November 2022
IFLA Webinar

Autocategorization Projects: A Taxonomist's Perspective

Bob Kasenchak, Factor
bob.kasenchak@factorfirm.com

Outline

- Introduction
- Types of Autocategorization (Classification of Autoclassification methodologies)
- Roles and collaboration
- Defining success
- Questions

WHO AM I?

BOB KASENCHAK

Information Architect
& Taxonomist

I am a taxonomist with an interest in ontolgies and Linked Data. I have worked for over a decade building and implementing taxonomy and auto-classification projects for publishing, enterprise, technology, and e-commerce clients.

Factor is an information architecture and human experience consultancy focused on the challenge of bringing user-centered design principles and practice to enterprise-scale information problems.

Contact
bob.kasenchak@factorfirm.com
@taxobob
www.factorfirm.com

factor

Auto-classification

Autoclassification

UF: Auto-tagging, auto-categorization, autocat, text classification

Automated (or semi-automated) methodologies for applying tags to content.

In addition to Subject tags, this can also include other tags (entities etc.)

May or may not be from formal taxonomies/vocabularies.

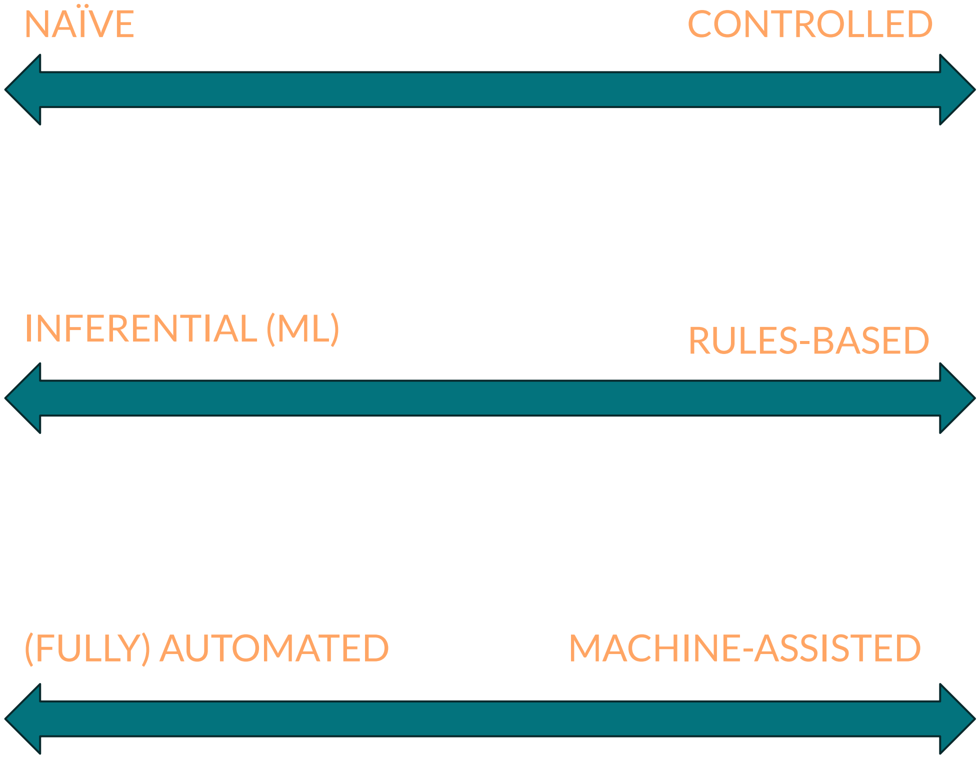
Types of Auto-classification

Types of AUTOCAT

We can think about auto-classification along three axes.

All three axes are used in combination(s), and more than one methodology may be appropriate.

Each approach has pros and cons



NAÏVE

CONTROLLED

INFERENTIAL (ML)

RULES-BASED

(FULLY) AUTOMATED

MACHINE-ASSISTED

NAÏVE VS CONTROLLED

NAÏVE CLASSIFICATION

Sometimes called “concept extraction” and includes Entity Extraction

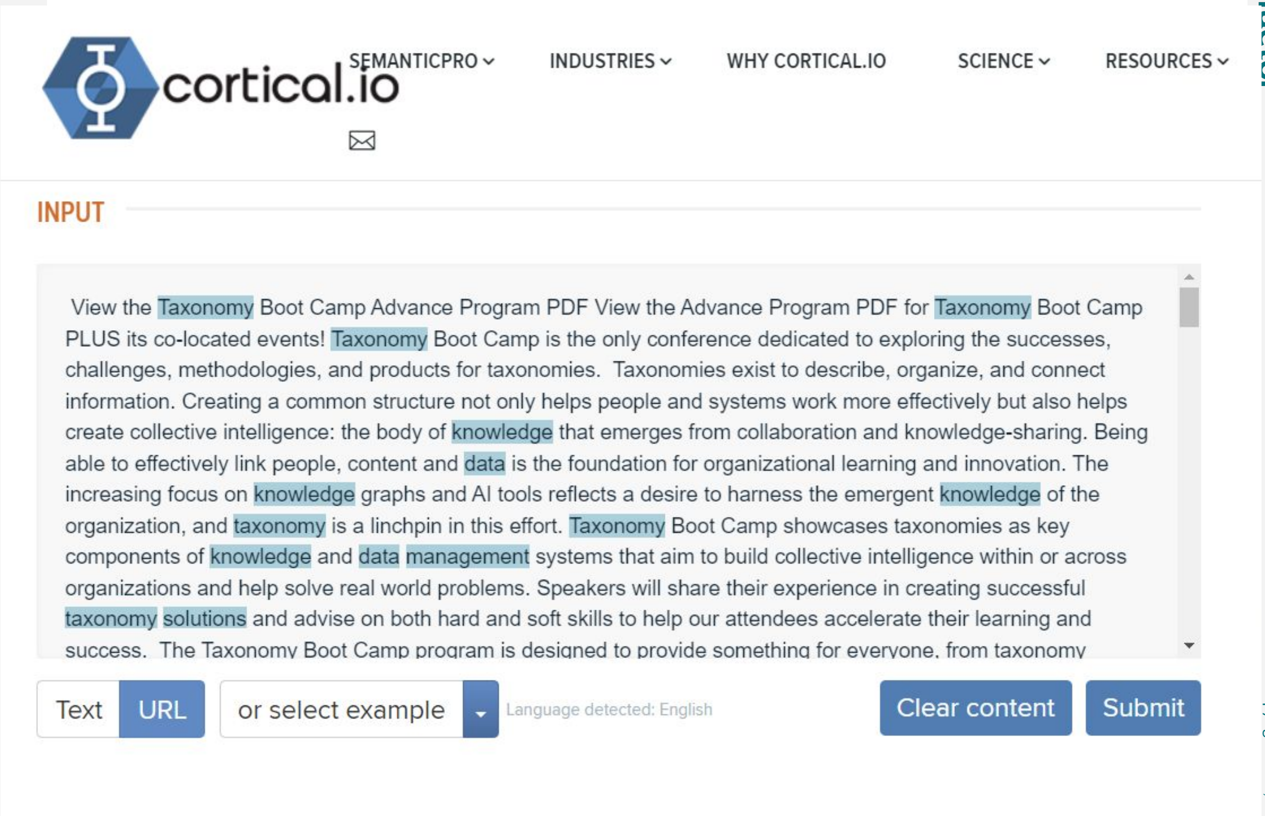
Concepts/entities are identified (using NLP techniques) and extracted from a document without any/much reference to existing lists of specific topics/entities (or very general topic clustering)

CLASSIFYING WITH CONTROLLED VOCABULARIES

One or more semantic structures (taxonomies, thesauri, authority files, ontologies) are in use; classification seeks to match text strings (concepts in a document) to existing lists of topics, entities, etc.

These methods are often combined with Naive Classification (to find gaps in existing vocabularies)

NAÏVE CLASSIFICATION (Entity/Concept Extraction)

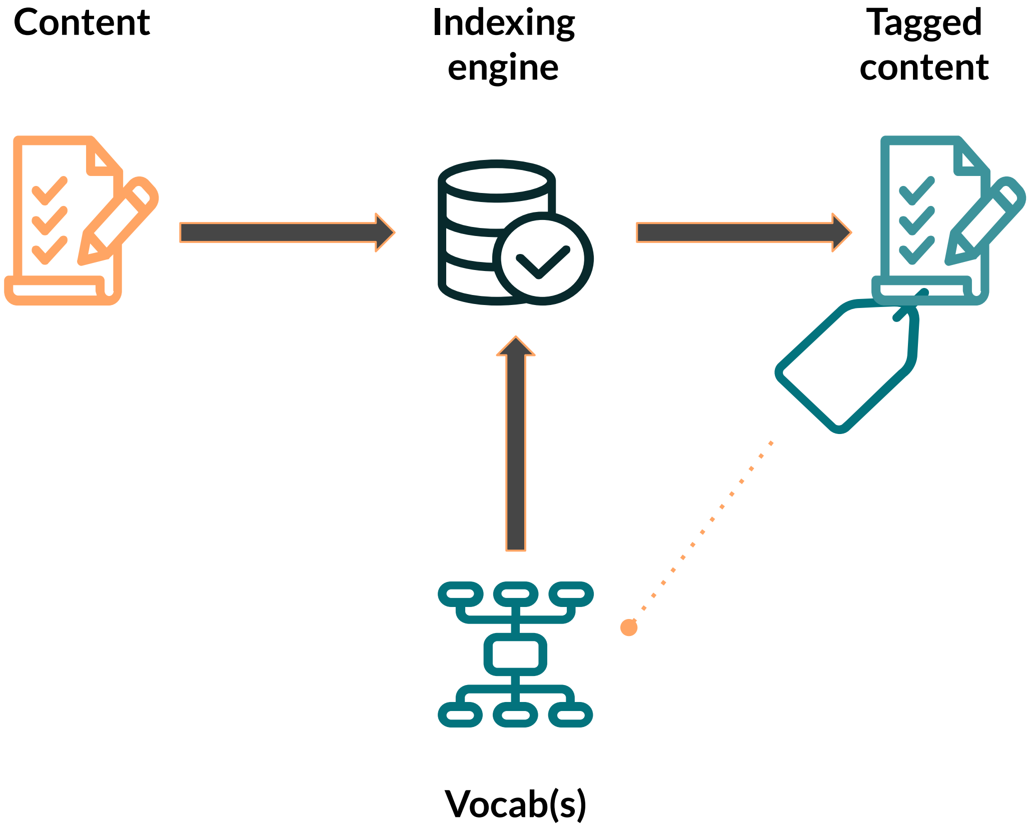


View the Taxonomy Boot Camp Advance Program PDF View the Advance Program PDF for Taxonomy Boot Camp PLUS its co-located events! Taxonomy Boot Camp is the only conference dedicated to exploring the successes, challenges, methodologies, and products for taxonomies. Taxonomies exist to describe, organize, and connect information. Creating a common structure not only helps people and systems work more effectively but also helps create collective intelligence: the body of knowledge that emerges from collaboration and knowledge-sharing. Being able to effectively link people, content and data is the foundation for organizational learning and innovation. The increasing focus on knowledge graphs and AI tools reflects a desire to harness the emergent knowledge of the organization, and taxonomy is a linchpin in this effort. Taxonomy Boot Camp showcases taxonomies as key components of knowledge and data management systems that aim to build collective intelligence within or across organizations and help solve real world problems. Speakers will share their experience in creating successful taxonomy solutions and advise on both hard and soft skills to help our attendees accelerate their learning and success. The Taxonomy Boot Camp program is designed to provide something for everyone, from taxonomy

https://www.cortical.io/freetools/extract-keywords/

factor

CONTROLLED CLASSIFICATION (Using Vocabulary(s))



factor

MACHINE LEARNING (INFERENTIAL) VS RULES-BASED

RULES-BASED METHODS

Humans (with perhaps some light machine assistance) create Boolean-type rules to specify contextual clues for matching text strings to concepts.

This generates a human-readable and -editable set of classification rules which are easy to test and change.

INFERENTIAL METHODS

A pre-tagged set of documents is fed into a system, which will infer a connection between words found in the sample texts to their metadata tags.

This generates some kind of automated process to tag other (new) documents with similar word(s) with the same tags.

Training sets must be substantially large, contain every possible tag (with multiple examples per tag), and be very accurately and specifically tagged.

RULES-BASED CLASSIFICATION

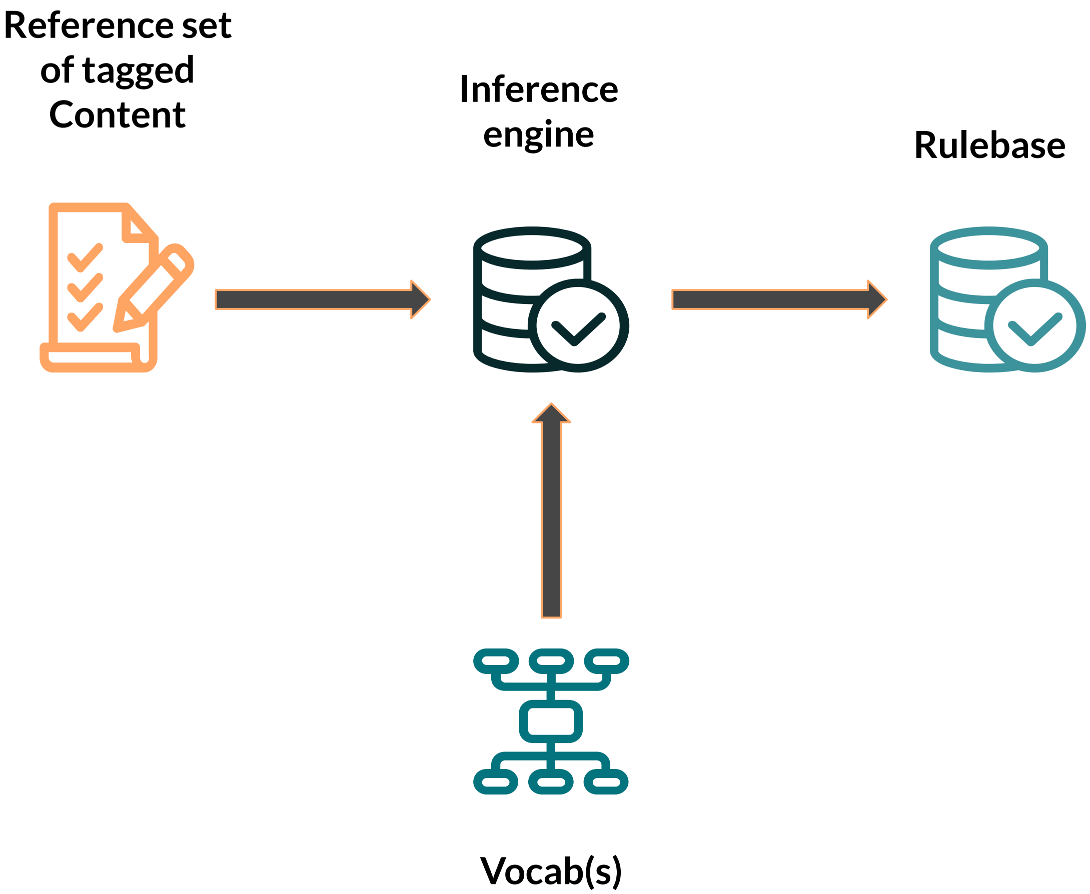
```
mercury

IF [same paragraph] as
    astronom*
    planet*
    orbit*
        TAG Mercury (planet)

ELSE IF [same paragraph] as
    Ford
    Detroit
    automobile
    brand
        TAG Mercury (car)

ELSE
    TAG Mercury (element)
```

INFERENTIAL (ML) CLASSIFICATION



Reference set of tagged Content

Inference engine

Rulebase

Vocab(s)

factor

AUTOMATIC VS MACHINE-ASSISTED

(FULLY) AUTOMATIC CLASSIFICATION

Documents are classified (against one or more vocabs) and tags are automatically applied to content. This method admits/requires spot-checking of applied tags for human-validated QC.

MACHINE-ASSISTED CLASSIFICATION

- Programs which narrow down and/or suggest relevant tags to a human tagger
- May also review human-applied tags to machine-applied tags for QC.
- Speeds up the human-based classification process and increase the accuracy of human tagging.

Faceted Taxonomies, Tag Limits, Weighting, & Hierarchical Tagging

How **many taxonomies** should be used for tagging?

How **many tags** should be applied to a content object?

How many **times** is a concept invoked?

In which **sections** of the document does the concept appear?

What BT-NT **relationships** can be leveraged for accurate retrieval?

Collaboration and Success

factor

PEOPLE, PROCESSES, SYSTEMS

People

Information specialists

Domain Experts

Developers

Content owners

Project Mgmt

Processes

Indexing

Rulebuilding

QC/Review

Governance

Systems

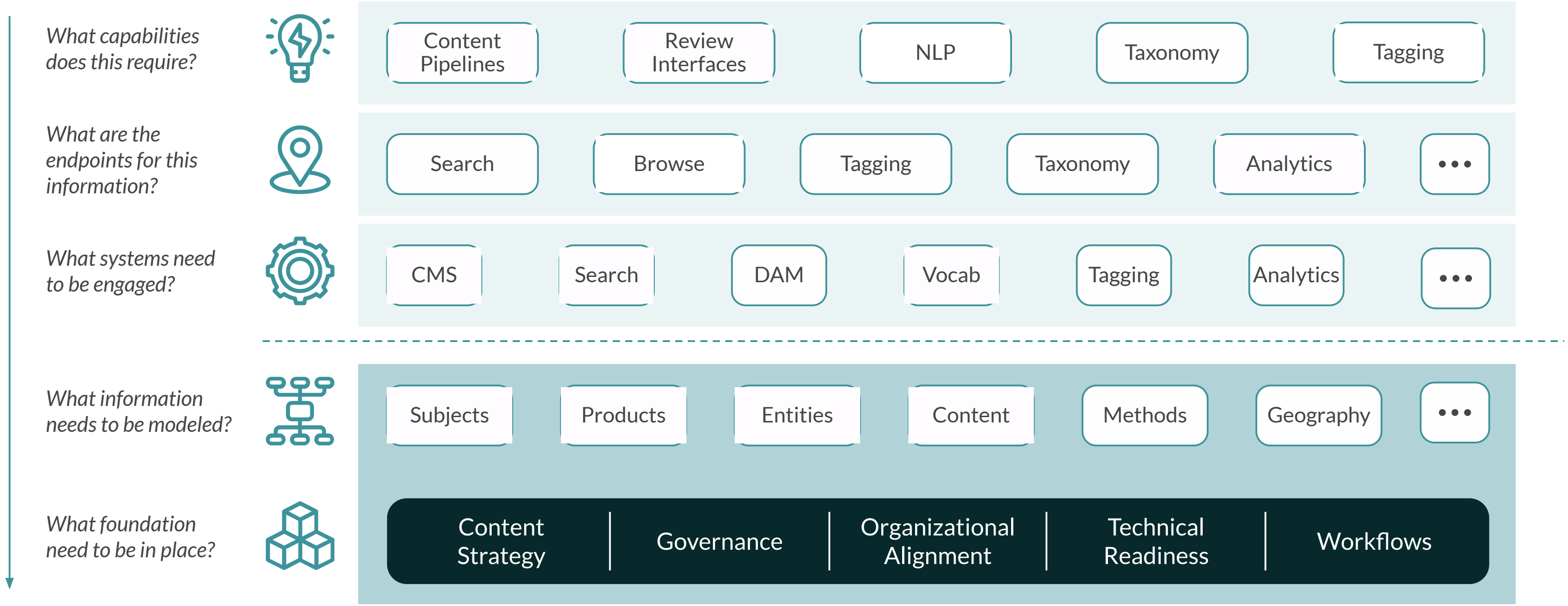
Vocab systems

Indexing Systems

CMS/DAM repositories

Search

WORKING FROM CAPABILITIES TO FOUNDATION



What capabilities does this require?

Content Pipelines · Review Interfaces · NLP · Taxonomy · Tagging

What are the endpoints for this information?

Search · Browse · Tagging · Taxonomy · Analytics · •••

What systems need to be engaged?

CMS · Search · DAM · Vocab · Tagging · Analytics · •••

What information needs to be modeled?

Subjects · Products · Entities · Content · Methods · Geography · •••

What foundation need to be in place?

Content Strategy | Governance | Organizational Alignment | Technical Readiness | Workflows

WHAT DOES SUCCESS LOOK LIKE?

- Accuracy
  - How do we measure this? What is “good”?
- Implementation(s)
  - How can we leverage the tagging?
- Governance and expansion
  - How to maintain and expand capabilities?

WHAT DOES ACCURACY LOOK LIKE?

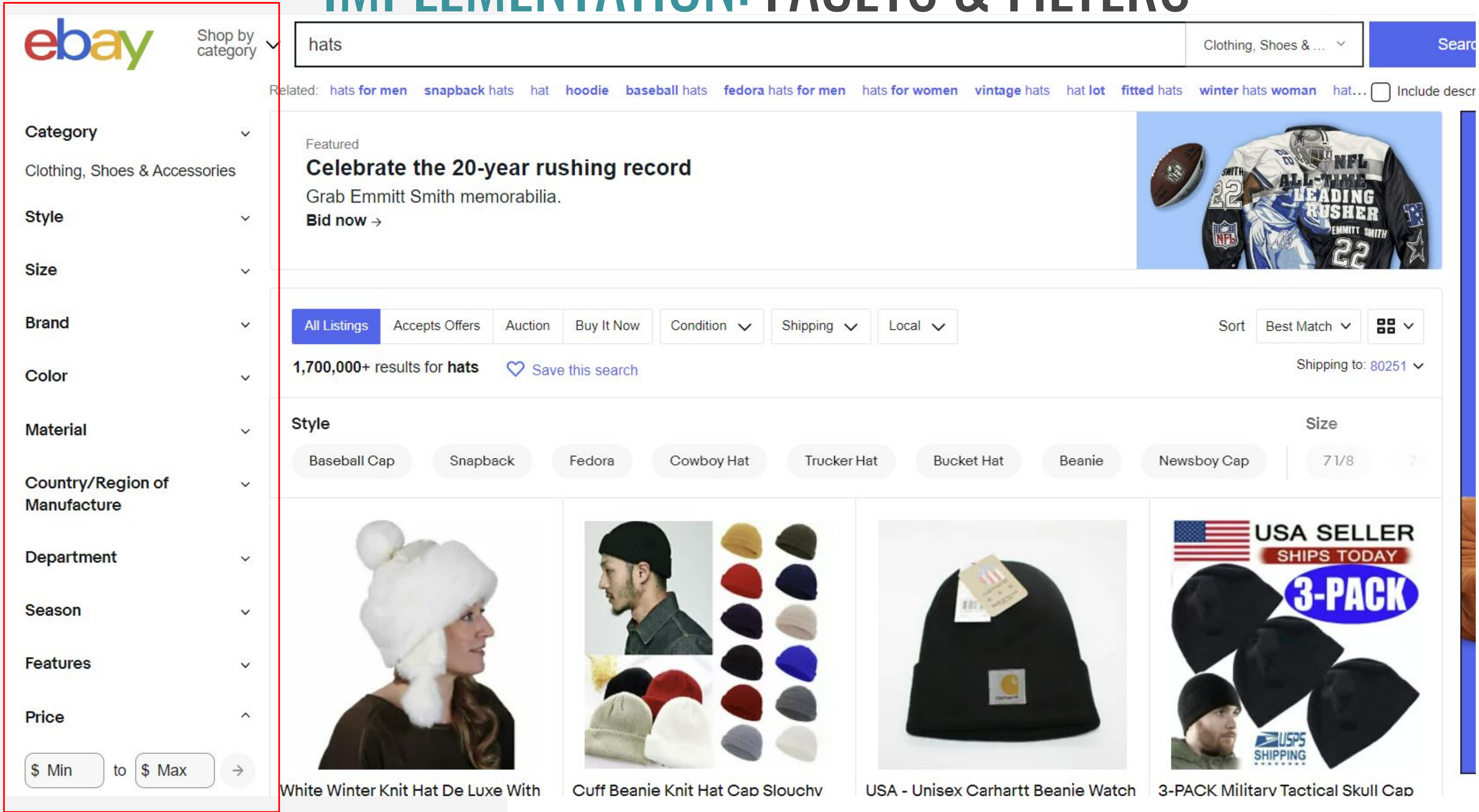
- Accurate (to some threshold) categorization
  - **85%** is VERY good in fully automated systems
- QC of the tagging is a feedback loop, not just a corrective
  - Improve tagging performance
  - Capture new concepts for taxonomy

factor

WHAT DOES IMPLEMENTATION LOOK LIKE?

- Enhanced search/browse experience
- Tags exploded as search queries
- How to surface in an interface?
  - Facets and filters, browse options, type-ahead

IMPLEMENTATION: FACETS & FILTERS



ebay Shop by category

hats | Clothing, Shoes & ... | Searc

Related: hats for men snapback hats hat hoodie baseball hats fedora hats for men hats for women vintage hats hat lot fitted hats winter hats woman hat... □ Include descr

Category

Clothing, Shoes & Accessories

Style

Size

Brand

Color

Material

Country/Region of Manufacture

Department

Season

Features

Price

$ Min to $ Max

Featured

Celebrate the 20-year rushing record

Grab Emmitt Smith memorabilia.

Bid now →

All Listings | Accepts Offers | Auction | Buy It Now | Condition | Shipping | Local

Sort Best Match

1,700,000+ results for **hats** ♡ Save this search

Shipping to: 80251

Style

Baseball Cap | Snapback | Fedora | Cowboy Hat | Trucker Hat | Bucket Hat | Beanie | Newsboy Cap

Size

7 1/8

White Winter Knit Hat De Luxe With

Cuff Beanie Knit Hat Cap Slouchy

USA - Unisex Carhartt Beanie Watch

3-PACK Military Tactical Skull Cap

WHAT DOES EXTENSIBILITY LOOK LIKE?

- Processes should be repeatable/extensible once capabilities are in place
    - Providing autocat as a service
    - Adding other content sets
        - (This might require separate tagging rulebases!)
    - Adding taxonomies for tagging
    - Re-tagging backfile after taxonomy changes

factor

THANK YOU!

Questions?

Bob Kasenchak
bob.kasenchak@factorfirm.com

@taxobob

www.factorfirm.com