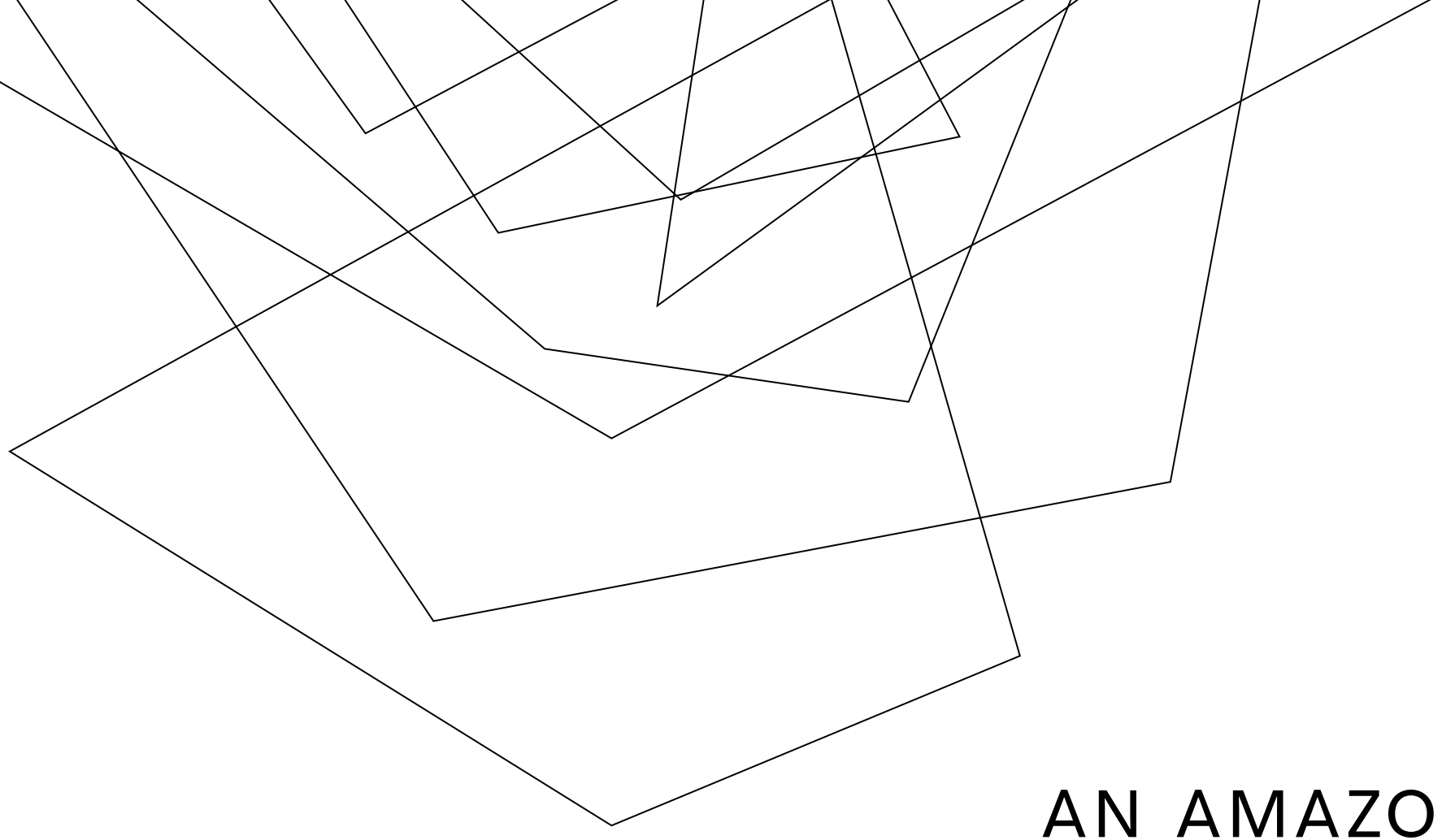


AN AMAZON PRODUCT CLASSIFIER

Nick Gigliotti

BUSINESS PROBLEM

Amazon has asked me to build a product classifier for two purposes:

1) Integrating new products

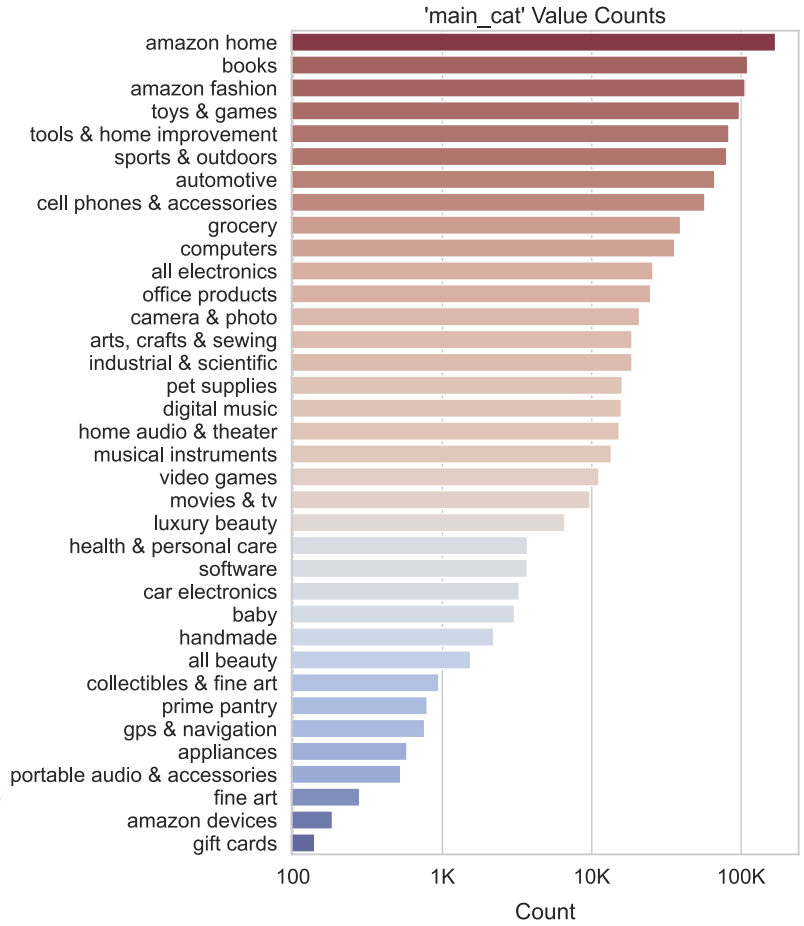
2) Flagging misclassified products

They would like some recommendations pertaining to the product classifier and its uses.

The dataset was originally gathered by three AI researchers, Jianmo Ni, Jiacheng Li, and Julian McAuley, for their 2019 paper, **"Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects."**

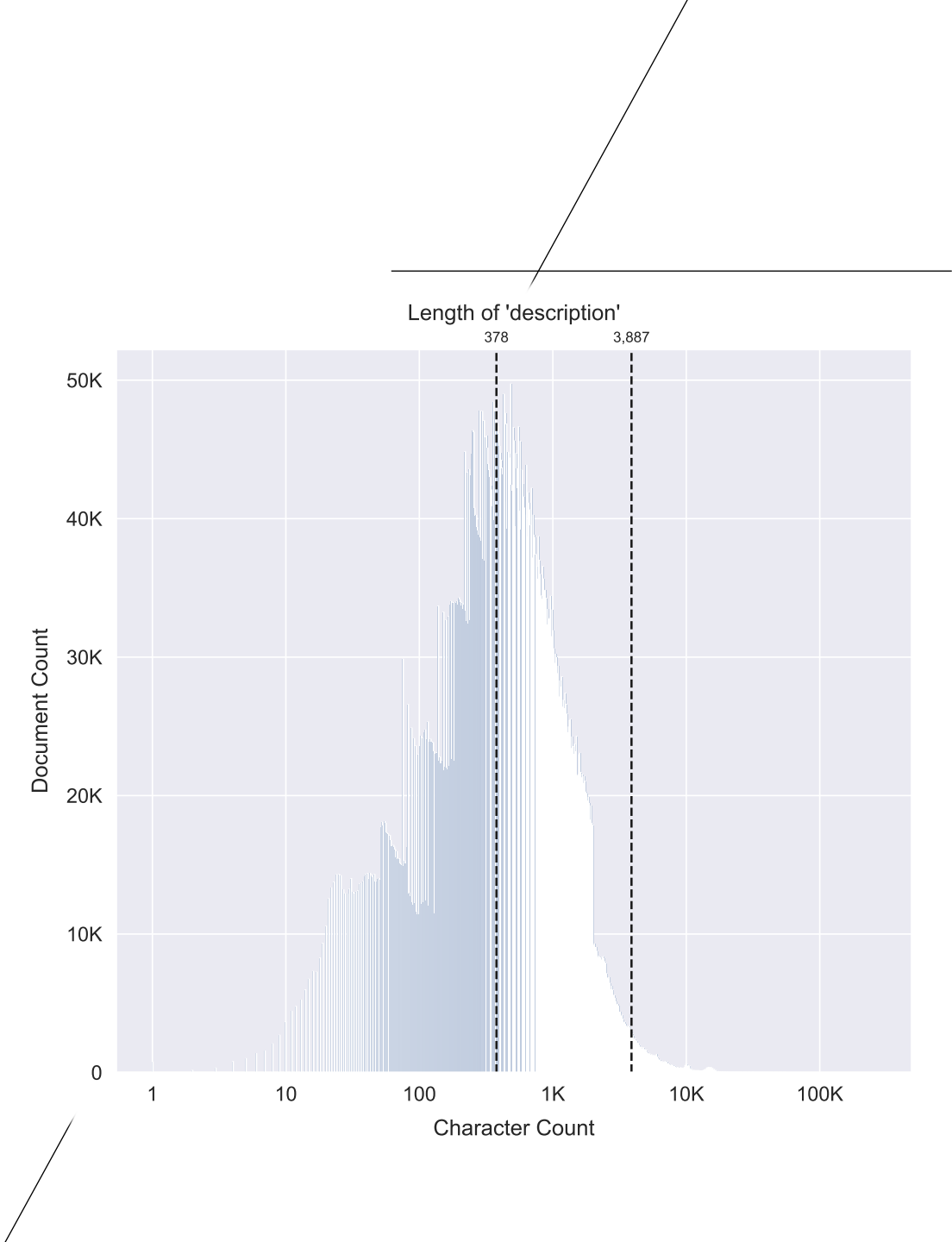
DATASET

- Originally contains ~15M Amazon products (~13GB)
  - Reduced to ~1M by extensive scrubbing and engineering
  - 36 classes, imbalanced
- Includes title, description, brand, feature, and category
- Probably scraped (contains HTML tags)
- Released/updated in 2018



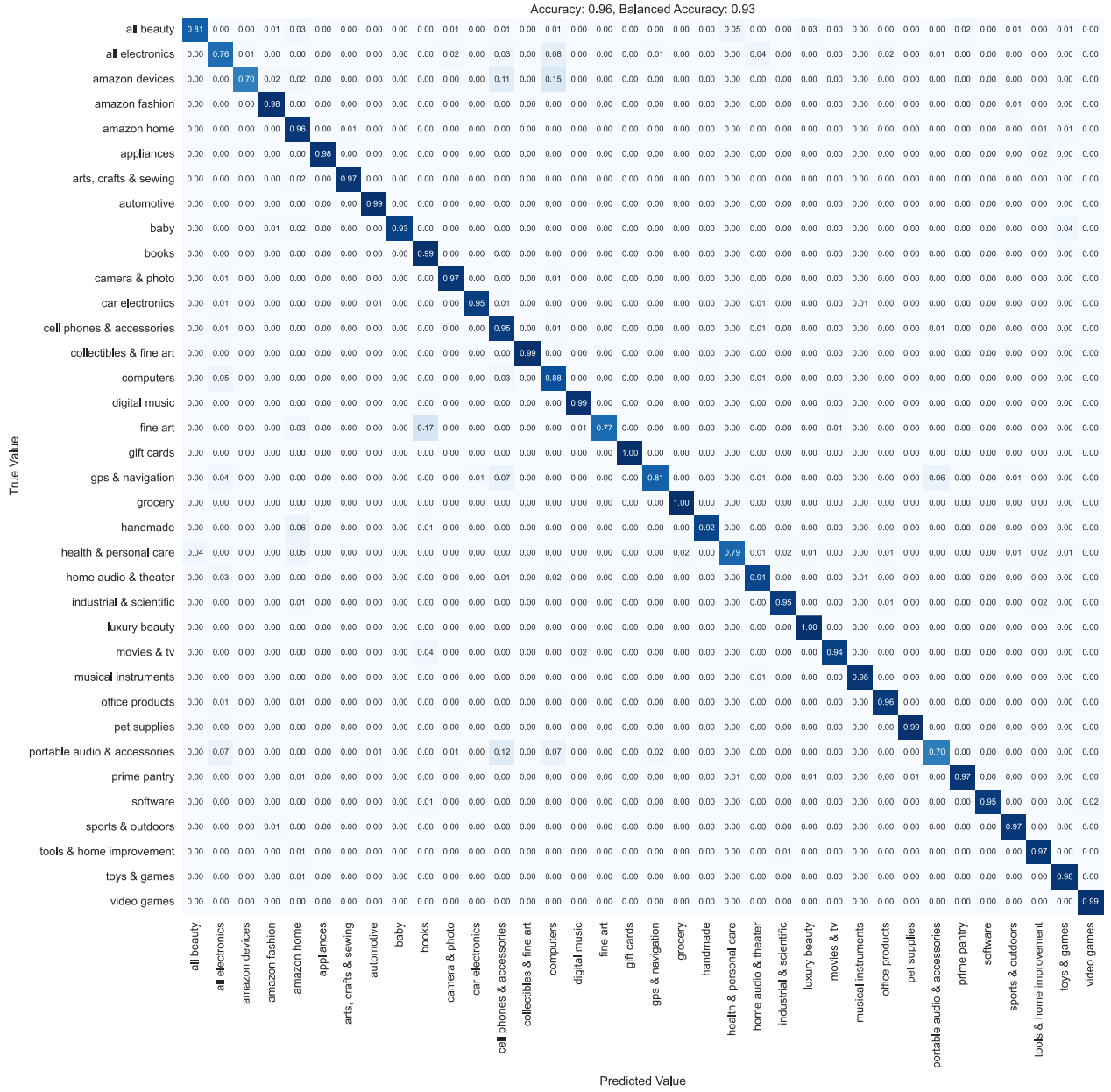
METHODS

- Curate corpus for model training:
  - Drop shortest 50% of descriptions
  - Drop extremely long outliers
  - Select only **top brands** in each category
- Rigorous preprocessing:
  - Filter out non-ASCII, numerals, punctuation
  - Filter out stop words, repetitive sequences
  - Restrict token length (2-16 characters)
- Feature engineering:
  - Multi-word brand terms
  - Extract bigrams (2-word phrases) from each category
- Binary TF*IDF vectorization
  - Binary occurrence markers {0, 1}
  - IDF weighting to emphasize rare terms
  - Normalize to reduce effect of document length
- Support Vector Machine with SGD
  - Efficient on large datasets with 1M or more documents
  - Normalize to reduce effect of document length



FINAL MODEL

- Correct 96% of the Time
  - Does well on small classes
  - Best class recall is 100%
  - Worst class recall is 70%
- Understandable Mistakes
  - No egregious errors
  - Portable Audio → Cell Phones
  - Fine Art → Books
  - Amazon Devices → Computers, Cell Phones
- Perfectionistic Learning
  - Learning rate is “adaptive”
  - Whenever quality of fit plateaus, learning continues at 1/5 of the rate

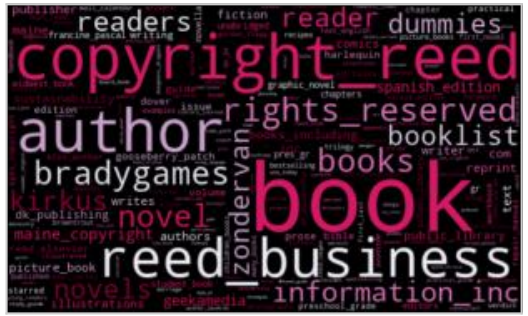


Fine Print

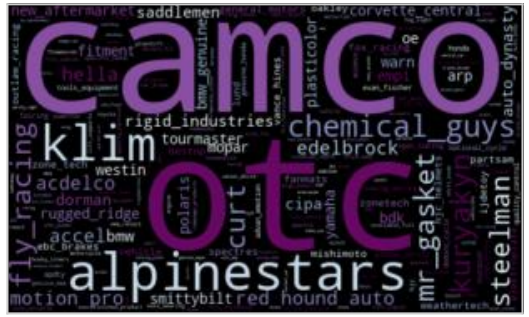
Highest F_1-Scores

Grocery

Books

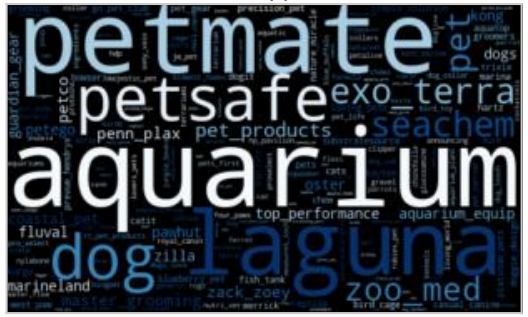


Automotive



Luxury Beauty

Pet Supplies



Amazon Fashion

Video Games

Toys & Games

Digital Music

Brands

Mixture

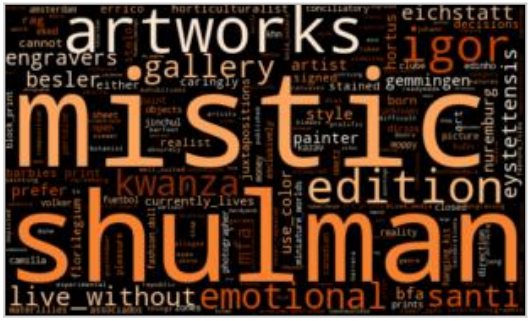
Lowest F_1-Scores

Amazon Devices

Portable Audio & Accessories

Gps & Navigation

Fine Art



All Electronics

Health & Personal Care

All Beauty

Appliances

Home Audio & Theater

- Top terms are mostly brand names
- Still look good
- Not too different from the top scores

CONCLUSION

- **If you're looking to classify products with NLP, lead with the brand terms.**

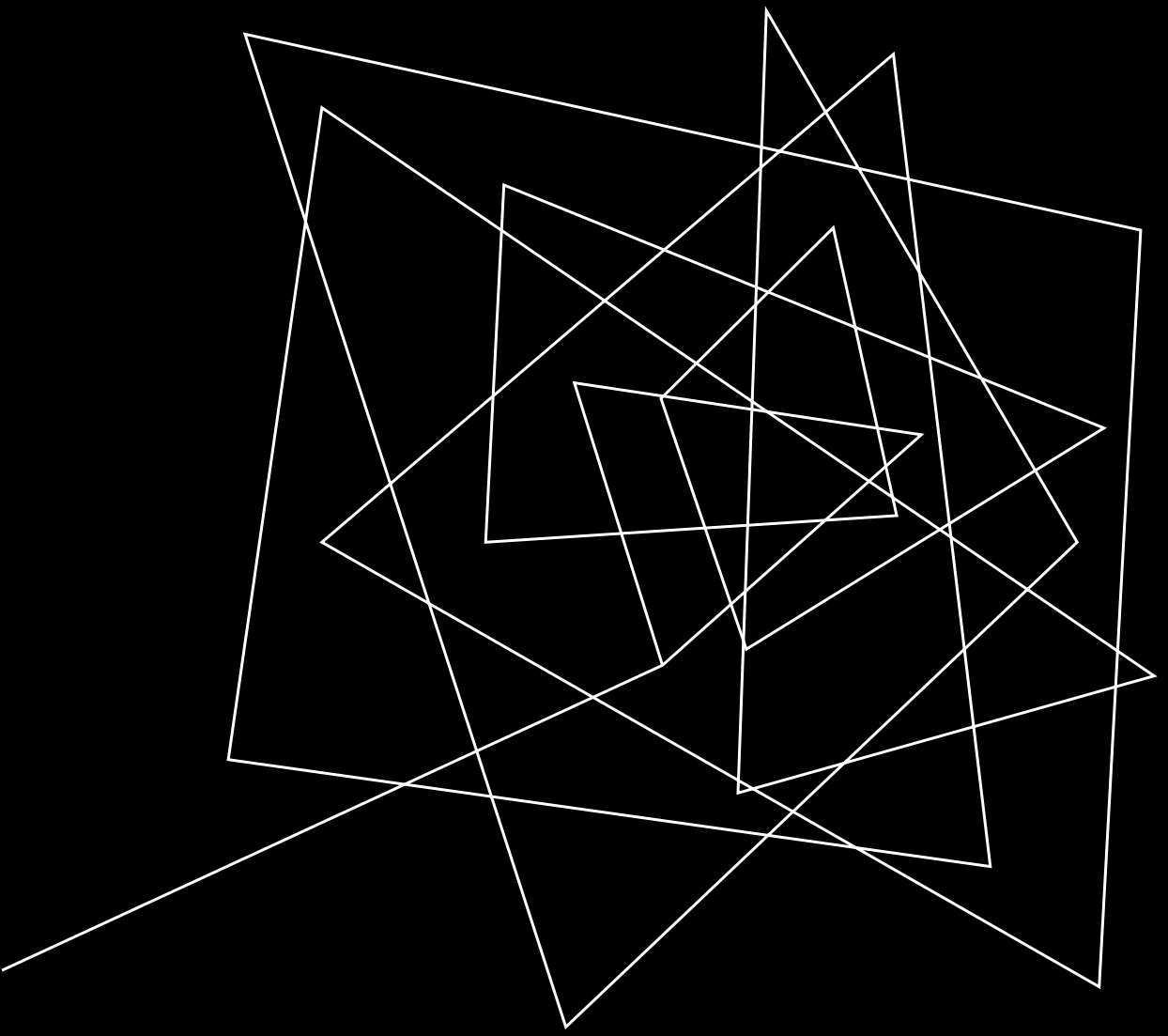
  Brand terms ranked high in nearly every category. One could build a decent model with *only* brand terms, though I wouldn't recommend going that far. Even if you wanted an image-based classifier, brands are the first place I'd start.

- **Don’t ignore boilerplate legalistic text, because sometimes it’s category-specific.**

  In fact, I recommend you gather up all the legalistic caveats and copyright statements you can get. This text is sometimes very distinctive of its category.

- **Use the model to study your competitors and scope out new suppliers.**

  This model can be used to analyze other business' inventories, including those of competitors. Discover new products and suppliers by directly comparing their inventories to yours under your classification scheme.



LOOKING FORWARD

- Gather data on brands concerning their relationships and parent companies.

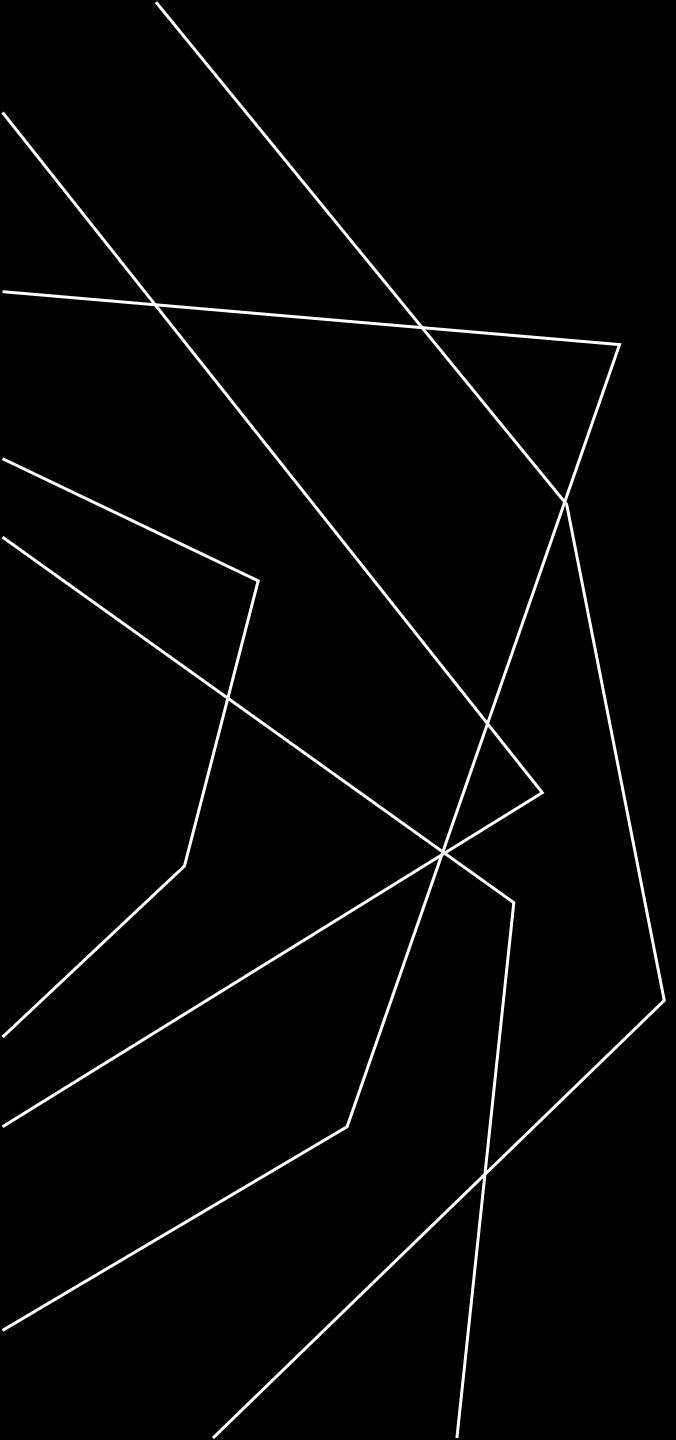
    - Try to expand the model's coverage to more obscure brands.

- Develop a workflow to create specialized subcategory models for each major category.

    - These will be **multilabel** classification models.

- Create a dashboard to demonstrate the accuracy and rich interpretability of the model.

- Obtain a new, unseen dataset to test the model's generalizability.



THANKS

Nick Gigliotti

github.com/ndgigliotti