ACM India Industry Webinar on

WHY DO WE NEED DATA SCIENCE IN E-COMMERCE?

by

Shourya Roy

Senior Research Director, Flipkart

President, ACM IKDD

19 October 2021

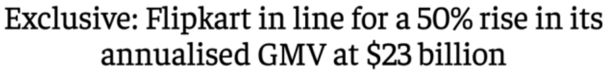
ACM India at a Glance

- **ACM**: world's largest educational and scientific computing society
    - **Mission**: advancing computing as science and profession
    - **Members**: ~100,000 worldwide, ~11000 in India
    - Comprising students, faculty, professionals
- **ACM India Chapters**: ~200 student chapters, ~20 professional chapters
- **ACM-W India**: empowering women in computing
- **Research Initiatives**
    - Student research: ARCS Symposium, best doctoral dissertation, partial travel grant, PhD clinic and Anveshan Setu
    - Research conferences: CODS-COMAD, ISEC, AIMS
- **ACM India Annual Event**
    - Discuss recent trends in technology and celebrate India's achievements in computing
- **Education Initiatives**
    - Summer and winter schools: ~2 week full-time course on technology area
    - Compute: Symposium on computing education
    - Expert Teacher Program: External experts offering a course
    - CSpathshala: inculcate computational thinking in schools
- **Learning and Professional Development**
    - Eminent Speaker Program
    - Industry Webinars, Education Webinars
    - Minigraphs: Comprehensive coverage of a tech area
    - ACM global resources: Digital Library, ACM Learning Center
- **New prestigious awards instituted**
    - Acknowledge and celebrate outstanding contributions
- **ACM Membership in India**
    - Student? student member form
    - Professional? professional member form

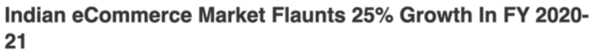
- Senior Research Director at Flipkart
- Prior roles in IBM Research, Xerox Research, and as head of AI Labs American Express
- ACM Distinguished Member
- President of ACM IKDD
- PhD from IISc Bangalore in Machine Learning and Computational Linguistics; Masters from IIT Bombay; Bachelors from Jadhavpur University

Growth of e-commerce

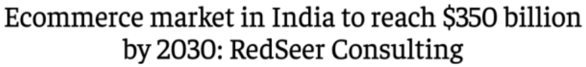
- The growth and spread of e-commerce has been a steady story over a decade or so
  - The last couple of years have been even steeper rise
- e-commerce companies have almost become `The Everything Store’* and the starting point of all purchase intents
- No longer they are only digitizing retail commerce, rather inventing new ways
- Spearheading by leveraging data, software and communication technologies



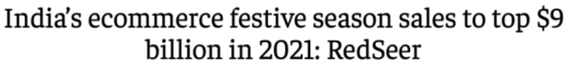
Exclusive: Flipkart in line for a 50% rise in its annualised GMV at $23 billion



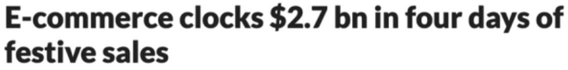
Indian eCommerce Market Flaunts 25% Growth In FY 2020-21



Ecommerce market in India to reach $350 billion by 2030: RedSeer Consulting



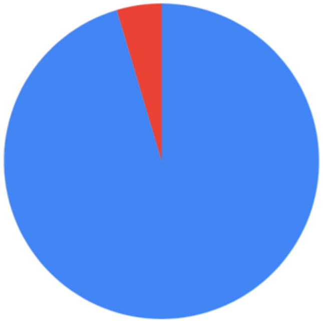
India’s ecommerce festive season sales to top $9 billion in 2021: RedSeer



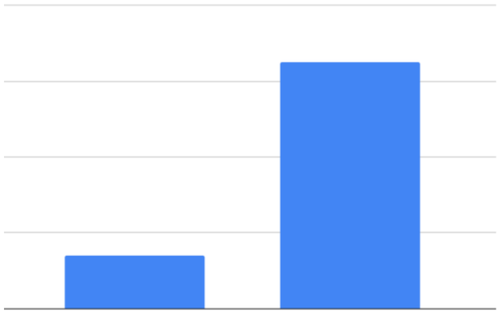
E-commerce clocks $2.7 bn in four days of festive sales

* Brad Stone. 2013. The Everything Store: Jeff Bezos and the Age of Amazon (2nd.ed.). Little, Brown and Company

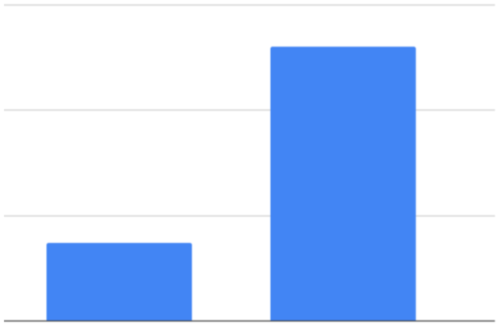
Growth of e-commerce in India



e-commerce penetration is 4.6% of $810B Retail market in India (FY21)



Online shopper base is 140M of internet user base of 625-675M (FY21)

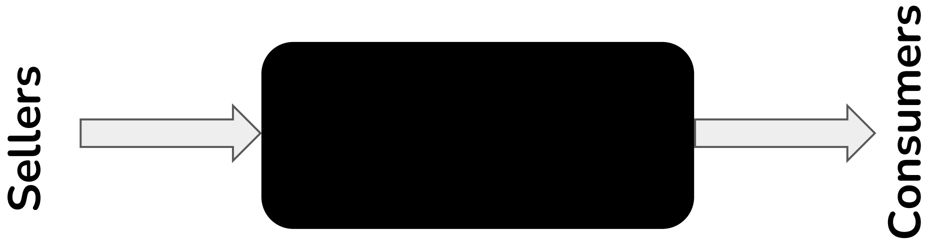


e-commerce market is expected to grow to $120-140B (FY26)

Data and Technology are going to be the two key enablers for growth of e-commerce platforms in the next decade and beyond

“How India Shops Online”; A Study by Bain and Flipkart; https://www.bain.com/insights/how-india-shops-online-2021/

Well, but what’s the big deal in an e-commerce platform?

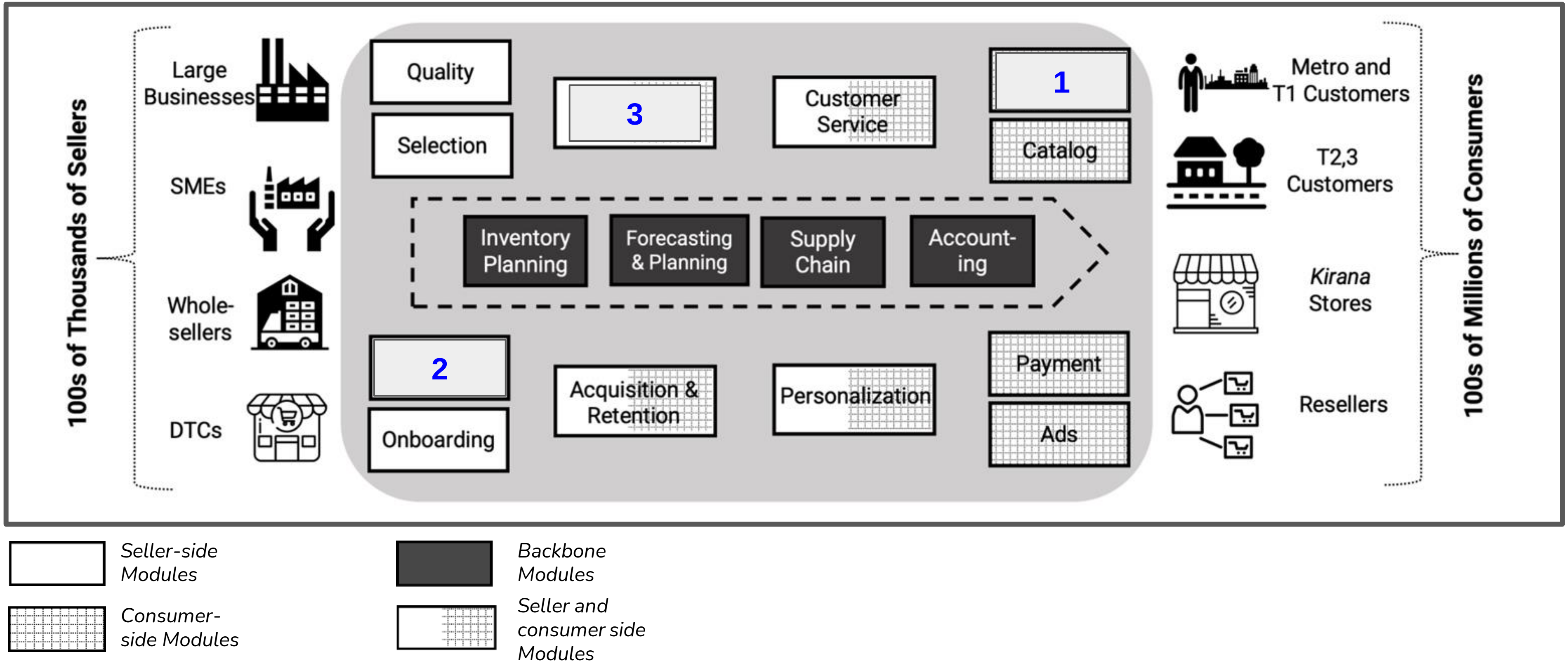


Sellers

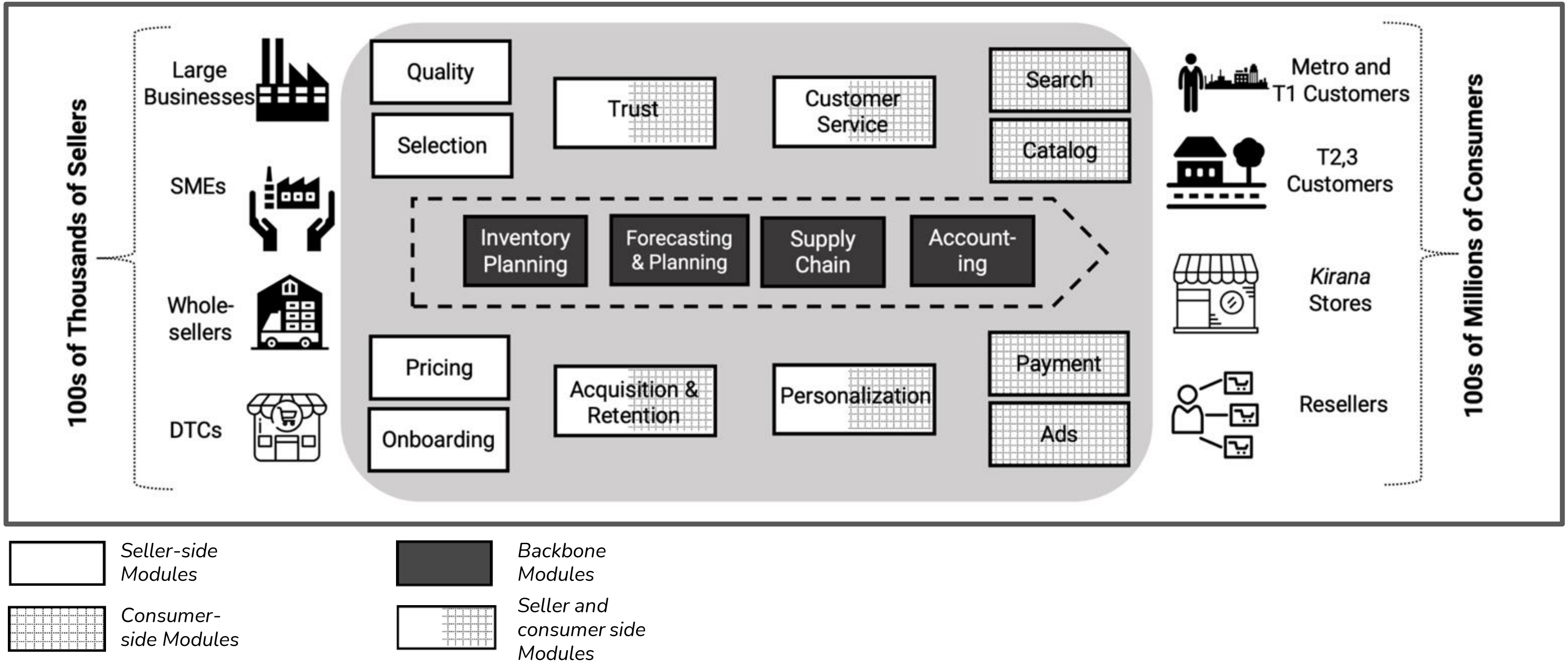
Consumers

Display, Payment, Delivery, …

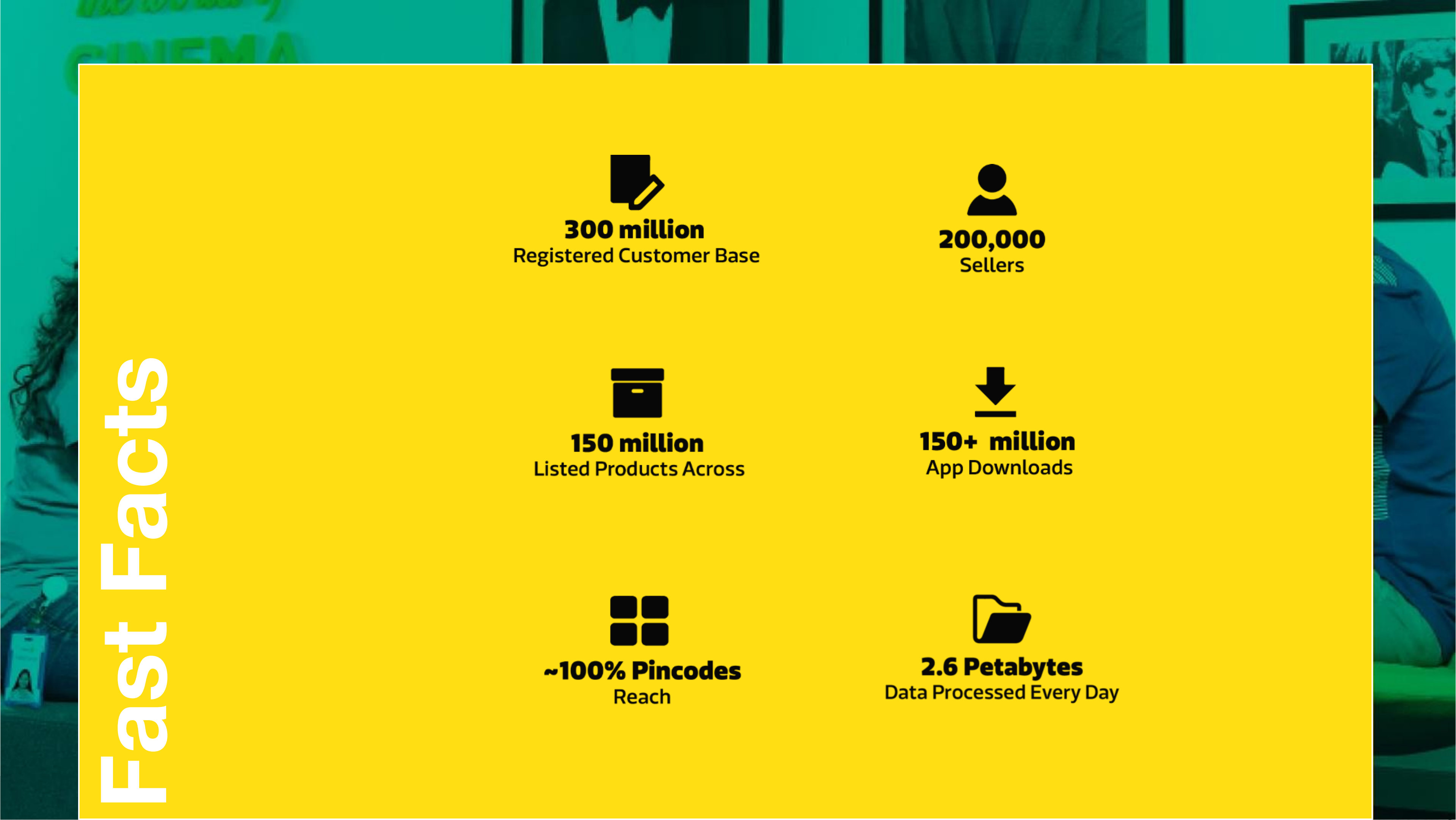
A Schematic of a Large Scale e-commerce Platform



A Schematic of a Large Scale e-commerce Platform



Fast Facts



300 million
Registered Customer Base

200,000
Sellers

150 million
Listed Products Across

150+ million
App Downloads

~100% Pincodes
Reach

2.6 Petabytes
Data Processed Every Day

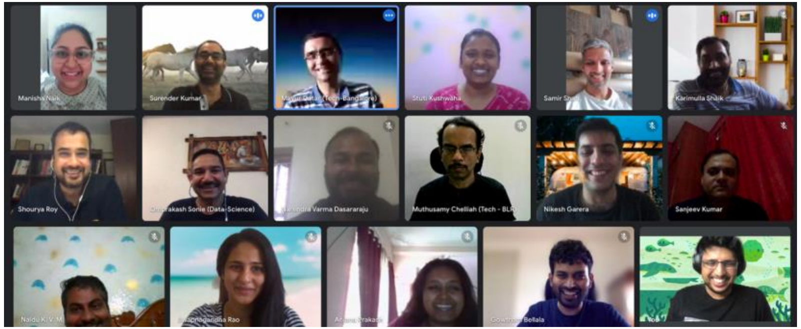
Agenda

- We will double click on a few areas:
  - Forecasting
  - Pricing
  - Trust
  - Catalog
  - User Generated Content

- For each of these:
  - Business Context and Importance
  - Opportunities and Challenges for Data Science
  - Samples of Data Science SOTA

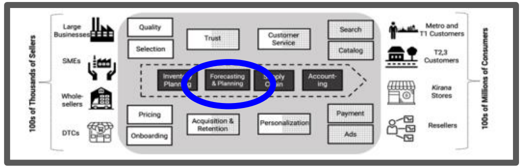
- The talk will be broad and not deep (*except references to deep learning* 😜)

- For convenience, I will be using AI/ML/DS interchangeably (*which is wrong*)



While the talk is mostly based on public domain content, I would like to acknowledge my colleagues at Flipkart from whom I have learnt quite a lot

Forecasting and Planning



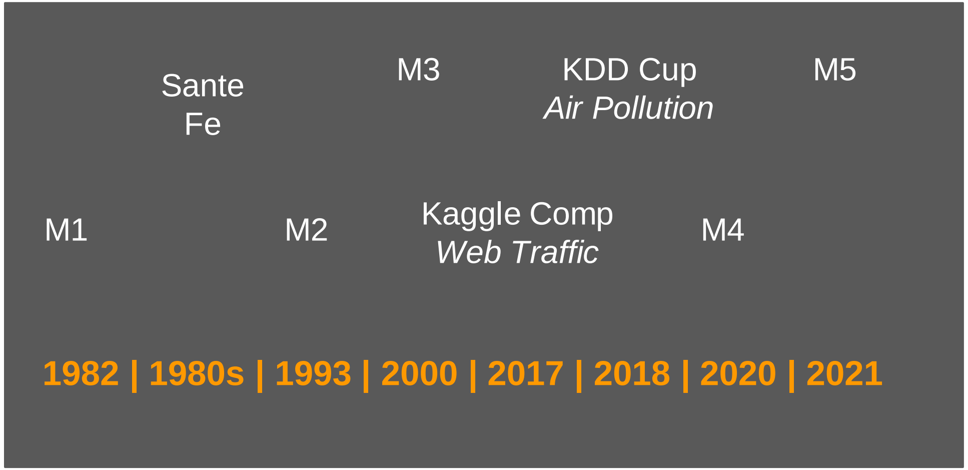
- Prediction of demand and supply
  - Based on the demand, {*what*, *how much*, *when*, *where* and *from whom*} to stock

- Consumers get their products and sellers have predictable shipping schedules
- Critical for ensuring **in-stock** and **speed** of delivery

- The sources of complexity
  - *Infinite* selection of e-commerce platforms
  - Large number of correlated and non-stationary time-series
    - Diversity of products having different life cycles and trends
  - Range of granularities along product, time and geography dimensions
  - Seasonal variations, bundle offers, promotions, sales, out-of-fashion and new product

What: *FastColors; Full Sleeve Solid Men Sweatshirt; Black-red; XL*
How Many: *Three*
When and Where: *17/10 (Srinagar); 20/10 (Guntur); 20/10 (Amritsar)*
From Whom: *Supp-N in North and Supp-S in South*

Back to the Forecast!



Sante Fe

M3

KDD Cup
Air Pollution

M5

M1

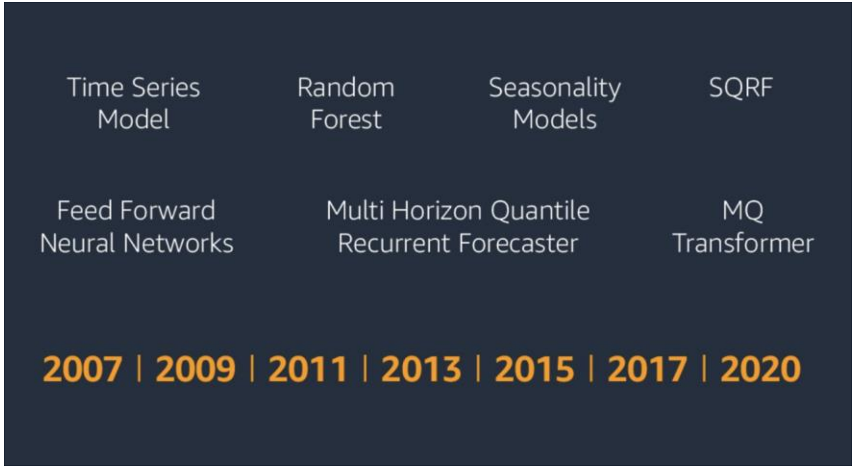
M2

Kaggle Comp
Web Traffic

M4

1982 | 1980s | 1993 | 2000 | 2017 | 2018 | 2020 | 2021

Timeline of Major Forecasting Competitions



Time Series Model

Random Forest

Seasonality Models

SQRF

Feed Forward Neural Networks

Multi Horizon Quantile Recurrent Forecaster

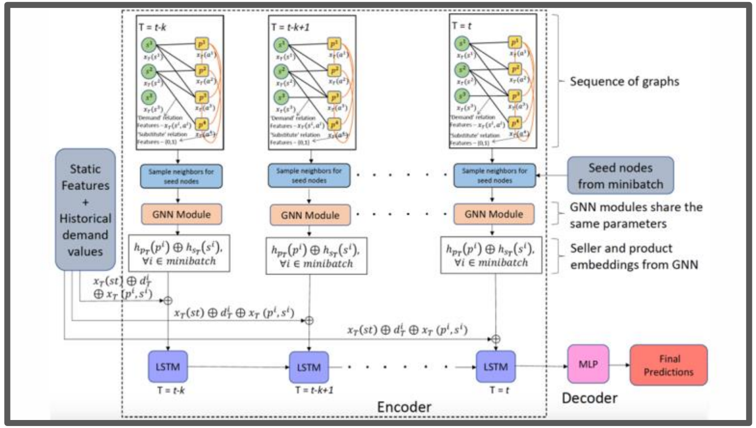
MQ Transformer

2007 | 2009 | 2011 | 2013 | 2015 | 2017 | 2020

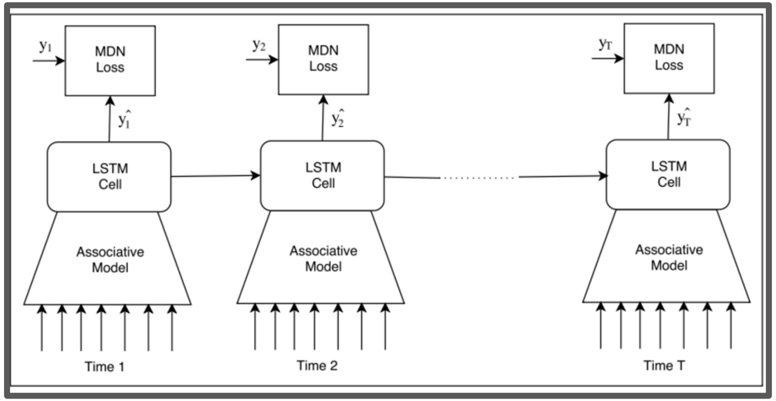
Evolution of Forecasting Techniques

A brief history of forecasting competitions;Rob J.Hyndman;https://www.sciencedirect.com/science/article/abs/pii/S016920701930086X
The history of Amazon's forecasting algorithm; https://www.amazon.science/latest-news/the-history-of-amazons-forecasting-algorithm

Samples of Recent Research


T = t-k
T = t-k+1
T = t
s^1 $x_T(s^1)$
s^2 $x_T(s^2)$
s^3 $x_T(s^3)$
p^1 $x_T(a^1)$
p^2 $x_T(a^2)$
p^3 $x_T(a^3)$
p^4 $x_T(a^4)$
'Demand' relation Features – $x_T(s^i, a^i)$
'Substitute' relation Features – {0,1}
Sequence of graphs
Static Features + Historical demand values
Sample neighbors for seed nodes
Seed nodes from minibatch
GNN Module
GNN modules share the same parameters
$h_{p_T}(p^i) \oplus h_{s_T}(s^i)$, $\forall i \in minibatch$
Seller and product embeddings from GNN
$x_T(st) \oplus d_T^i \oplus x_T(p^i, s^i)$
LSTM
MLP
Final Predictions
Encoder
Decoder

1. GNNs on sequence of hypergraphs (***Amazon***)



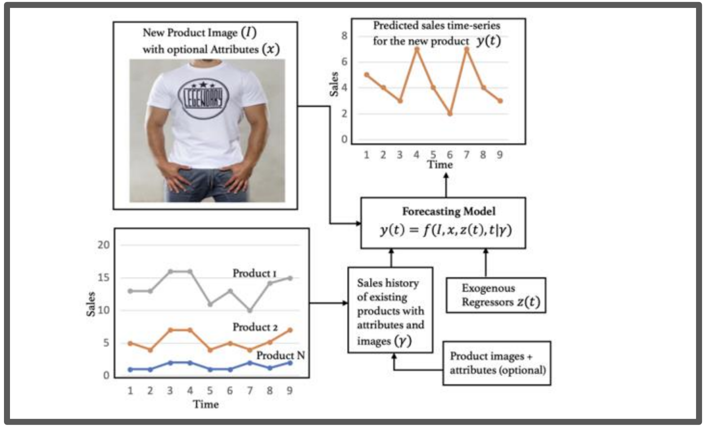
2. Associative and Recurrent Mixture Density Networks (***Flipkart***)

- Demand prediction considering interdependencies between seller and products
- Hypergraph with product and seller nodes with demand and substitute relations
- Time series of GNN and LSTM followed by a final layer of MLP

- Attempts to address similar dependency factors between products
- Modeled using an ensemble of MLP and LSTM
- Output is probability distribution over demands as a mixture of Gaussians

[1] Spatio-temporal multi-graph networks for demand forecasting in online marketplaces; Gandhi et al; 2021;
[2] ARMDN: Associative and Recurrent Mixture Density Networks for eRetail Demand Forecasting; Mukherjee et al; 2018

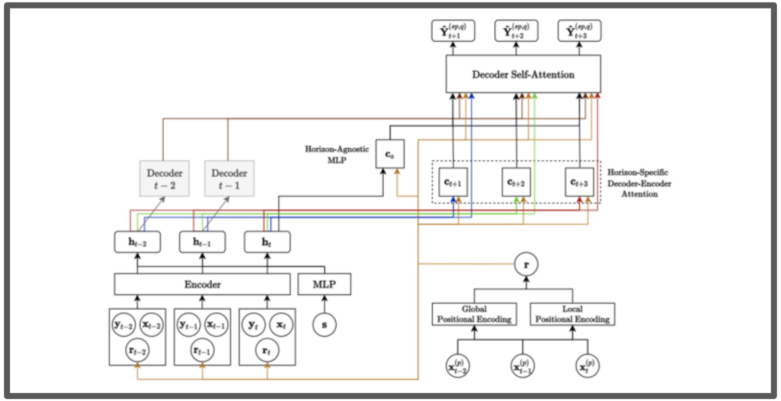
Samples of Recent Research



New Product Image (I)
with optional Attributes (x)
Predicted sales time-series
for the new product $y(t)$
Sales
Time
Forecasting Model
$y(t) = f(I, x, z(t), t|\gamma)$
Product 1
Product 2
Product N
Sales history of existing products with attributes and images (γ)
Exogenous Regressors $z(t)$
Product images + attributes (optional)

3. Seq2seq using image and structured features (*IBM*)

- Demand prediction for fashion apparels
- Additional challenges of huge dead unsold inventory, higher volume and velocity of introduction of new products
- Multiple models based on textual and image features e.g. kNN, encoder-decoder based models etc.



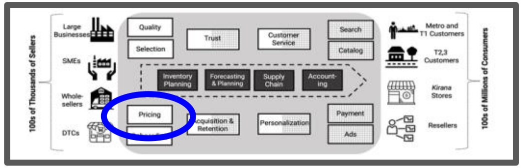
4. Seq2seq with attention and positional encoding (*Amazon*)

- Application of Transformer architecture to demand forecasting
- Interesting application of concepts viz. self-attention, positional encoding
- Provides SOTA results on multiple datasets with improvements in situations such as seasonal peaks and promotions

[3] Spatio-temporal multi-graph networks for demand forecasting in online marketplaces; Gandhi et al; 2021;

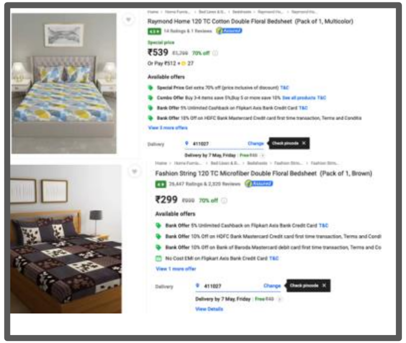
[4] MQTransformer: Multi-Horizon Forecasts with Context Dependent and Feedback-Aware Attention; Eisenach at al; 2020

Pricing



- Predict optimal price increasing the likelihood of a product to be sold
  - Recommendation for seller at the time of listing
  - Shape customer demand
  - Gain competitive advantage
- Increase profitability of sellers and platform while providing the best price to consumers

- The sources of complexity
  - Dependent on various factors *brand*, *quality*, *supply & demand*, *production cost & volume*, *competition*
  - Short duration price fluctuations due to sales, promotional events etc
  - Inadequate, incomplete and poor quality of data



Home › Home Furnis... › Bed Linen & B... › Bedsheets › Raymond Ho... › Raymond Ho...

Raymond Home 120 TC Cotton Double Floral Bedsheet (Pack of 1, Multicolor)

4.1 ★ 14 Ratings & 1 Reviews Assured

Special price

₹539 ₹1,799 70% off

Or Pay ₹512 + 27

Available offers

- **Special Price** Get extra 70% off (price inclusive of discount) T&C
- **Combo Offer** Buy 3-4 items save 5%;Buy 5 or more save 10% See all products T&C
- **Bank Offer** 5% Unlimited Cashback on Flipkart Axis Bank Credit Card T&C
- **Bank Offer** 10% Off on HDFC Bank Mastercard Credit card first time transaction, Terms and Conditio

View 3 more offers

Delivery 411027 Change Check pincode ✕

Delivery by 7 May, Friday | Free ₹40

Home › Home Furnis... › Bed Linen & B... › Bedsheets › Fashion Strin... › Fashion Strin...

Fashion String 120 TC Microfiber Double Floral Bedsheet (Pack of 1, Brown)

4 ★ 26,447 Ratings & 2,320 Reviews Assured

₹299 ₹999 70% off

Available offers

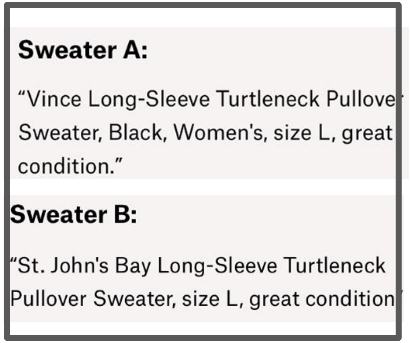
- **Bank Offer** 5% Unlimited Cashback on Flipkart Axis Bank Credit Card T&C
- **Bank Offer** 10% Off on HDFC Bank Mastercard Credit card first time transaction, Terms and Condi
- **Bank Offer** 10% Off on Bank of Baroda Mastercard debit card first time transaction, Terms and Co
- No Cost EMI on Flipkart Axis Bank Credit Card T&C

View 1 more offer

Delivery 411027 Change Check pincode ✕

Delivery by 7 May, Friday | Free ₹40

View Details



Sweater A:

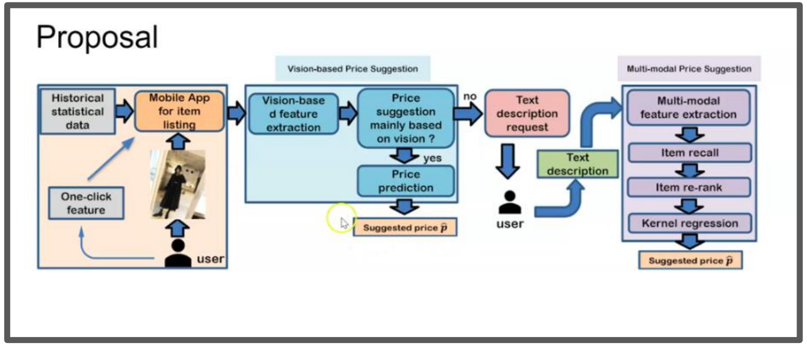
"Vince Long-Sleeve Turtleneck Pullove Sweater, Black, Women's, size L, great condition."

Sweater B:

"St. John's Bay Long-Sleeve Turtleneck Pullover Sweater, size L, great condition

https://medium.com/swlh/mercari-price-suggestion-challenge-an-end-to-end-machine-learning-case-study-4a6d833fa1c7

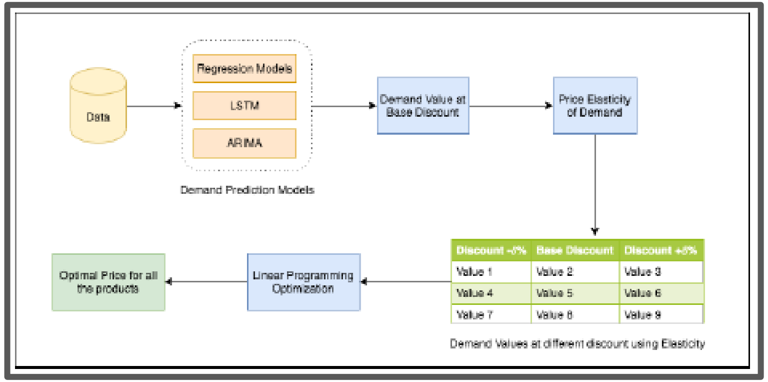
Samples of Recent Research



Proposal

Vision-based Price Suggestion
Multi-modal Price Suggestion
Historical statistical data
Mobile App for item listing
One-click feature
user
Vision-based feature extraction
Price suggestion mainly based on vision ?
no
yes
Price prediction
Suggested price $\hat{p}$
Text description request
user
Text description
Multi-modal feature extraction
Item recall
Item re-rank
Kernel regression
Suggested price $\hat{p}$

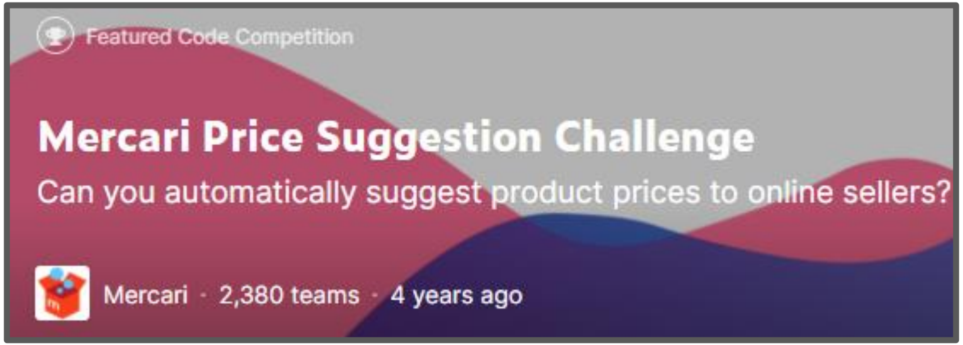
[1] For 2nd-hand items (from?)



Data
Regression Models
LSTM
ARIMA
Demand Prediction Models
Demand Value at Base Discount
Price Elasticity of Demand
Optimal Price for all the products
Linear Programming Optimization
Demand Values at different discount using Elasticity

Discount -δ%	Base Discount	Discount +δ%
Value 1	Value 2	Value 3
Value 4	Value 5	Value 6
Value 7	Value 8	Value 9

[2] For fashion e-commerce (from?)



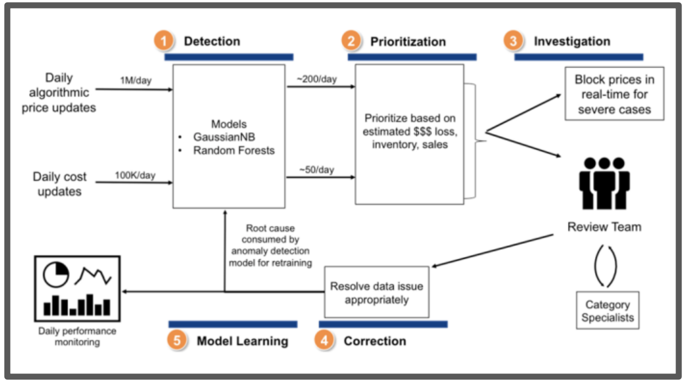
Featured Code Competition

Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Mercari · 2,380 teams · 4 years ago

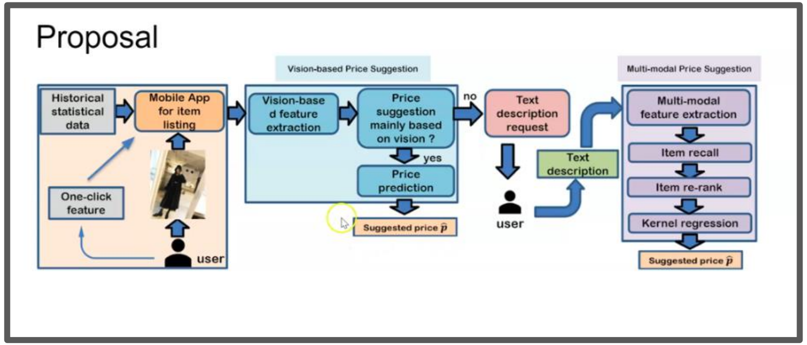
[3] For a community marketplace (from?)



1 Detection
2 Prioritization
3 Investigation
Daily algorithmic price updates
1M/day
Daily cost updates
100K/day
Models
• GaussianNB
• Random Forests
~200/day
~50/day
Prioritize based on estimated $$$ loss, inventory, sales
Block prices in real-time for severe cases
Review Team
Category Specialists
Root cause consumed by anomaly detection model for retraining
Resolve data issue appropriately
Daily performance monitoring
5 Model Learning
4 Correction

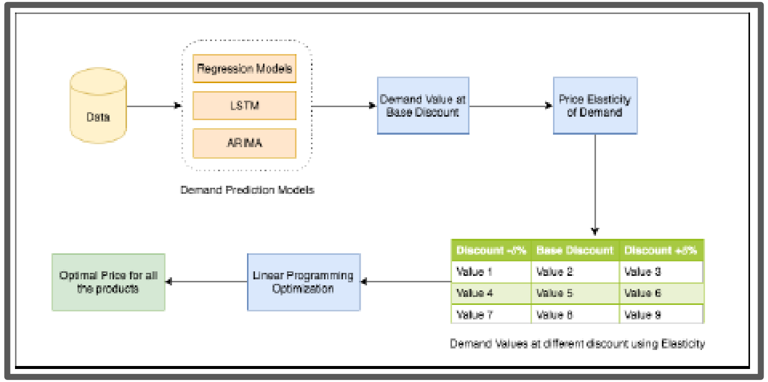
[4] Pricing anomaly detection (from?)

Samples of Recent Research


Proposal
Vision-based Price Suggestion
Multi-modal Price Suggestion
Historical statistical data
Mobile App for item listing
One-click feature
user
Vision-based feature extraction
Price suggestion mainly based on vision ?
no
yes
Price prediction
Suggested price $\hat{p}$
Text description request
user
Text description
Multi-modal feature extraction
Item recall
Item re-rank
Kernel regression
Suggested price $\hat{p}$

[1] For 2nd-hand items (from?)

Liang Han, Zhaozheng Yin, Zhurong Xia, Mingqian Tang, Rong Jin

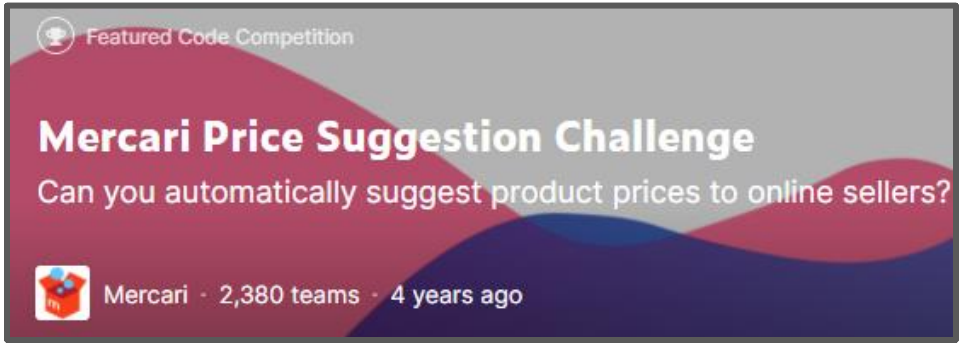

Data
Regression Models
LSTM
ARIMA
Demand Prediction Models
Demand Value at Base Discount
Price Elasticity of Demand
Linear Programming Optimization
Optimal Price for all the products

Discount -δ%	Base Discount	Discount +δ%
Value 1	Value 2	Value 3
Value 4	Value 5	Value 6
Value 7	Value 8	Value 9

Demand Values at different discount using Elasticity

[2] For fashion e-commerce (from?)

Sajan Kedia, Samyak Jain, Abhishek Sharma



Featured Code Competition

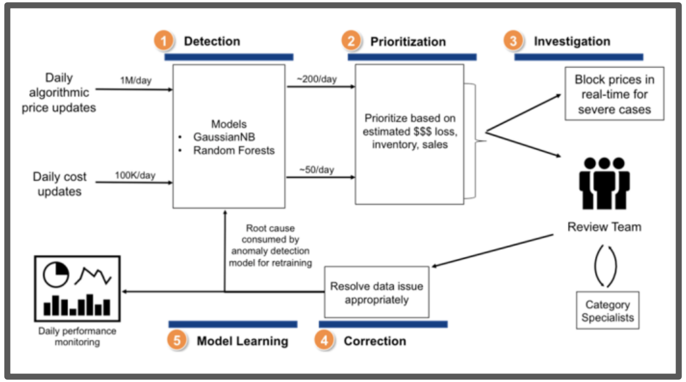
Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Mercari · 2,380 teams · 4 years ago

[3] For a community marketplace (from?)

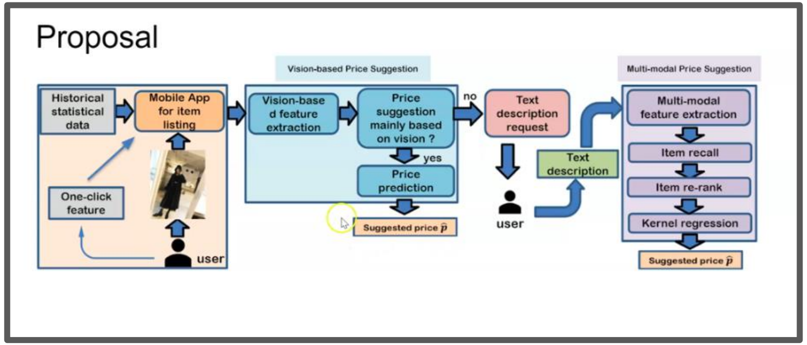
Kaggle Competition


1 Detection
2 Prioritization
3 Investigation
Daily algorithmic price updates
1M/day
Daily cost updates
100K/day
Models
• GaussianNB
• Random Forests
~200/day
~50/day
Prioritize based on estimated $$$ loss, inventory, sales
Block prices in real-time for severe cases
Review Team
Category Specialists
Root cause consumed by anomaly detection model for retraining
Resolve data issue appropriately
Daily performance monitoring
5 Model Learning
4 Correction

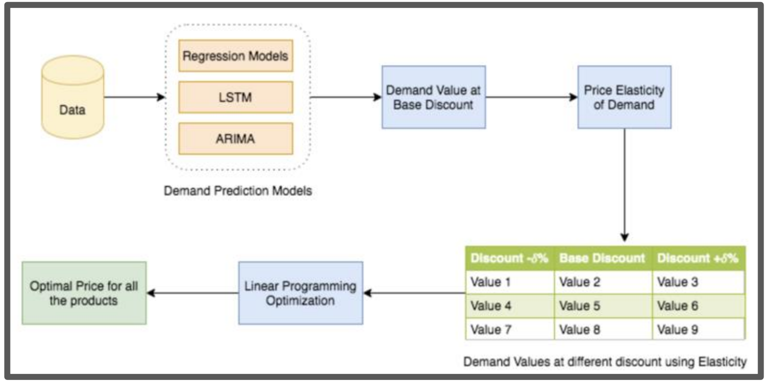
[4] Pricing anomaly detection (from?)

Jagdish Ramakrishnan, Elham Shaabani, Chao Li, Mátyás A. Sustik

Samples of Recent Research


Proposal
Vision-based Price Suggestion
Multi-modal Price Suggestion
Historical statistical data
Mobile App for item listing
One-click feature
user
Vision-base d feature extraction
Price suggestion mainly based on vision ?
no
yes
Price prediction
Suggested price $\hat{p}$
Text description request
user
Text description
Multi-modal feature extraction
Item recall
Item re-rank
Kernel regression
Suggested price $\hat{p}$

[1] For 2nd-hand items (***Alibaba***)


Data
Regression Models
LSTM
ARIMA
Demand Prediction Models
Demand Value at Base Discount
Price Elasticity of Demand
Linear Programming Optimization
Optimal Price for all the products

Discount -δ%	Base Discount	Discount +δ%
Value 1	Value 2	Value 3
Value 4	Value 5	Value 6
Value 7	Value 8	Value 9

Demand Values at different discount using Elasticity

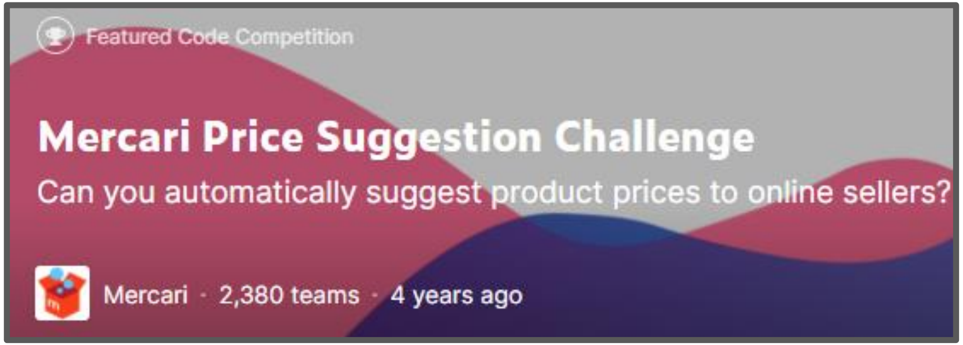
[2] For fashion e-commerce (***Myntra***)

- Multi-modal inputs: textual, visual and statistical item features
- Binary classification followed by regression for price suggestion
- Additional constraints for different demand, customized loss function to facilitate more transactions

- Three stage technique for optimal pricing for clothing and apparels
  - Demand prediction at different discount levels
  - Price-elasticity based model to obtain different demand values
  - Choosing the most optimal permutation of demand-price pairs

[1] Price Suggestion for Online Second-hand Items with Texts and Images; Han et al.; 2020; [2] Price Optimization in Fashion E-commerce; Kedia et al.; 2020

Samples of Recent Research



Featured Code Competition

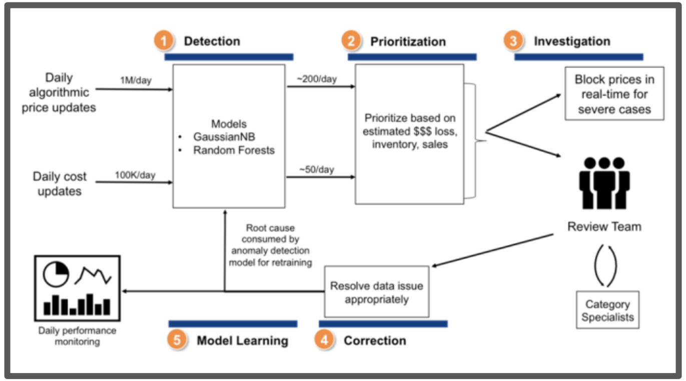
Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Mercari · 2,380 teams · 4 years ago

[3] For a community marketplace (*Mercari, Kaggle*)

- Kaggle competition for predicting prices of second-hand items based on textual and structured features
- Winning team had an ensemble of multiple models MLP, LGBM at different granularities
- Good dataset to explore and understand pricing nuances

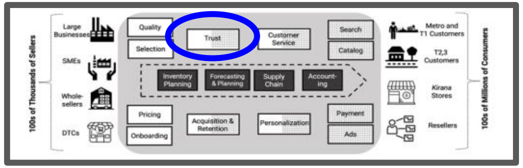

1 Detection
2 Prioritization
3 Investigation
Daily algorithmic price updates
1M/day
Daily cost updates
100K/day
Models
• GaussianNB
• Random Forests
~200/day
~50/day
Prioritize based on estimated $$$ loss, inventory, sales
Block prices in real-time for severe cases
Review Team
Category Specialists
Root cause consumed by anomaly detection model for retraining
Resolve data issue appropriately
Daily performance monitoring
5 Model Learning
4 Correction

[4] Pricing anomaly detection (*Walmart*)

- Comparison of a number of supervised and unsupervised approaches
- Novelty is mostly in terms of retail-specific feature engineering
- Addresses multiple practical considerations e.g. business-led prioritization, manual review/override etc

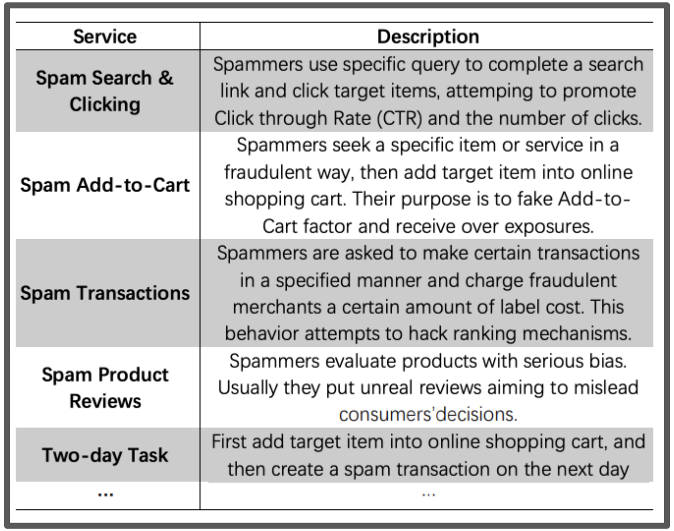
[3] Mercari Price Suggestion Challenge; Mercari; 2017; [4] Anomaly Detection for an E-commerce Pricing System; Ramakrishnan et al.; 2019;

Trust



- Making e-commerce platforms trusted by mitigating fraud and curtailing abuse
  - Fraud and abuse can happen by any party in the e-commerce ecosystem viz. *consumers*, *sellers*, *platform*
  - **Fraud**: Stolen card, missing products, return-to-origin, supply chain fraud, delivery fraud
  - **Abuse**: Excessive return, seller cancellation, reseller
- Leads to negative (bottomline) monetary impact and poor customer experience
- Uber goals are elimination of fraudsters and behaviour shaping

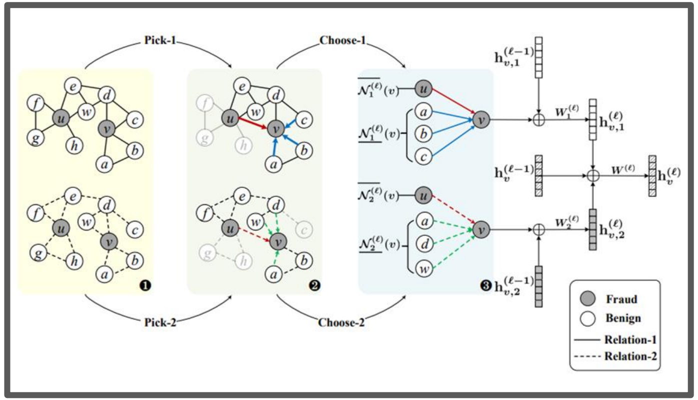
- The sources of complexity
  - Continuously evolving nature of fraud and abuse
  - Scale and diversity of data and lack of sacrosanct labels
  - *Walking on the thin ice* of decision making



Service	Description
Spam Search & Clicking	Spammers use specific query to complete a search link and click target items, attemping to promote Click through Rate (CTR) and the number of clicks.
Spam Add-to-Cart	Spammers seek a specific item or service in a fraudulent way, then add target item into online shopping cart. Their purpose is to fake Add-to-Cart factor and receive over exposures.
Spam Transactions	Spammers are asked to make certain transactions in a specified manner and charge fraudulent merchants a certain amount of label cost. This behavior attempts to hack ranking mechanisms.
Spam Product Reviews	Spammers evaluate products with serious bias. Usually they put unreal reviews aiming to mislead consumers'decisions.
Two-day Task	First add target item into online shopping cart, and then create a spam transaction on the next day
...	...

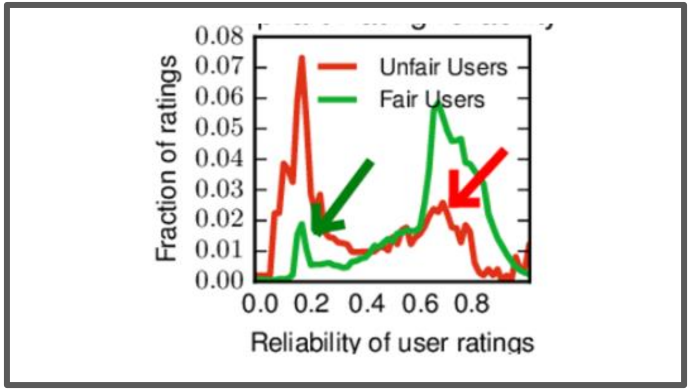
Collaboration Based Multi-Label Propagation for Fraud Detection; Wang et al; 2020

Samples of Recent Research



[1] GNN-based fraud detection (*Alibaba*)

- Label propagation on a network of users through sampling of nodes for handling class imbalance
  - Two step process - *pick* and *choose*
- Graph Neural Network to obtain user/node embeddings followed by classification

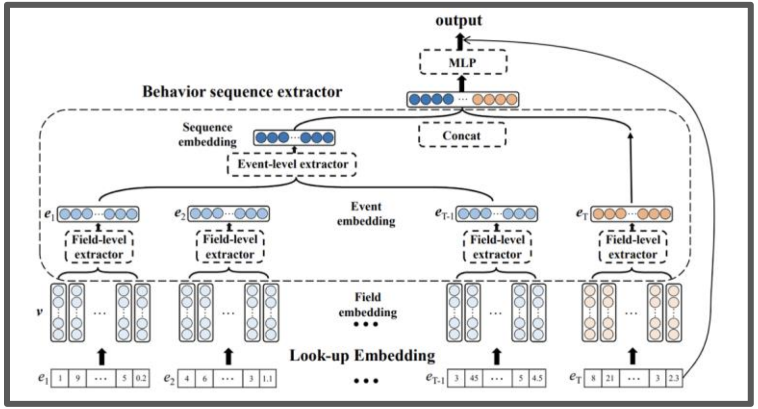


[2] Fraudulent rating detection (*Flipkart*)

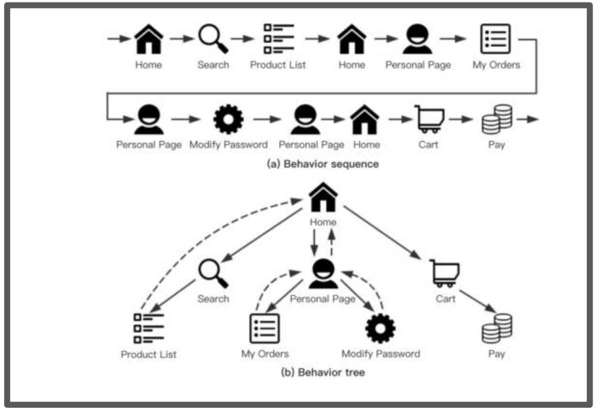
- Recursive modeling of three intrinsic quality metrics
  - fairness of a user
  - reliability of a rating
  - goodness of a product

[1] Pick and Choose: A GNN-based Imbalanced Learning Approach for Fraud Detection; Liu et al; 2021 [2] REV2: Fraudulent User Prediction in Rating Platforms; Kumar et al; 2018;

Samples of Recent Research



[3] Hierarchical Explainable Network (HEN) (***Alibaba***)



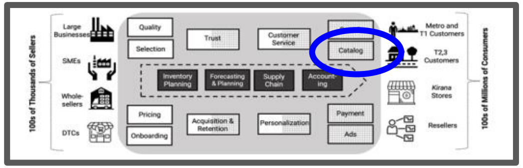
[4] Behavior Tree with Local Intention(***Alibaba***)

- Seq2seq modeling of users' historical behaviours
- Predict if a future payment is fraudulent

- Leverage webpage hierarchy reflected in page-jumps capturing user intents
- Modeled through LSTM with behaviour tree as input and predict if a payment is fraudulent

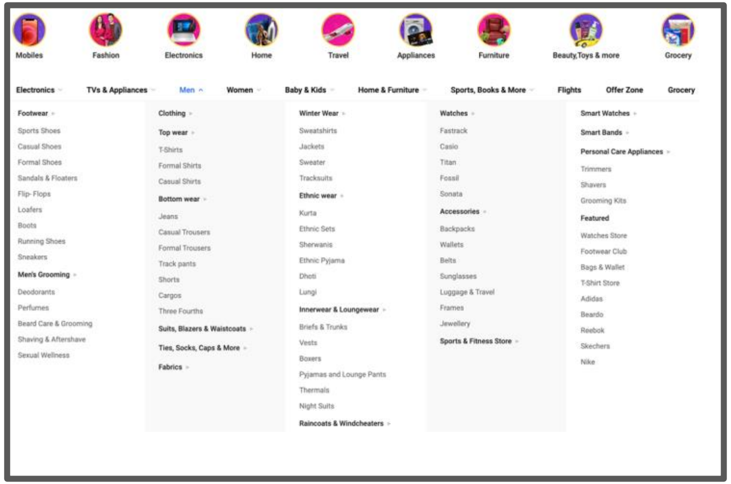
[3] Modeling Users' Behavior Sequences with Hierarchical Explainable Network for Cross-domain Fraud Detection; Zhu et al; 2020
[4] Fraud Transactions Detection via Behavior Tree with Local Intention Calibration; Liu et al; 2020

Product Catalog



- Large taxonomy of all products available on the platform where sellers add for consumers to explore/browse/purchase
- Huge size with thousands of leaf nodes; 5-10 levels of depth; hundreds of millions of products with 10-100+ attributes
- High velocity of addition/deletion and updates

- The sources of complexity
  - Very large scale hierarchical classification with highly imbalanced number of samples
  - *Vocabulary gap* between product descriptions and consumer search intent
  - Maintaining *correctness*, *uniqueness* and *recency* with constant addition/deletion/updation
  - Supporting emerging applications such as visual search, multimodal search, code-mixed search etc.



Mobiles Fashion Electronics Home Travel Appliances Furniture Beauty, Toys & more Grocery

Electronics | TVs & Appliances | Men | Women | Baby & Kids | Home & Furniture | Sports, Books & More | Flights | Offer Zone | Grocery

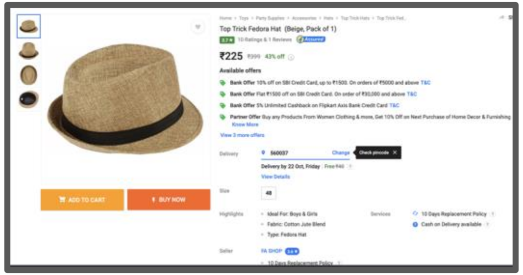
Footwear
Sports Shoes
Casual Shoes
Formal Shoes
Sandals & Floaters
Flip- Flops
Loafers
Boots
Running Shoes
Sneakers
Men's Grooming
Deodorants
Perfumes
Beard Care & Grooming
Shaving & Aftershave
Sexual Wellness

Clothing
Top wear
T-Shirts
Formal Shirts
Casual Shirts
Bottom wear
Jeans
Casual Trousers
Formal Trousers
Track pants
Shorts
Cargos
Three Fourths
Suits, Blazers & Waistcoats
Ties, Socks, Caps & More
Fabrics

Winter Wear
Sweatshirts
Jackets
Sweater
Tracksuits
Ethnic wear
Kurta
Ethnic Sets
Sherwanis
Ethnic Pyjama
Dhoti
Lungi
Innerwear & Loungewear
Briefs & Trunks
Vests
Boxers
Pyjamas and Lounge Pants
Thermals
Night Suits
Raincoats & Windcheaters

Watches
Fastrack
Casio
Titan
Fossil
Sonata
Accessories
Backpacks
Wallets
Belts
Sunglasses
Luggage & Travel
Frames
Jewellery
Sports & Fitness Store

Smart Watches
Smart Bands
Personal Care Appliances
Trimmers
Shavers
Grooming Kits
Featured
Watches Store
Footwear Club
Bags & Wallet
T-Shirt Store
Adidas
Beardo
Reebok
Skechers
Nike



Home > Toys > Party Supplies > Accessories > Hats > Top Trick Hats > Top Trick Fed...

Top Trick Fedora Hat (Beige, Pack of 1)

3.7★ 10 Ratings & 1 Reviews Assured

₹225 ~~₹399~~ 43% off

Available offers

- **Bank Offer** 10% off on SBI Credit Card, up to ₹1500. On orders of ₹5000 and above T&C
- **Bank Offer** Flat ₹1500 off on SBI Credit Card. On order of ₹30,000 and above T&C
- **Bank Offer** 5% Unlimited Cashback on Flipkart Axis Bank Credit Card T&C
- **Partner Offer** Buy any Products From Women Clothing & more, Get 10% Off on Next Purchase of Home Decor & Furnishing Know More

View 3 more offers

Delivery 560037 Change Check pincode ✕

Delivery by 22 Oct, Friday | Free ₹40

View Details

Size 48

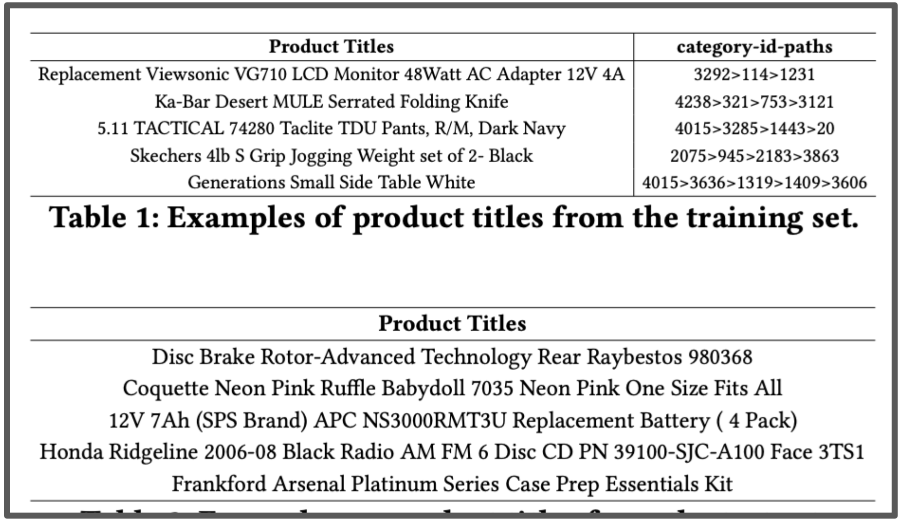
Highlights
- Ideal For: Boys & Girls
- Fabric: Cotton Jute Blend
- Type: Fedora Hat

Services
- 10 Days Replacement Policy
- Cash on Delivery available

Seller FA SHOP 3.6★
- 10 Days Replacement Policy

Taxonomies for E-commerce: Best Practices and Design Challenges; 2012; https://www.slideshare.net/HeatherHedden/taxonomies-for-ecommerce

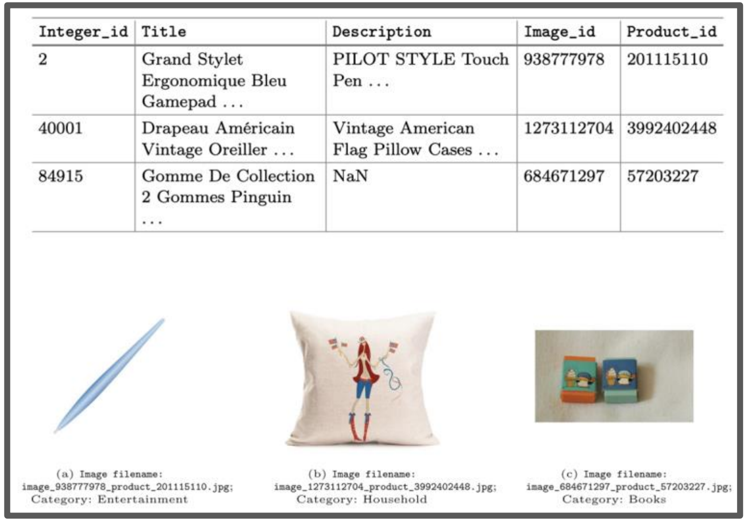
Catalog Datasets



Product Titles	category-id-paths
Replacement Viewsonic VG710 LCD Monitor 48Watt AC Adapter 12V 4A	3292>114>1231
Ka-Bar Desert MULE Serrated Folding Knife	4238>321>753>3121
5.11 TACTICAL 74280 Taclite TDU Pants, R/M, Dark Navy	4015>3285>1443>20
Skechers 4lb S Grip Jogging Weight set of 2- Black	2075>945>2183>3863
Generations Small Side Table White	4015>3636>1319>1409>3606

Table 1: Examples of product titles from the training set.

Product Titles
Disc Brake Rotor-Advanced Technology Rear Raybestos 980368
Coquette Neon Pink Ruffle Babydoll 7035 Neon Pink One Size Fits All
12V 7Ah (SPS Brand) APC NS3000RMT3U Replacement Battery (4 Pack)
Honda Ridgeline 2006-08 Black Radio AM FM 6 Disc CD PN 39100-SJC-A100 Face 3TS1
Frankford Arsenal Platinum Series Case Prep Essentials Kit



Integer_id	Title	Description	Image_id	Product_id
2	Grand Stylet Ergonomique Bleu Gamepad ...	PILOT STYLE Touch Pen ...	938777978	201115110
40001	Drapeau Américain Vintage Oreiller ...	Vintage American Flag Pillow Cases ...	1273112704	3992402448
84915	Gomme De Collection 2 Gommes Pinguin ...	NaN	684671297	57203227

(a) Image filename: image_938777978_product_201115110.jpg; Category: Entertainment

(b) Image filename: image_1273112704_product_3992402448.jpg; Category: Household

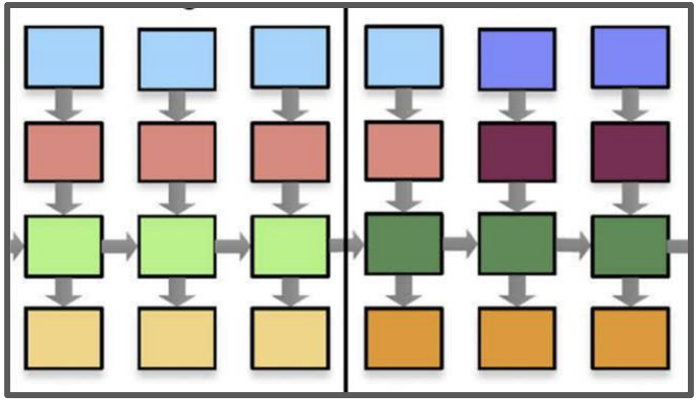
(c) Image filename: image_684671297_product_57203227.jpg; Category: Books

- **Textual dataset** of one million product titles and the corresponding anonymized category paths from their entire product catalog
  - Over three thousand leaf level nodes (#classes)
- **Evaluation**: weighted-precision, weighted-recall and weighted-F1 for the test set of exact "category-id-path" match

- A **multi-modal dataset** of ~100K product listings comprising textual titles and description and product image
- **Tasks**: [1] large-scale multi-modal classification and [2] cross-modal retrieval
- **Evaluation**: [1] macro-F1 score and [2] recall@1

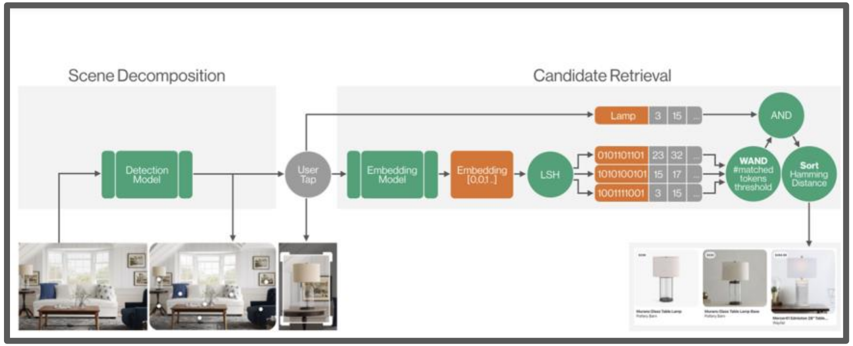
Overview of the SIGIR 2018 eCom Rakuten Data Challenge; Lin et al; 2018
An E-Commerce Dataset in French for Multi-modal Product Categorization and Cross-Modal Retrieval; Amoualian et al; 2020

Samples of Recent Research



[1] Product categorization as NMT (*Rakuten*)

- Large-scale hierarchical categorization problem
    - primarily text (title, description); multi-modal (product images)
- NMT Formulation: *text -> path-in-catalog-tree*
- Improvement in benchmark datasets

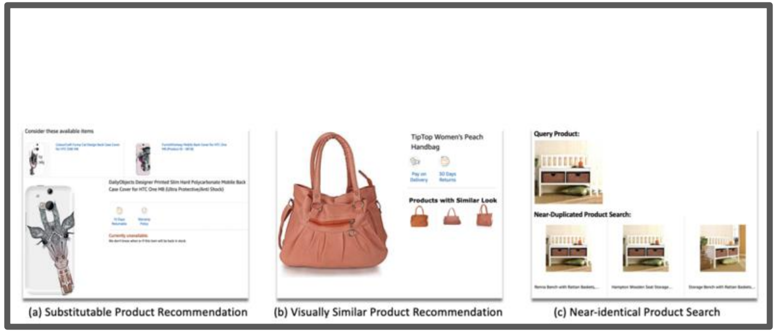


[2] Visual search through OD (*Pinterest*)

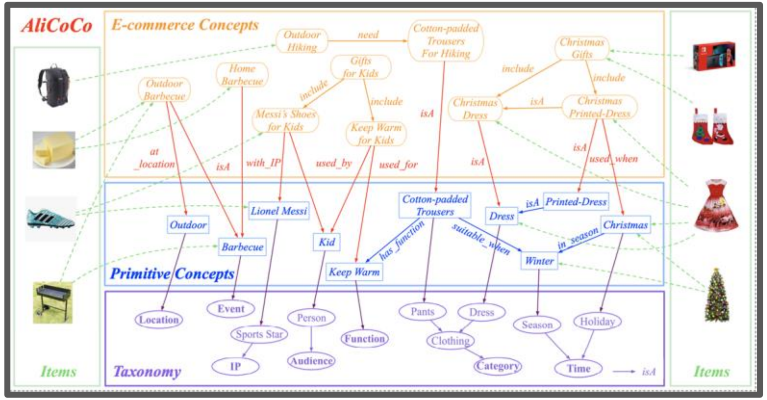
- Multiple use-cases of object detection e.g. shop-the-look, complete-the-look
- Common approach: Object detection followed by candidate retrieval
- e-commerce specific customization e.g. category filter
- Other downstream applications include auto-enrichment, catalog completion

[1] E-Commerce Product Categorization via Machine Translation;Tan et al; 2020; [2] Shop The Look: Building a Large Scale Visual Shopping System at Pinterest;Shiau et al.; 2020;

Samples of Recent Research


Consider these available items
DailyObjects Designer Printed Slim Hard Polycarbonate Mobile Back Case Cover for HTC One M8 (Ultra Protective/Anti Shock)
Currently unavailable.
TipTop Women's Peach Handbag
Pay on Delivery
30 Days Returns
Products with Similar Look
Query Product:
Near-Duplicated Product Search:

(a) Substitutable Product Recommendation (b) Visually Similar Product Recommendation (c) Near-identical Product Search



[3] Product matching platform (*Amazon*)

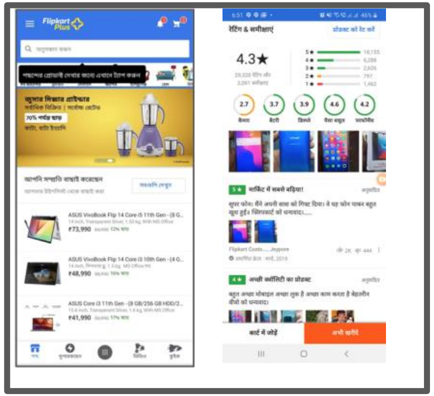
- Duplicate and near-duplicate product detection (and elimination)
- Textual and multi-modal similarity techniques
- Has applications in search, recommendation, fake detection

[4] Knowledge Graph/Concept Nets (*Alibaba*)

- Semantic gap between user-intent and product-catalog in e-commerce search
- Development and adoption of knowledge graphs are on the rise
- Tasks include attribute and relation extraction, KG embeddings, evaluation, KG alignment and merging

[3] A Flexible Large-Scale Similar Product Identification System in E-commerce;Zuo et al; 2020 [4] AliCoCo: Alibaba E-commerce Cognitive Concept Net; Luo et al; 2020

User-generated Content: *the rise of 3Vs*



Flipkart Plus

रेटिंग & समीक्षाएं

प्रोडक्ट को रेट करें

4.3★

5★ 18,155
4★ 6,288
3★ 2,626
2★ 797
1★ 1,462

2.7 कैमरा · 3.7 बैटरी · 3.9 डिस्प्ले · 4.6 पैसा वसूल · 4.2 परफॉर्मेंस

5★ मार्केट में सबसे बढ़िया!

सुपर फोन। मैंने अपनी सास को गिफ्ट दिया। वे यह फोन पाकर बहुत खुश हुईं। फ्लिपकार्ट को धन्यवाद।....

Flipkart Custo..., Jeypore

4★ अच्छी क्वॉलिटी का प्रोडक्ट

बहुत अच्छा मोबाइल अच्छा लुक है अच्छा काम करता है फ्लिपकार्ट जी को धन्यवाद।

कार्ट में जोड़ें

अभी खरीदें

ASUS VivoBook Flip 14 Core i5 11th Gen - (8 G...

₹73,990

ASUS VivoBook Flip 14 Core i3 10th Gen - (4 G...

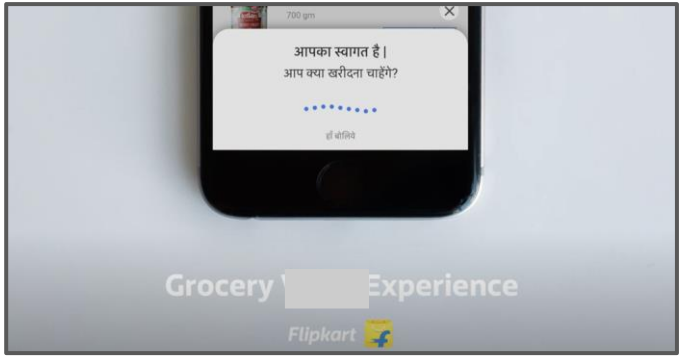
₹48,990

ASUS Core i3 11th Gen - (8 GB/256 GB HDD/2...

₹41,990

Vernacular

- Growing base of smartphone users who are more comfortable in vernacular languages
- Challenges faced:
  - Inability to comprehend English
  - Lack of trust and confidence



आपका स्वागत है।
आप क्या खरीदना चाहेंगे?

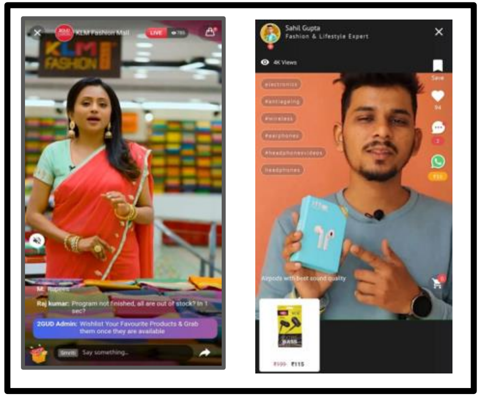
हाँ बोलिये

Grocery [redacted] Experience

Flipkart

Voice

- Voice is increasingly becoming the interface of choice especially with new-to-smartphone population



KLM Fashion Mall

LIVE

Raj kumar: Program not finished, all are out of stock? In 1 sec?

2GUD Admin: Wishlist Your Favourite Products & Grab them once they are available

Say something...

Sahil Gupta
Fashion & Lifestyle Expert

4K Views

#electronics
#antiageing
#wireless
#earphones
#headphonesvideos
#headphones

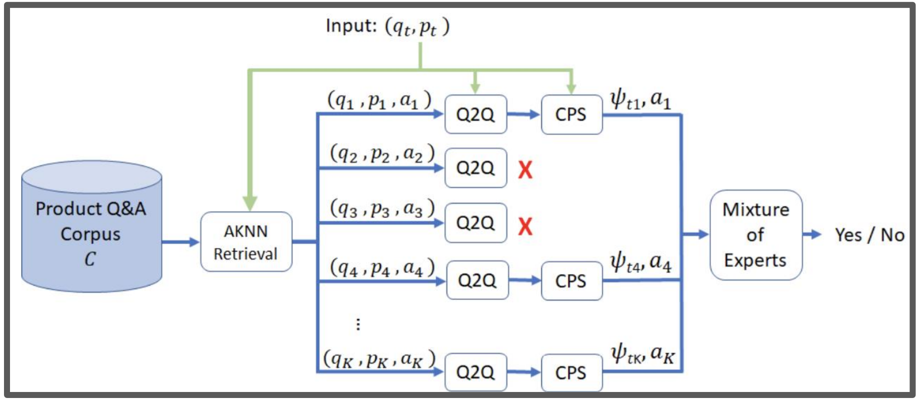
Airpods with best sound quality

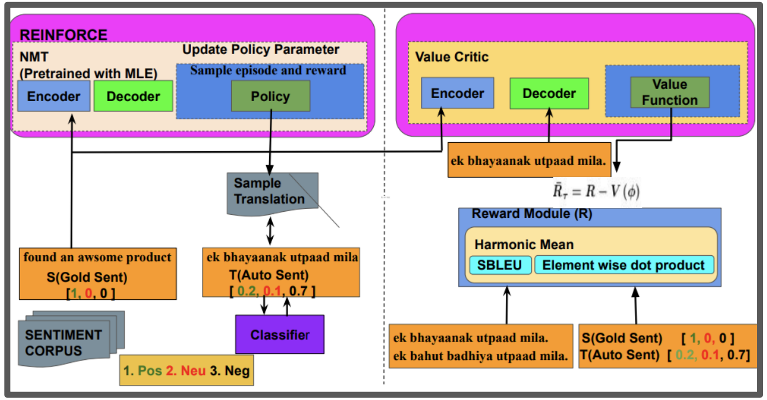
₹199 ₹115

Video

- Shoppable-videos offer an immersive and interactive experience
  - Product shoots
  - Mini-influencers led demonstrations
  - Livestreaming with celebrities

Samples of Recent Research





[1] Answering Questions Utilizing Product Similarity (*Amazon*)

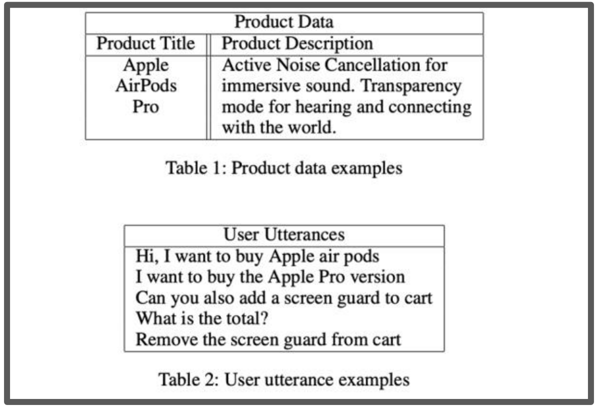
- Automatic answering questions about products leveraging prior QAs from contextually similar products
- Helpful for new products, products with less number of reviews
- Predict answer using a Mixture-of-expert framework to aggregate the answers from contextually similar products

[2] Sentiment-preserving Review Translation (*Flipkart*)

- NMT may lack from preserving stylistic and pragmatic properties of text
  - More prominent for not well-structured text e.g. product reviews
- Deep RL framework to fine-tune the parameters of a NMT system
  - Encoding underlying sentiment as well as without compromising the adequacy

[1] Answering Product-Questions by Utilizing Questions from Other Contextually Similar Products ; Rozen et al; 2021 [2] Sentiment Preservation in Review Translation using Curriculum-based Re-inforcement Framework; Kumari et al; 2021

Samples of Recent Research



Product Data	
Product Title	Product Description
Apple AirPods Pro	Active Noise Cancellation for immersive sound. Transparency mode for hearing and connecting with the world.

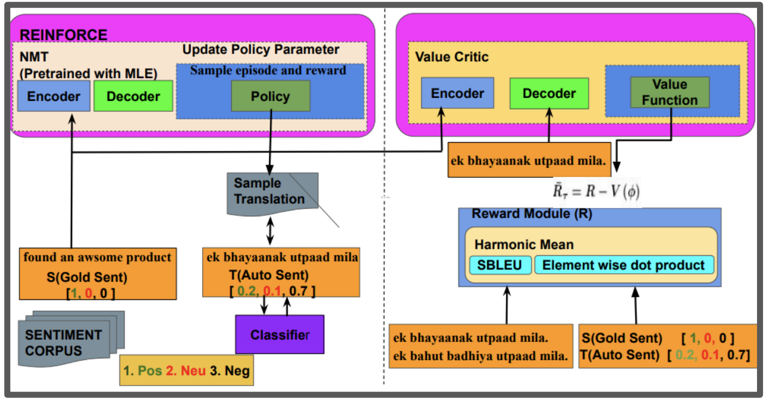
Table 1: Product data examples

User Utterances
Hi, I want to buy Apple air pods
I want to buy the Apple Pro version
Can you also add a screen guard to cart
What is the total?
Remove the screen guard from cart

Table 2: User utterance examples

[3] Retrained Distilled BERT for Shopping Assistant (*Walmart*)

- Retrained distilled BERT for Retail domain
  - Product titles, descriptions etc. and chat logs
- Showed improvement on downstream tasks such as intent detection, sequence tagging etc.

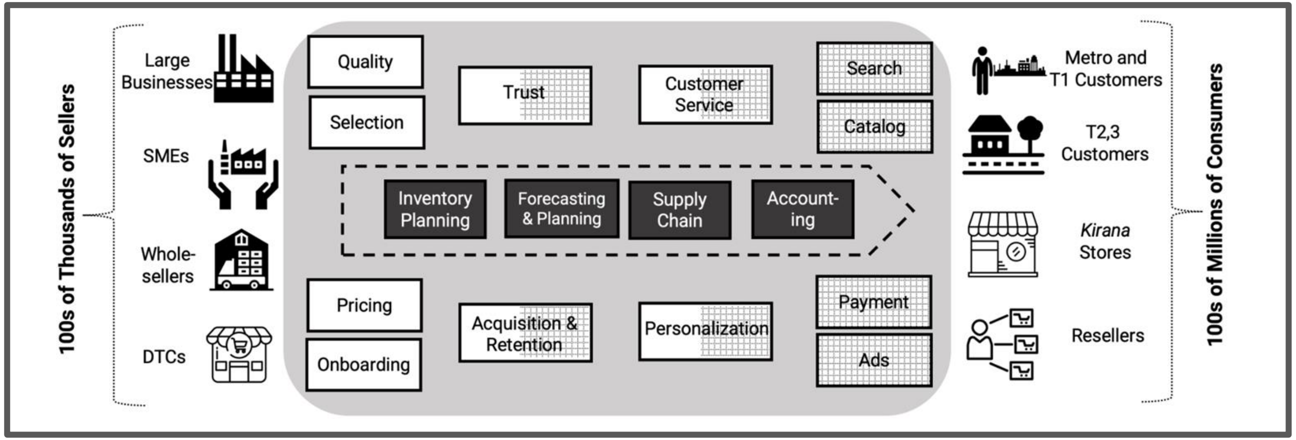


[2] Sentiment-preserving Review Translation (*Flipkart*)

- NMT may lack from preserving stylistic and pragmatic properties of text
  - More prominent for not well-structured text e.g. product reviews
- Deep RL framework to fine-tune the parameters of a NMT system
  - Encoding underlying sentiment as well as without compromising the adequacy

[3] Retraining DistilBERT for a Voice Shopping Assistant by Using Universal Dependencies; Jayarao and Sharma; 2021

Summary



- E-commerce is poised for big(ger) growth in the coming years with data and technology can play strong enabling roles
- The domain is rich in data and richer in problem statements
- Advancements in various (sub)-fields of AI/ML/DS have made significant breakthroughs and more to come