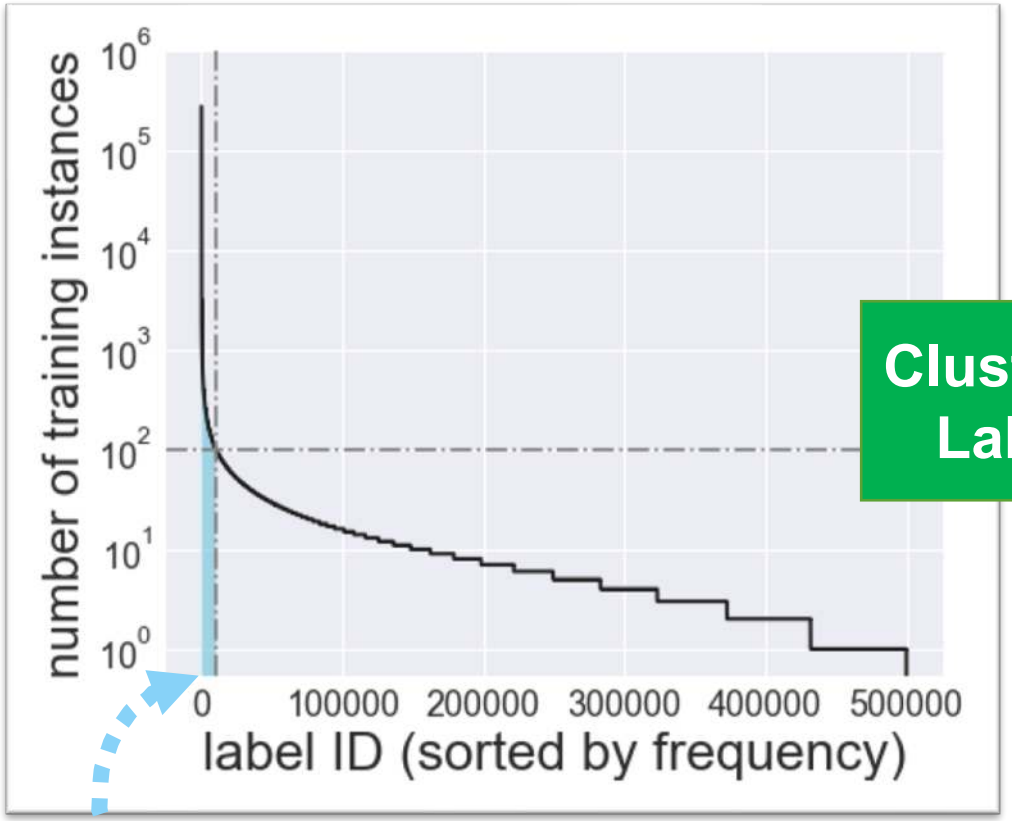
State-of-the-Art Method (cont.)

A small number of labels need to be classified; the simple OVR works.

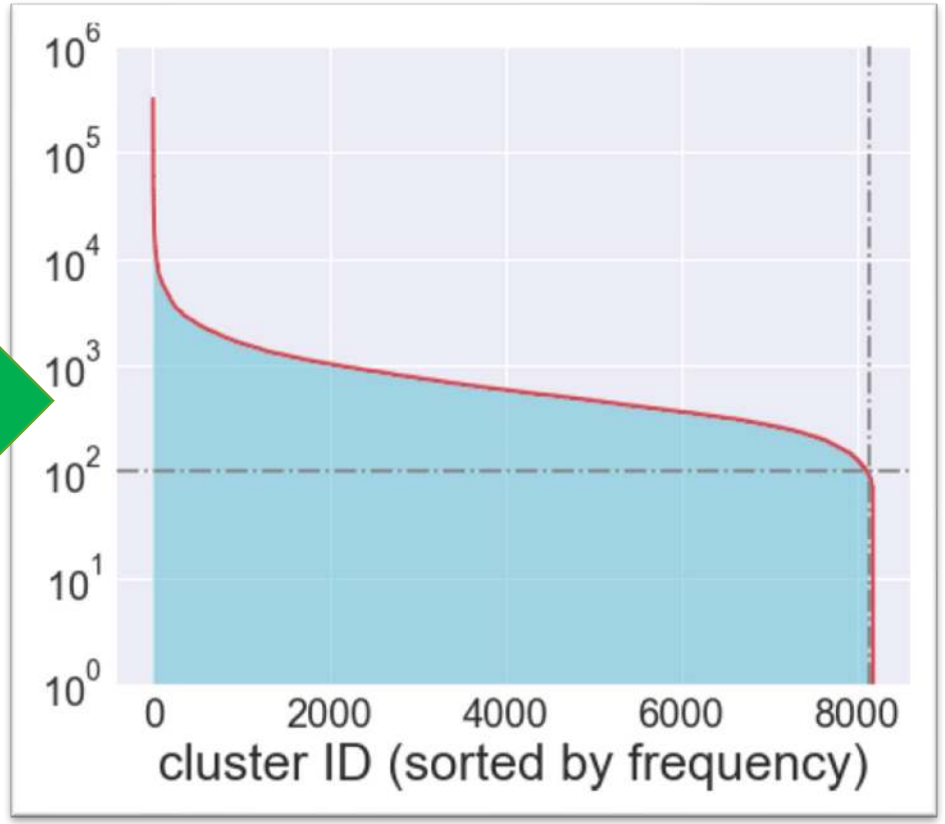


Tail labels may be grouped together with head ones, enabling "transfer learning" from head to tail.

State-of-the-Art Method (cont.)



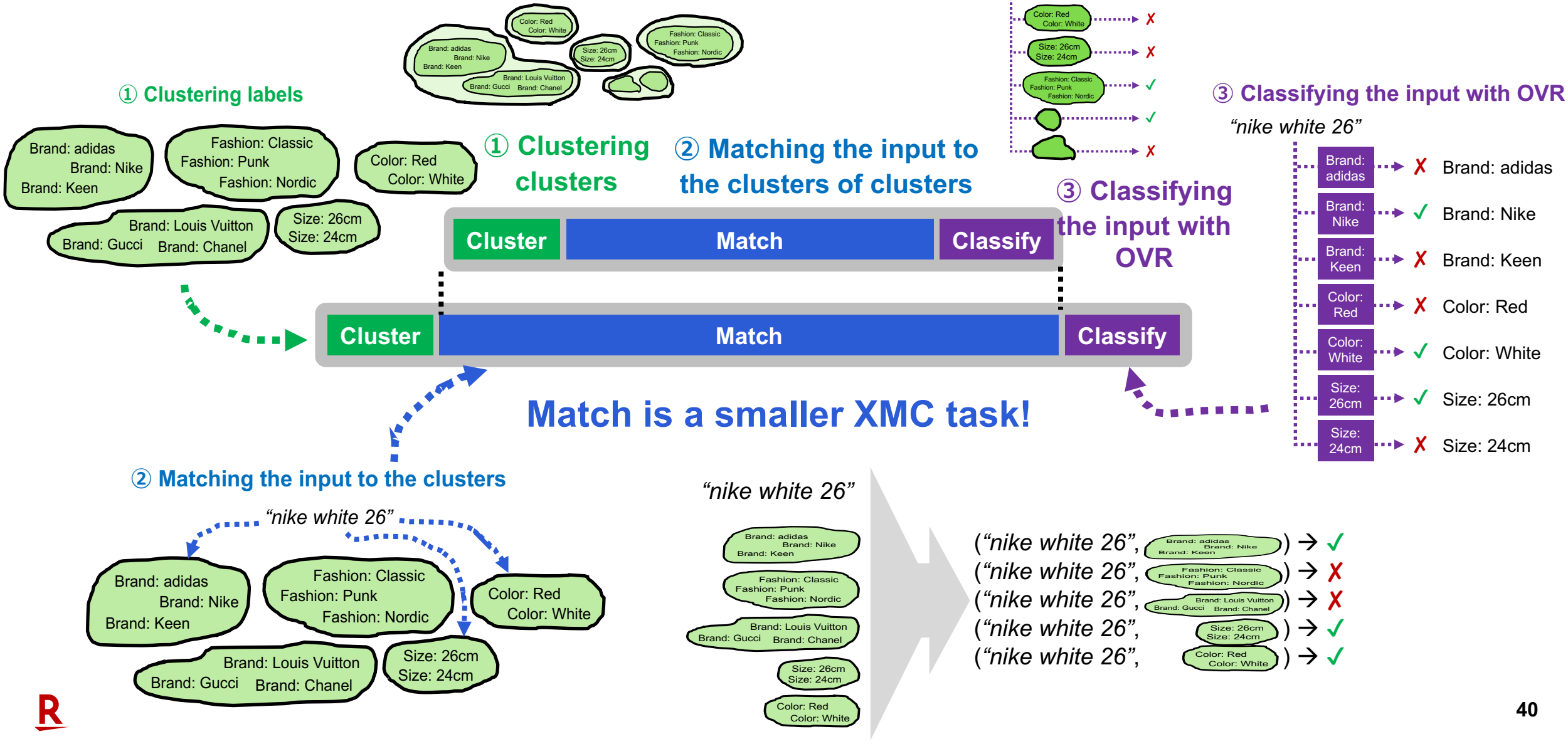
Clustering Labels



In Wiki-500K, only 2.1% of the labels have more than 100 training instances.

Clustering creates 8,192 label clusters; 99.4% of clusters have more than 100 training instances.

State-of-the-Art Method (cont.)



State-of-the-Art Method (cont.)

Suppose: Cluster #1 Cluster #2 Cluster #3

	Query	Feature 1	Feature 2	Feature 3	...	Label 1	Label 2	Label 3	Label 4	...
Clusters #1, #3	*"nike white 26"*	0.1	0.9	0.2	...	0	1	0	1	...
Clusters #1, #2	*"adidas red"*	0.3	0.4	0.7	...	1	0	1	0	...
Cluster #1	*"nike air jordan"*	0.1	0.2	0.8	...	0	1	0	0	...

Used to train a smaller XMC model.

Brand: adidas

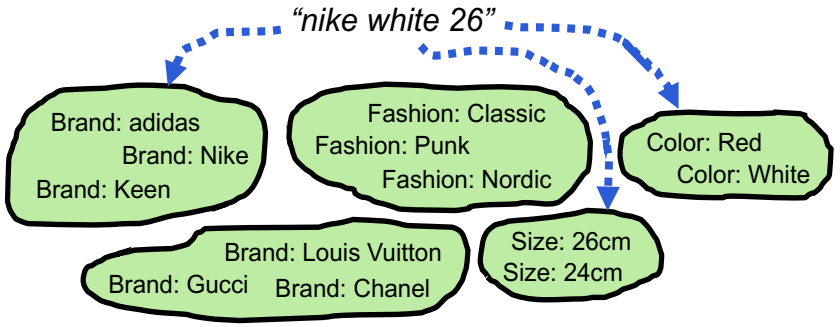
Brand: Nike

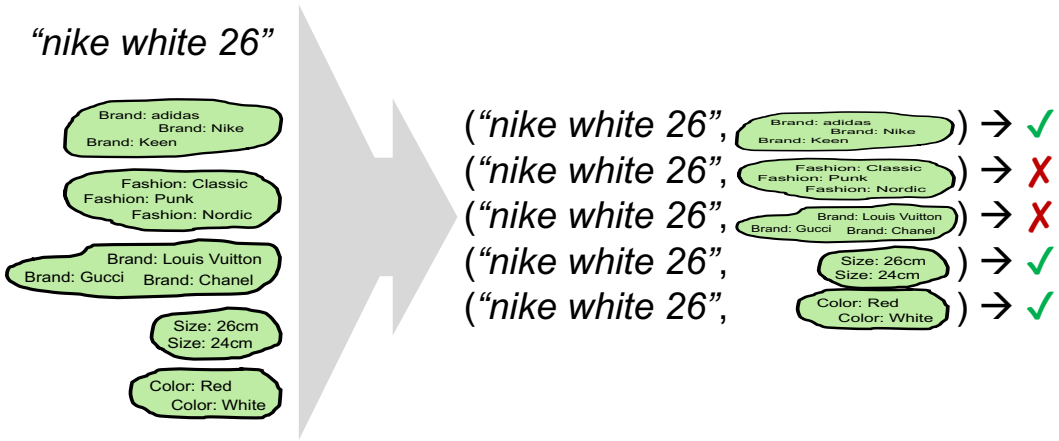
Color: Red

Size: 26cm

Match is a smaller XMC task!

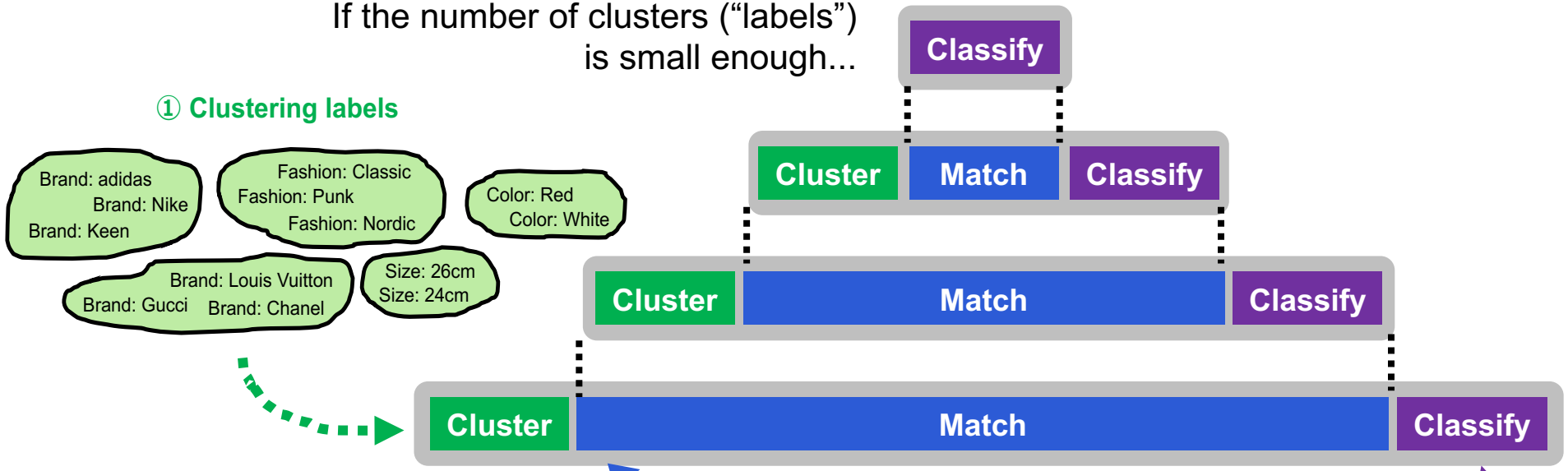
② Matching the input to the clusters





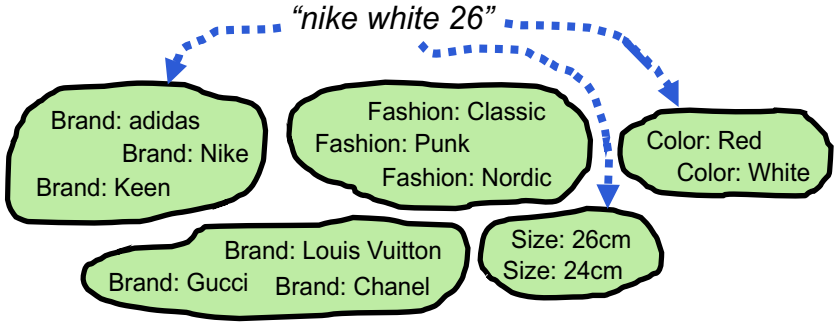
State-of-the-Art Method (cont.)

If the number of clusters (“labels”) is small enough...



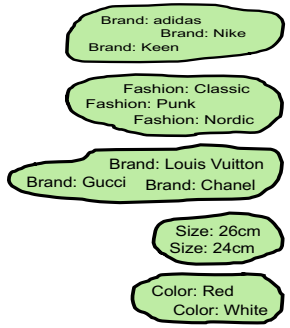
① Clustering labels

② Matching the input to the clusters



Match is a smaller XMC task!

“nike white 26”



(“nike white 26”, Brand: adidas, Brand: Nike, Brand: Keen) → ✓
(“nike white 26”, Fashion: Classic, Fashion: Punk, Fashion: Nordic) → ✗
(“nike white 26”, Brand: Louis Vuitton, Brand: Gucci, Brand: Chanel) → ✗
(“nike white 26”, Size: 26cm, Size: 24cm) → ✓
(“nike white 26”, Color: Red, Color: White) → ✓

③ Classifying the input with OVR

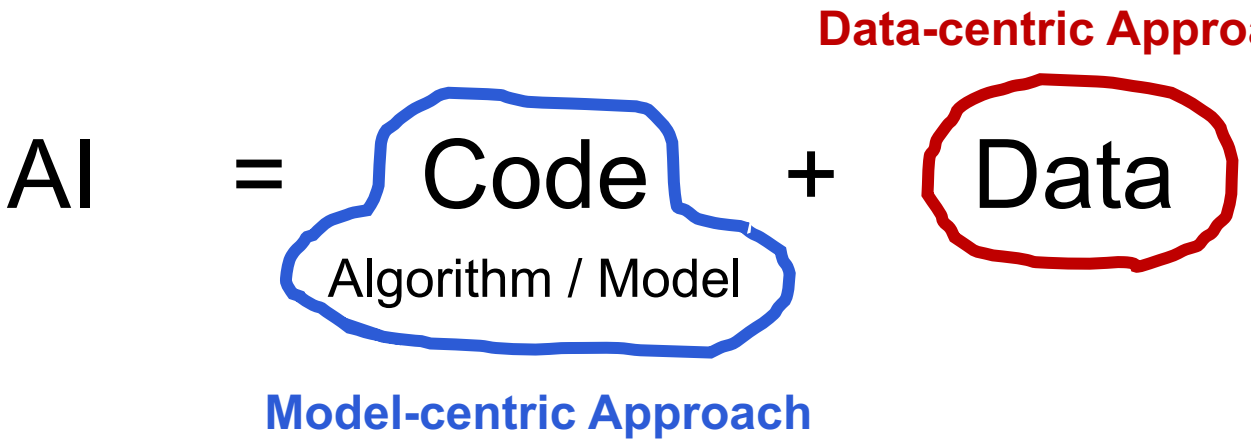
“nike white 26”

eXtreme Multi-label Classification (XMC)
A Research Direction

Data-centric Approach to AI

Discipline of systematically engineering the data used to build an AI system

https://datacentricai.org/



- Efficient data creation
- Label error correction
- ...

NeurIPS Data-centric AI Workshop (2021)

https://datacentricai.org/neurips21/

Model-centric Approach to XMC

Most studies on XMC are model-centric and use public datasets.

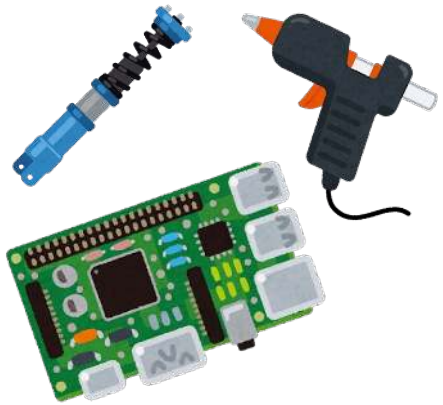
http://manikvarma.org/downloads/XC/XMLRepository.html

Dataset	Download	BoW Feature Dimensionality	Number of Labels	Number of Train Points	Number of Test Points	Avg. Points per Label	Avg. Labels per Point
LF-AmazonTitles-131K	BoW Features Raw text	40,000	131,073	294,805	134,835	5.15	2.29
LF-Amazon-131K	BoW Features Raw text	80,000	131,073	294,805	134,835	5.15	2.29
LF-WikiSeeAlsoTitles-320K	BoW Features Raw text	40,000	312,330	693,082	177,515	4.67	2.11
LF-WikiSeeAlso-320K	BoW Features Raw text	80,000	312,330	693,082	177,515	4.67	2.11
LF-WikiTitles-500K	BoW Features Raw text	80,000	501,070	1,813,391	783,743	17.15	4.74
LF-AmazonTitles-1.3M	BoW Features Raw text	128,000	1,305,265	2,248,619	970,237	38.24	22.20
AmazonCat-13K	BoW Features Raw text	203,882	13,330	1,186,239	306,782	448.57	5.04
AmazonCat-14K	BoW Features Raw text	597,540	14,588	4,398,050	1,099,725	1330.1	3.53
WikiSeeAlsoTitles-350K	BoW Features Raw text	91,414	352,072	629,418	162,491	5.24	2.33
WikiTitles-500K	BoW Features Raw text	185,479	501,070	1,699,722	722,678	23.62	4.89
Wikipedia-500K	BoW Features Raw text	2,381,304	501,070	1,813,391	783,743	24.75	4.77
AmazonTitles-670K	BoW Features Raw text	66,666	670,091	485,176	150,875	5.11	5.39
Amazon-670K	BoW Features Raw text	135,909	670,091	490,449	153,025	3.99	5.45
AmazonTitles-3M	BoW Features Raw text	165,431	2,812,281	1,712,536	739,665	31.55	36.18
Amazon-3M	BoW Features Raw text	337,067	2,812,281	1,717,899	742,507	31.64	36.17
Mediamill	BoW Features	120	101	30,993	12,914	1902.15	4.38
Bibtex	BoW Features	1,836	159	4,880	2,515	111.71	2.40
Delicious	BoW Features	500	983	12,920	3,185	311.61	19.03
RCV1-2K	BoW Features	47,236	2,456	623,847	155,962	1218.56	4.79
EURLex-4K	BoW Features	5,000	3,993	15,539	3,809	25.73	5.31
EURLex-4.3K	BoW Features	200,000	4,271	45,000	6,000	60.57	5.07
Wiki10-31K	BoW Features	101,938	30,938	14,146	6,616	8.52	18.64
Delicious-200K	BoW Features	782,585	205,443	196,606	100,095	72.29	75.54
WikiLSHTC-325K	BoW Features	1,617,899	325,056	1,778,351	587,084	17.46	3.19

Issues for Model-centric Approach to XMC

However, labeled datasets may have wrong and/or missing labels.

Human annotators make mistakes and get tired of long tedious work...



Deep knowledge of the target domain may be required...

In XMC, the number of labels can be millions; it is hard to manually verify datasets thoroughly...

Data-centric Approach to XMC: A Case for E-Commerce

Query-click logs

Clicked products and attribute values

Queries	Product Title	Attribute-Value
"louis vuitton bag"	"Louis Vuitton Bag Zebra print"	Brand: Louis Vuitton Pattern: Zebra
"lv bag zebra"	"LV Louis Vuitton Leather Bag"	Brand: Louis Vuitton Material: Leather

Automatically generated training data

(*"louis vuitton bag"*, Brand: Louis Vuitton)
(*"louis vuitton bag"*, Pattern: Zebra)
(*"louis vuitton bag"*, Material: Leather)
(*"lv bag zebra"*, Brand: Louis Vuitton)
(*"lv bag zebra"*, Pattern: Zebra)
(*"lv bag zebra"*, Material: Leather)

In an e-commerce service, there are various large datasets with weak supervision signals.

- Query-click logs
- Catalog data
- Item classification taxonomy

There are many wrong / missing labels...

- Inconsistent click behaviors of fickle users
- Erroneous retrieval results

How can we utilize these weak signals to generate quality training data?

➢ **We are working on this at RIT now!**

We are Hiring!

RIT Tokyo Language Program

- (Internship) Research Scientist
- (Entry Level) Research Scientist
- Senior Research Scientist

Position Details

- Design algorithms and build proto-type systems to solve defined, scientific problems
- Demonstrate algorithms / proto-type systems to business stakeholders
- Collaborate with engineers to deploy algorithms / proto-type systems into production services
- Collaborate with other scientists on research and paper publications
- Participate in scientific conferences and contribute to scientific community with paper publications

Mandatory Qualifications

- Bachelor's degree in computer science, related research field or equivalent experience
- English fluency for communicating with researchers, engineers, and business stakeholders
- Proficiency in reading and processing Japanese text (our product catalog is written in Japanese)
- Coding skills: fluency in Python
- Experience with Linux environment


Rakuten