letgo

Building a 1500-Class Listing Categorizer from Implicit User Feedback

Arnau Tibau - *arnau.tibau@letgo.com*

October 2019, Data Council Barcelona

Arnau

Luca

Julien

Philipp

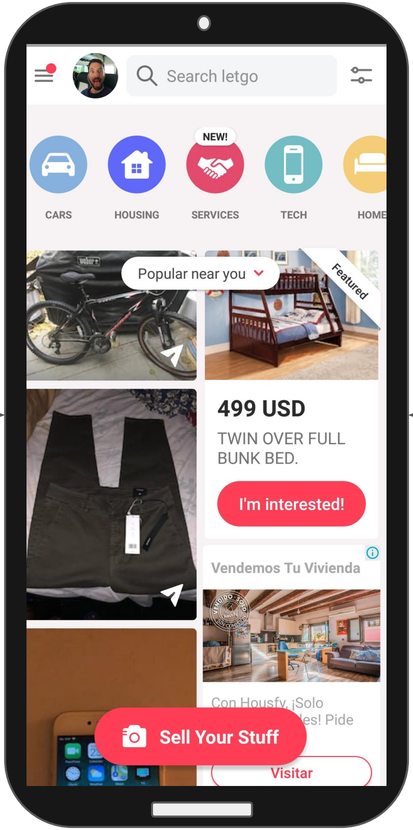
Antoine

Outline

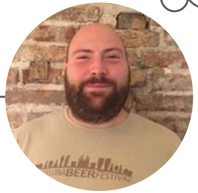
1. **The problem of listing categorization**
2. Building a listing categorization dataset
3. Building a listing taxonomy
4. Image-based listing categorization
5. Conclusions & next steps

A local two-sided marketplace - where sellers sell & buyers buy

Arnau
Hoboken,
New York



“bike”



Luca
Jersey City,
New York

A correct categorization is key for a good buyer experience

Moderation

Categorization

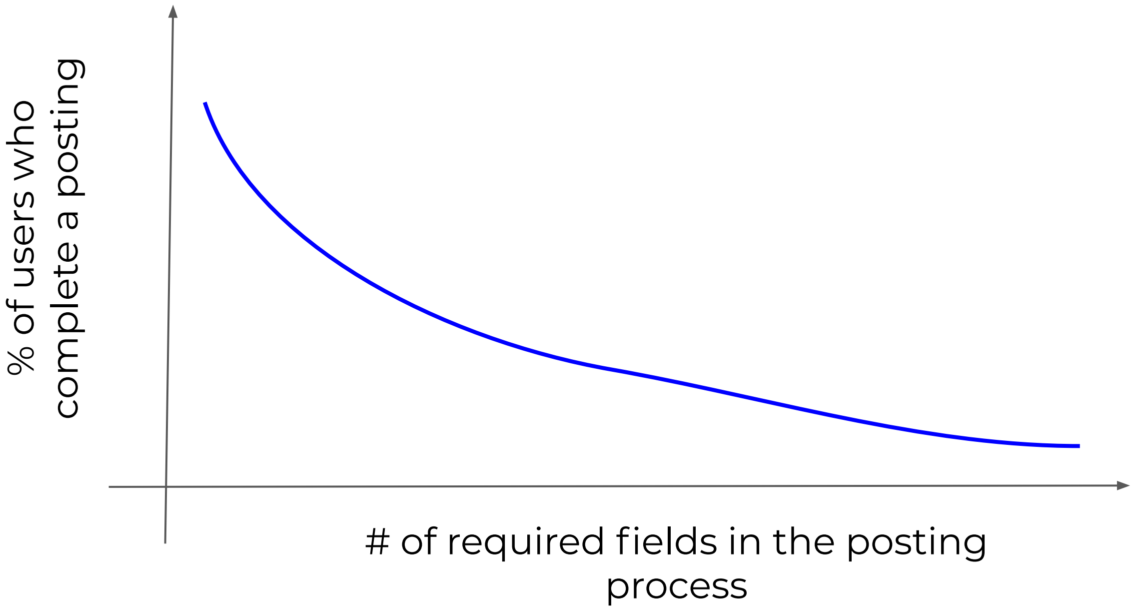
....

- Recommender System
- Search
- Filter
- Related listings

- Category: Sports > Bicycles > Road bike
- Illegal Content: No
- Brand: Fuji
- ...

Listing Catalog

Why don't you just ask sellers for the category?



Why don’t you just ask sellers for the category? (II)

Adversarial Sellers

Is this an “iphone”?

Posting mistakes

This is not a “Car”

Problem statement

Title: Joovy twin Roo+ Stroller

Description: Best stroller for infant twins in my opinion. Fits two infant car seats side by side. You pick the [...]

Price: $50

→ Listing categorizer → Baby & Child > Strollers > Twin stroller

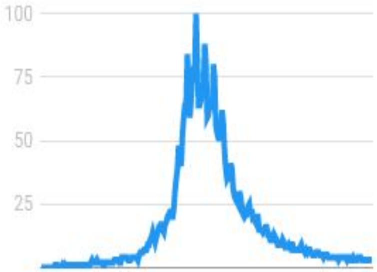
It's *just* a classical "supervised learning" setup

Challenges

Ambiguity	Noisy Data	Open-world	Adversarial Actors

A TV, a TV stand or a TV + TV Stand?

Does this look like a “Cardigan”?



“Fidget spinners” rise and demise

This is not an “iphone”

Outline

1. The problem of listing categorization
2. **Building a listing categorization dataset**
3. Building a listing taxonomy
4. Image-based listing categorization
5. Conclusions & next steps

To train a model we need data and labeled data is expensive

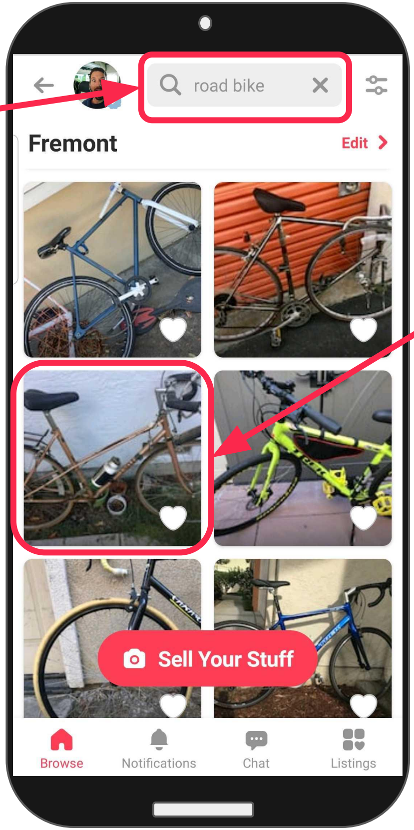
1500 categories x 1000 samples / category = 1.5M labeled samples

	$ / annotation (estimate)	Total Estimate*
AWS Groundtruth (internal workforce)	0.02 (aws) + 0.06 (labor)	$120k
AWS Groundtruth (Mechanical Turk)	0.02 (aws) + 0.02 (labor)	$60k
Google Data Labeling	? (google) + 0.025 (labor)	-

With a recommended replication factor of 3x, we're talking **$180k - $360k** for a moderately sized, static dataset.

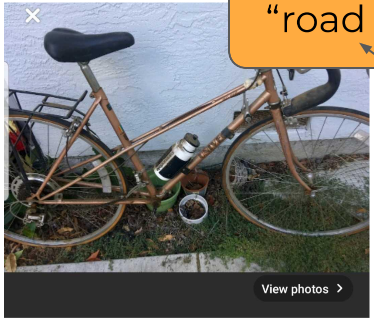
Implicit user feedback provides noisy labels

search query



Contacted listing (feedback)

Implicit user feedback provides noisy labels (II)



“road bike” (3)

$80.19

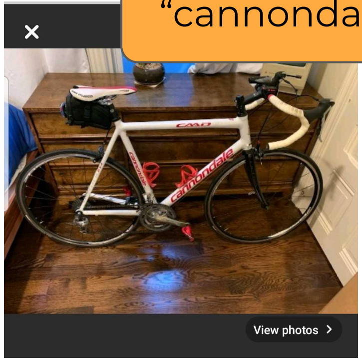
Univega mixte road bike project

Does not include the bell

label (search query)

consensus (# of users who agreed”)

“cannondale” (4)

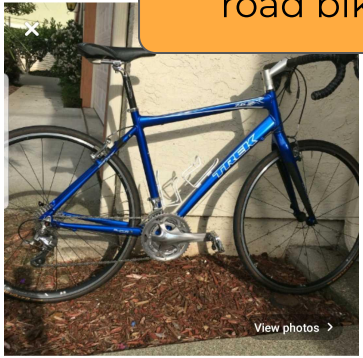


$1,350

Cannondale CAAD9 6 - 58 cm

Cannondale CAAD9 6 - 58 cm

“road bike” (1)



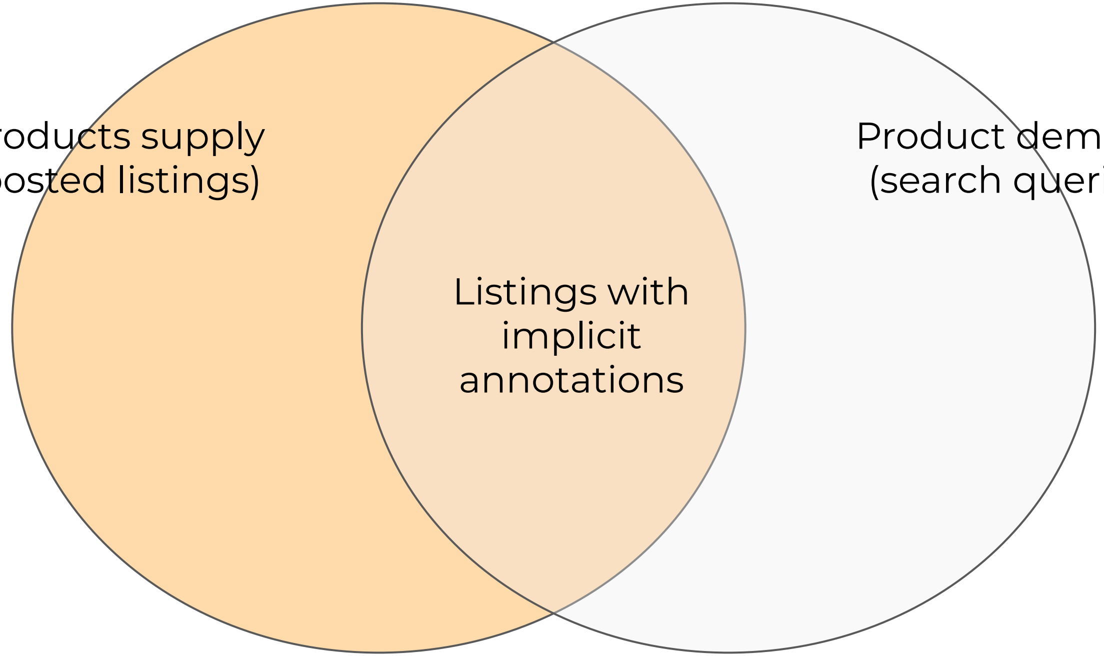
$500

Blue trek road bike

2009 Trek FX 7.4 Hybrid
20” frame
Shimano 105 5700 group set
Excellent condition

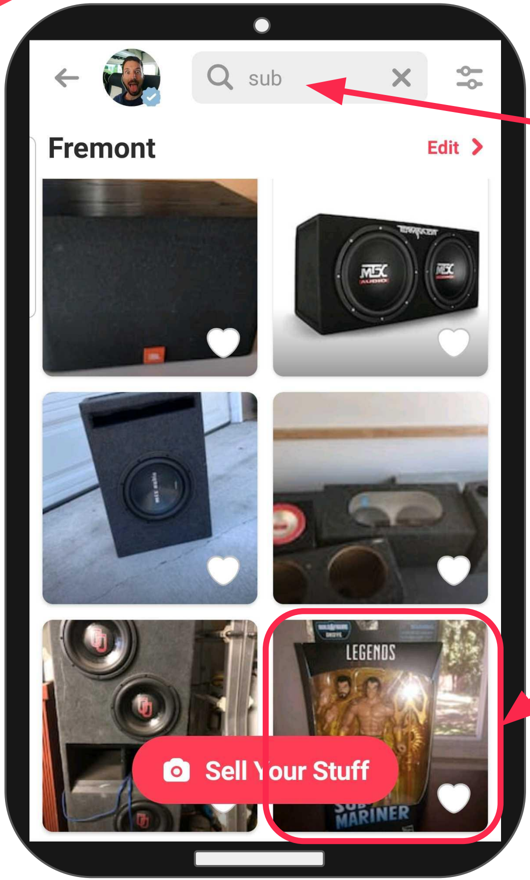
We have 10s of millions of **different** search queries a month

Challenges: Potential **Selection Bias**



Potential problems ahead if P(class | w/ annotations) ≠ P(class)

Challenges: **Uncorrelated label noise**

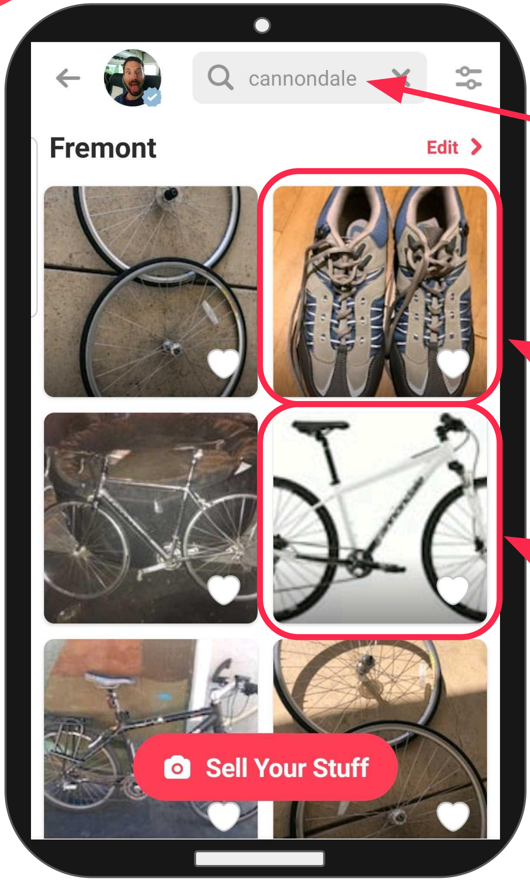


Query: “sub”
(for sub-woofer)

Contacted:
“Sub-mariner
action figure”

- “Uncorrelated” because label error depends *weakly* on the listing attributes
- Sources of this type of label noise:
  - Curiosity
  - Mistakes
  - ....
- **“Easy” to deal with through outlier detection**

Challenges: **Correlated label noise**



Query: “cannondale bike”

Contacted: “cannondale shoes”

Contacted: “cannondale mountain bike”

- “Correlated” because label error depends *strongly* on the listing attributes
- Sources of this type of label noise:
  - Listing similarity
  - User interest correlation
  - Taxonomy errors
- **Hard to deal with**

Outline

1. The problem of listing categorization
2. Building a listing categorization dataset
3. **Building a listing taxonomy**
4. Image-based listing categorization
5. Conclusions & next steps

What is a listing taxonomy?

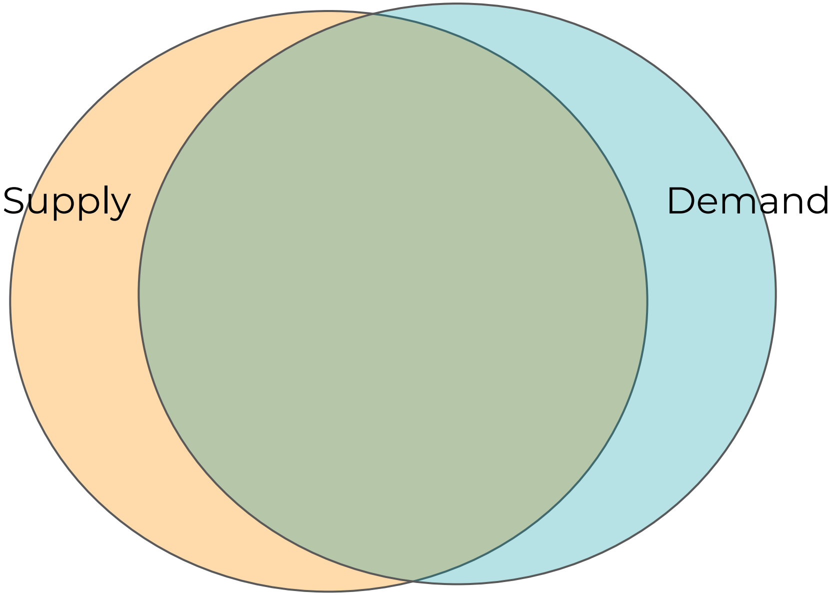
- adirondack
- bean bag
- bench
- camper chair
- commode chair
- cuddle chair
- dining chair
- feeding chair
- gaming chair
- **glider chair**
- highchair
- lounge chair
- massage chair
- office armchair
- office chair
- ottoman
- papasan chair
- parson chair
- ...

- armoire
- bed
- bed frame
- board
- cabinet
- couch
- **chair**
- chest
- desk
- dinning set
- ...

- cars
- fashion & accessories
- **furniture**
- ...

The leafs of the taxonomy are the classes in our classifier...
... and the tree structure encodes the relationship between labels

What makes a good marketplace taxonomy?

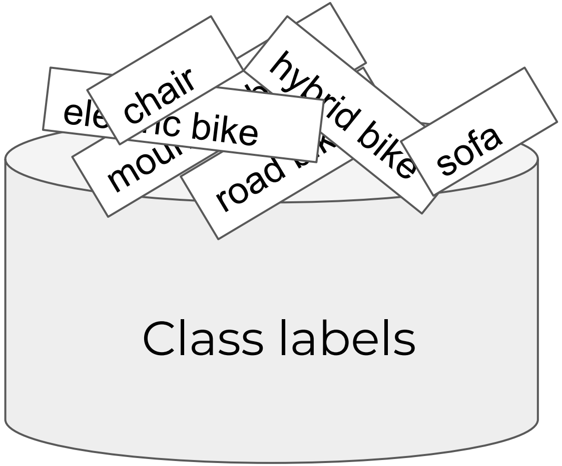


Good coverage of both supply and demand

	Electronics	Cars & Trucks	Other Auto & Parts	Home & Applian...	Tools & Gardening	Sports & Outdoors	Toys & Gaming	Movies, Books & ...	Baby & Child	Fashion & Acces...	Pet Supplies	Health & Beauty
Phones	32											
Cameras	31											
Trucks		32										
Motorcycle		4	25			1						
Furniture				32								
Home appliances				32								
Home Improvement				27	5							
Decor				31								
Plants				2	30							
Lawnmower			3	1	26	2						
Sports gear						32						
Bikes						31	1					
Fitness equipment				1		28						3
Boat		1	9			19						
Video games	5						26	1				
Video consoles	10			1			20	1				
Toys & Board games							32					
Collectibles				1			3					
Music								32				
Books								32				
Musical instruments								21				
Baby Toys							4		28			
Baby Clothing									32			
Clothing									1	31		
Jewelry										31		1
Bags & Handbags										32		
Footwear						1				31		
Pet food											32	
Cosmetics										1		31

Minimizes user confusion while offering sufficient granularity

Building a taxonomy is typically a manual process



road bike

tool chest

vanity tops

wall cabinets

welsh cabinets

adirondack chairs

hybrid bike

bmx bike

baseball card

business card

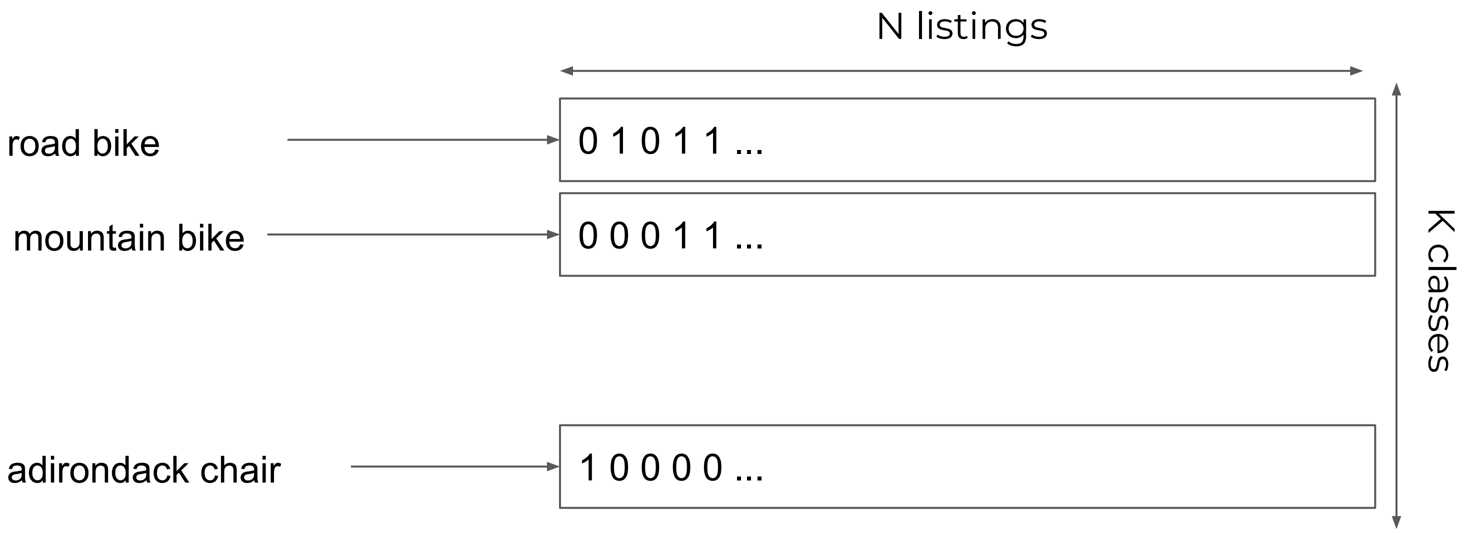
calling card

card collection

Bikes → road bike, hybrid bike, bmx bike, ….

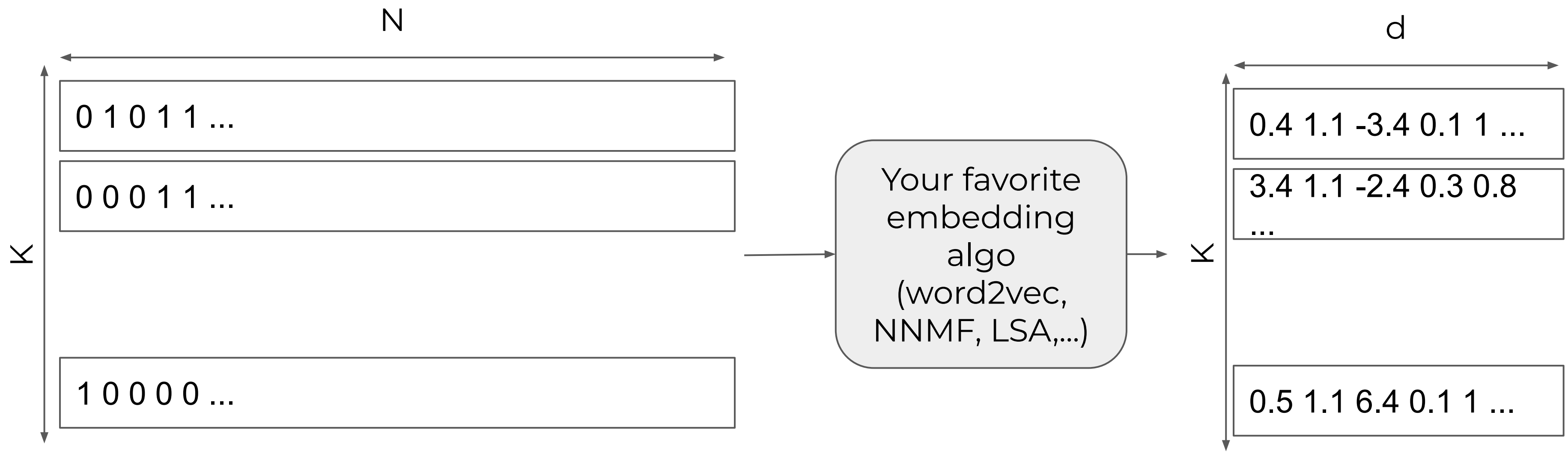
This is a very time-consuming process (O(k^2))… can we help speed it up?

A data-driven process to define large taxonomies



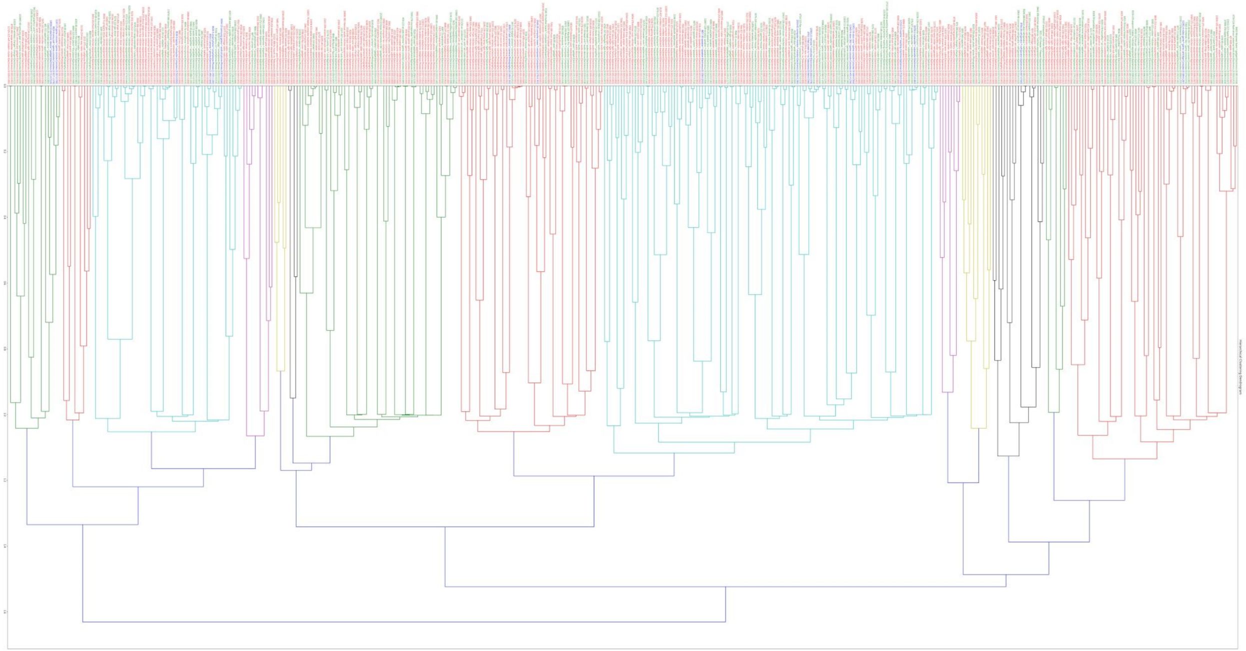
First we map potential classes to a vector space (K << N)

A data-driven process to define large taxonomies (II)



We “embed” these vectors into a lower-dimensional space

A data-driven process to define large taxonomies (II)

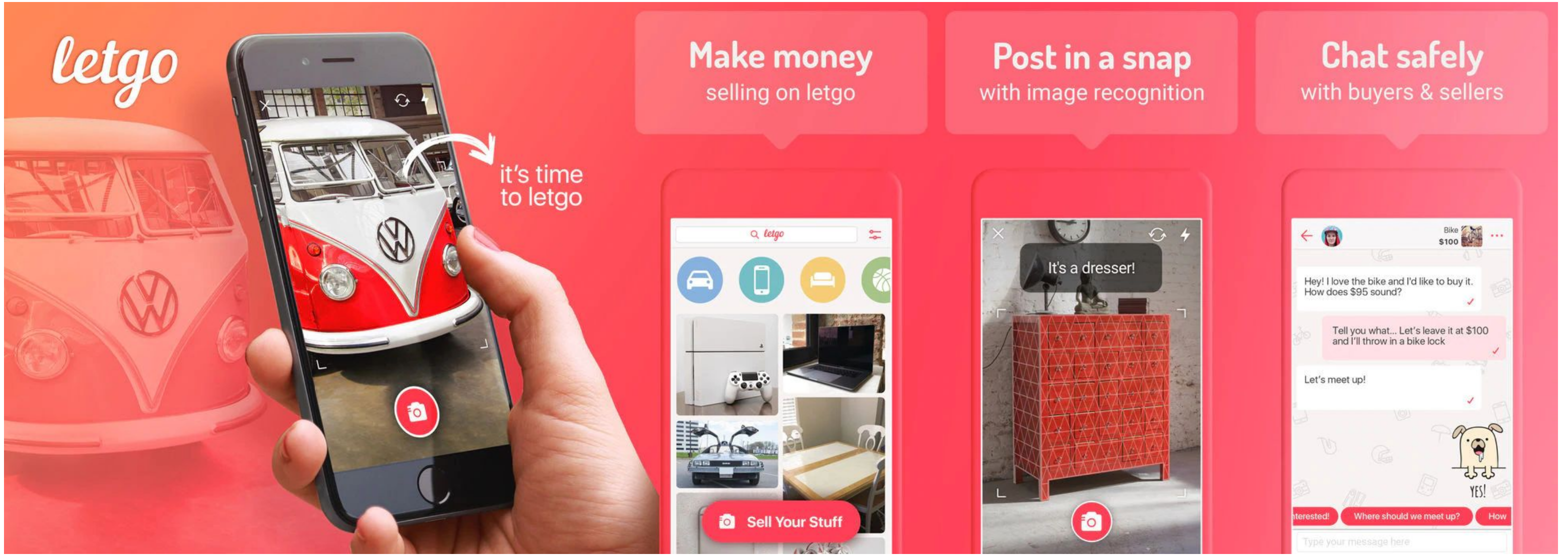


Finally we cluster those class-vectors
and inspect manually
(C clusters instead of K classes)

Outline

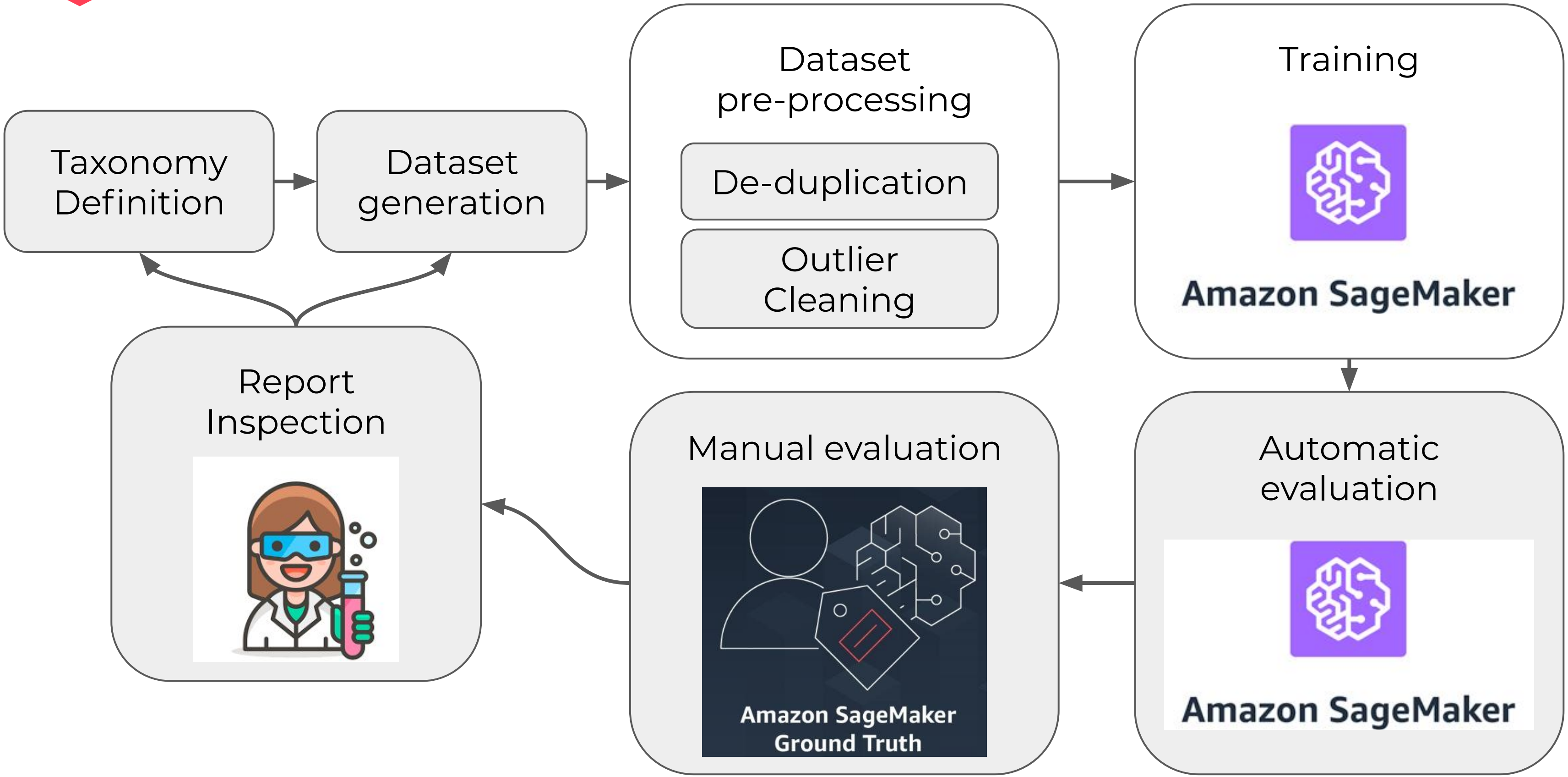
1. The problem of listing categorization
2. Building a listing categorization dataset
3. Building a listing taxonomy
4. **Image-based listing categorization**
5. Conclusions & next steps

Why image-based listing categorization?

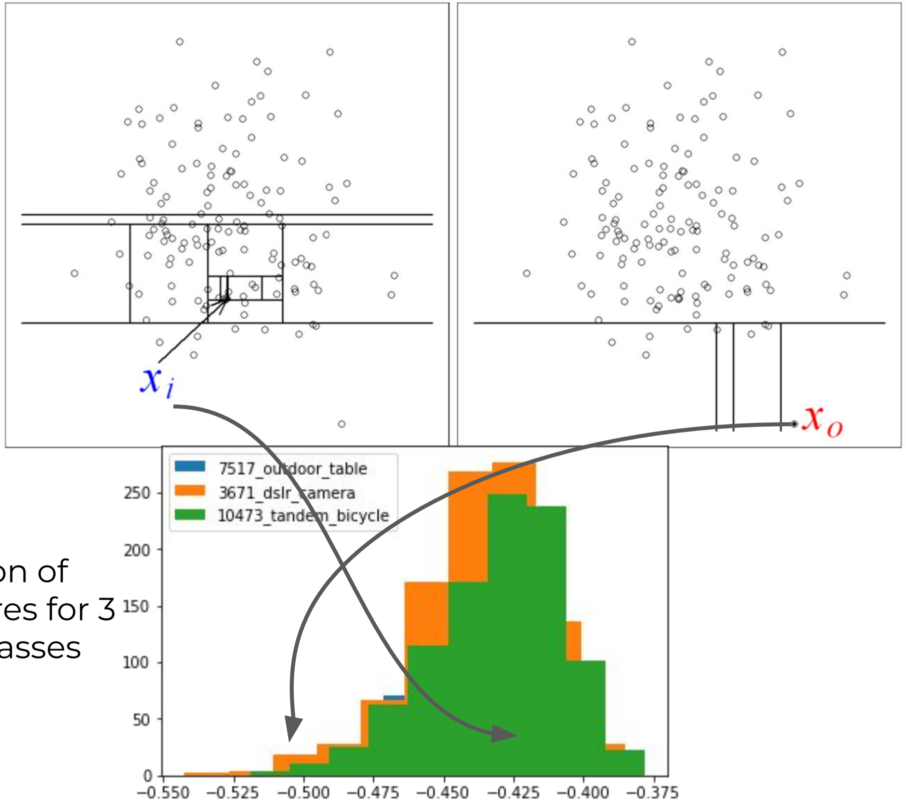


- At first, we only asked for one image at posting-time*
- Very few listings had a description
- Even if now we have more info (name, description, price), the picture and the other info should be coherent.

A virtuous training & debugging cycle

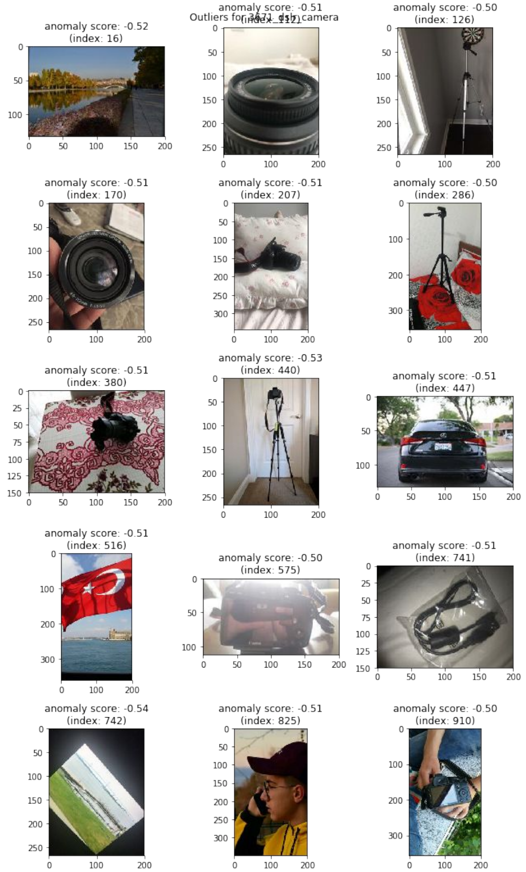


Uncorrelated label noise cleaning via Isolation Forests



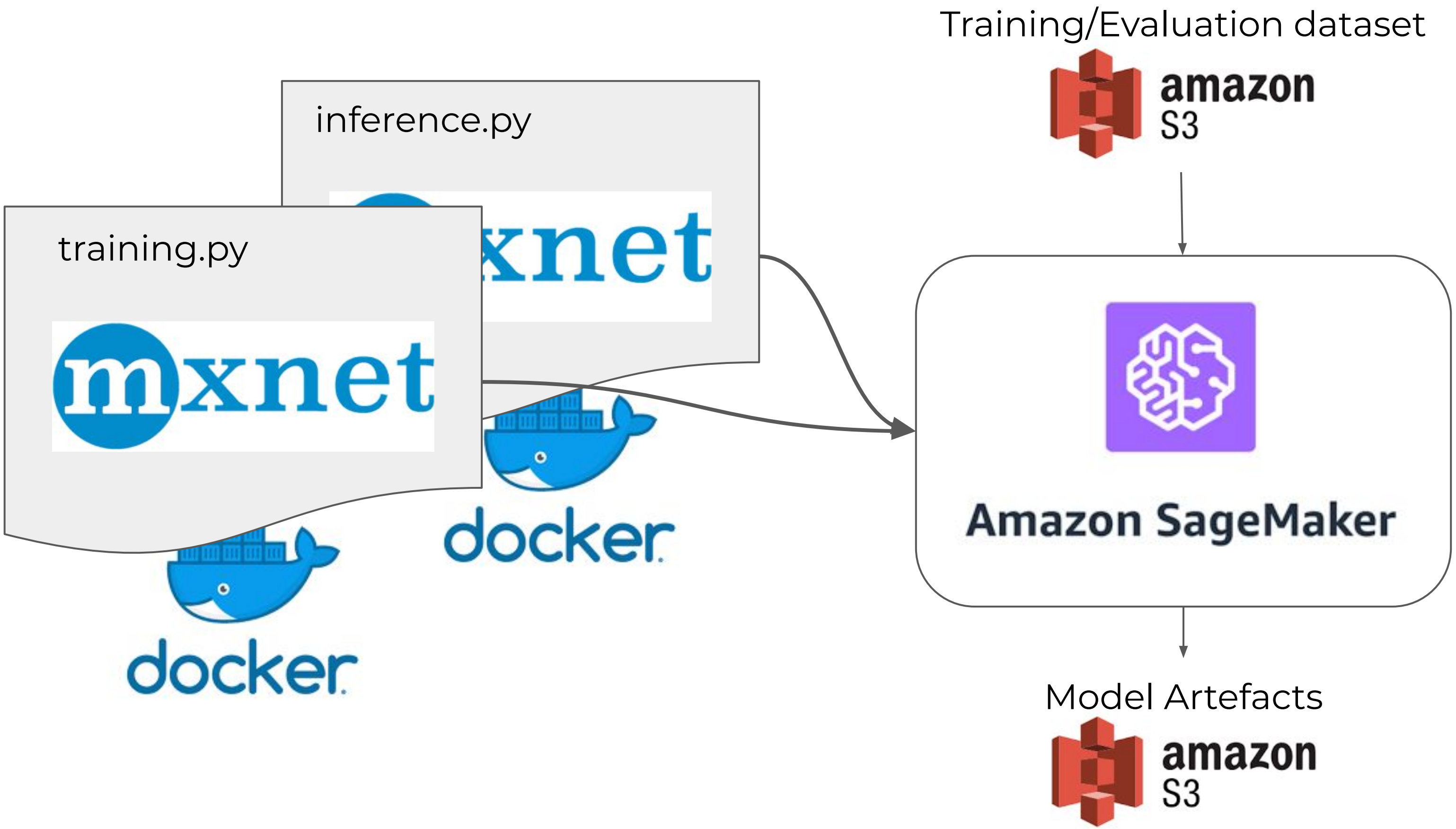
Distribution of anomaly scores for 3 different classes

Outliers for "dslr camera"



[Isolation-based Anomaly Detection, Liu, Ting and Zhou, ACM 2012](#)

Training & deploying with AWS Sagemaker’s



Training & deploying with AWS Sagemaker's (II)

Pros	Cons
• **Ease-of-use** • **Flexibility** (DL framework, model ser/deserialization, data transformation) • Endpoint **autoscaling** • Integration with **AWS ecosystem** (IAM roles, Cloudwatch, ELB, etc) • Fast & responsive **maintenance**	• **Debugging** • Inference **Cost** ($$$) • **Unstable** SDK • Integration **Testing** is hard

Model & dataset debugging (I)

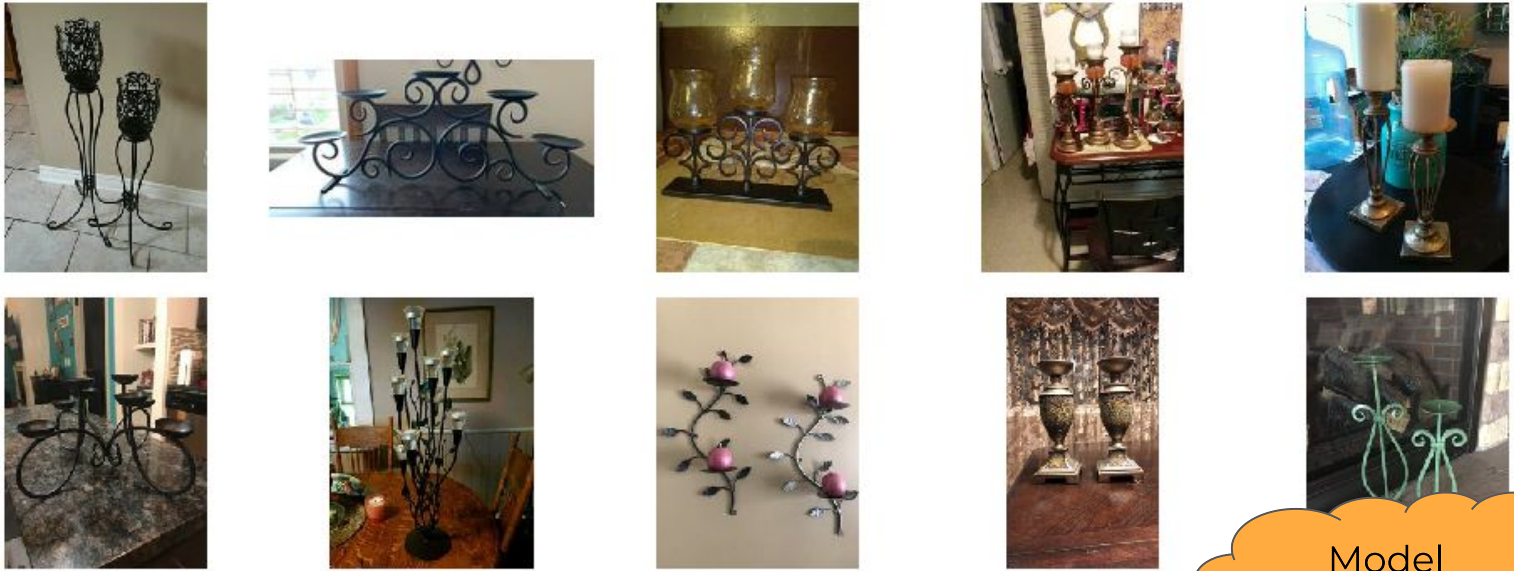
Global metrics	Precision	Recall	Accuracy
Category			
Subcategory			
Micro-cat			

Local metrics	Precision	Recall	f1-score
6901_masquerade_mask	95%	97%	96%
11381_vinyl_figure	92%	94%	93%
11361_video_game	89%	96%	93%
....			

Metrics are great to measure progress but not that great for fixing problems:
Look at the mistakes!

Model & dataset debugging (II): confusion report

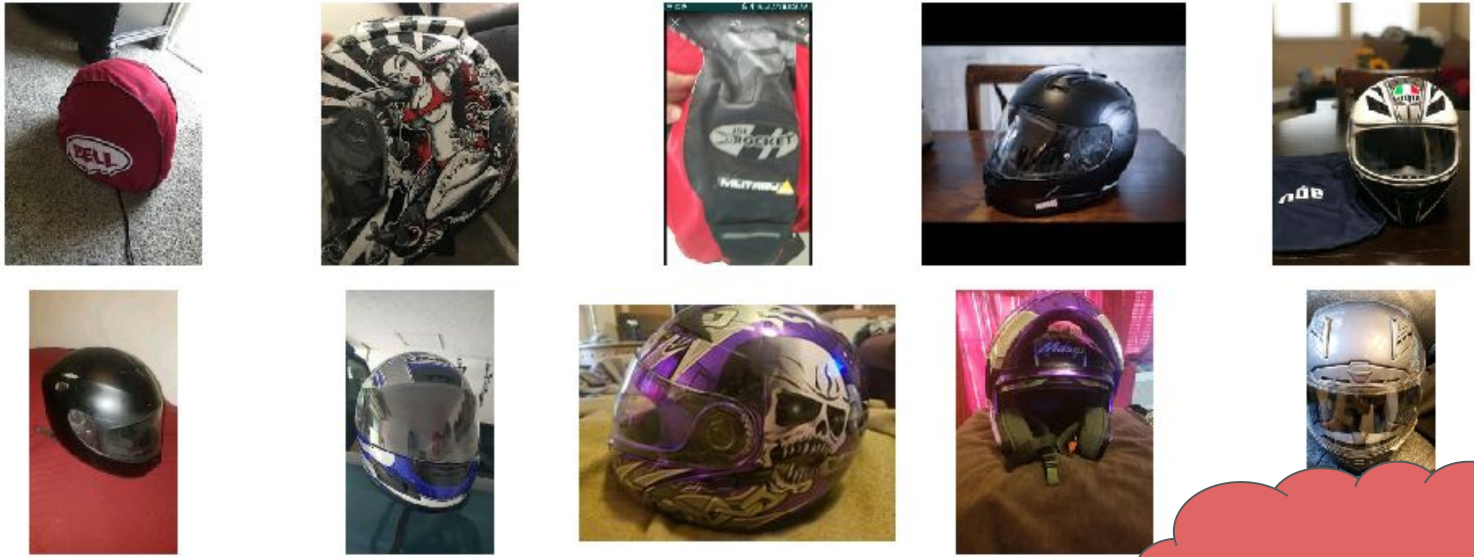
1848_candle_stand->1839_candle_holder (16.00% - 16)



Model capacity

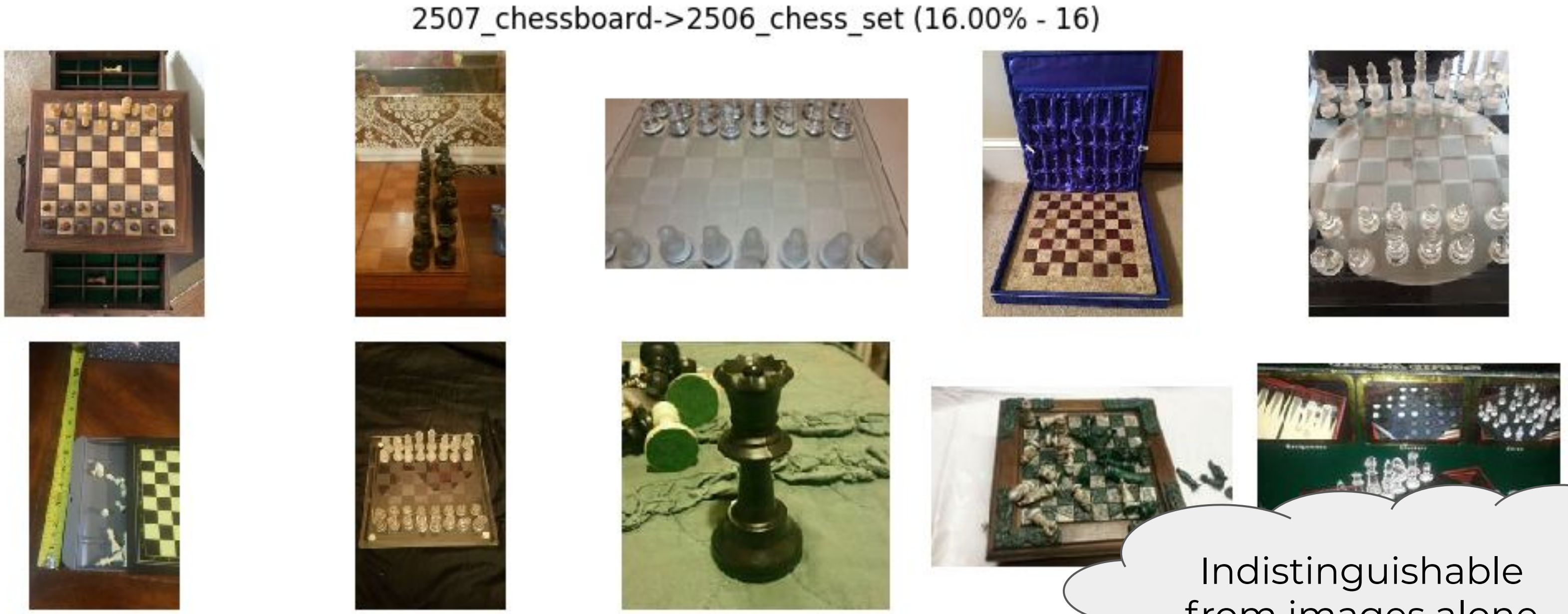
Model & dataset debugging (III): confusion report

7209_motorcycle_tire->7192_motorcycle_helmet (13.00% - 13)



Wrong label

Model & dataset debugging (IV): confusion report



Indistinguishable
from images alone

Other potential issues:

- Low sample size
- Non-disjoint classes

Manual evaluation

- Noisy labels means we can't trust our accuracy metrics
- Customer perception ≠ metrics
- Gathering human labels → EASY PEASY?

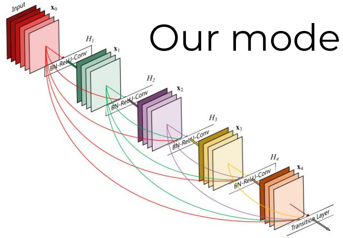


Title: Joovy twin Roo+ Stroller

Description: Best stroller for infant twins in my opinion. Fits two infant car seats side by side. You pick the [...]

Price: $50

Our model



Manual annotations

Manual evaluation challenges (II)

- What do we ask for? right/wrong, choose correct one?
- What do we show? All images, text, category tree?

Same predictions, different category tree → 10% difference in manually evaluated accuracy

Outline

1. The problem of listing categorization
2. Building a listing taxonomy
3. Building a listing categorization dataset
4. Image-based listing categorization
5. **Conclusions & next steps**

Conclusions & next steps

- Implicit feedback is a **cost-effective** way of annotating data
- Performance **evaluation** is not trivial
- Building pipelines with **iteration** in mind is KEY
- **AWS Sagemaker** provides a good trade-off between flexibility and ease of use

Next steps:

- **Resolving ambiguities** → multi-image and multi-modal models
- **Handling correlated label noise** → Noise-aware models

letgo

Thanks for your attention!

Any questions?

arnau.tibau@letgo.com

We're hiring!

Check out our openings here:

we.letgo.com/careers