

Practical machine learning to accelerate data intelligence

Pretty easy / fast (bespoke).
Pretty good (no innovation).
Pretty cheap (scalable).
... but "enterprise-y"
(immediate value)

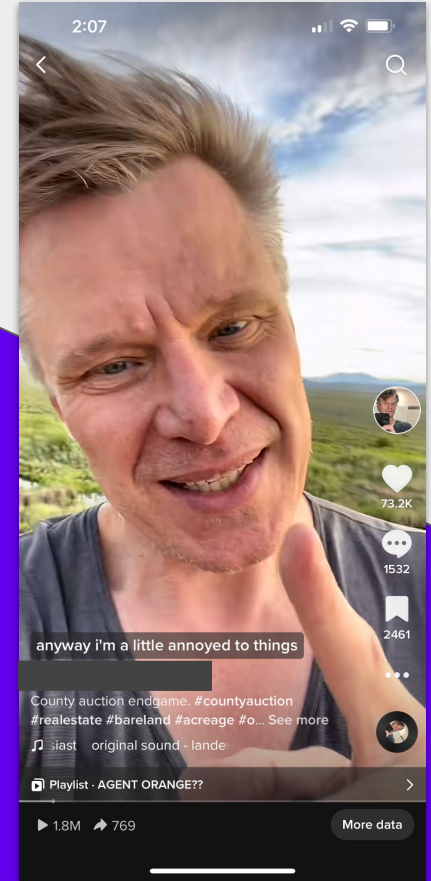
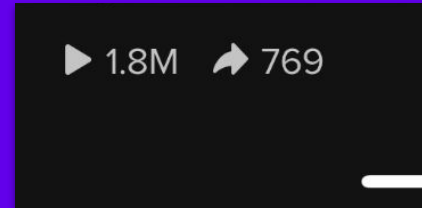
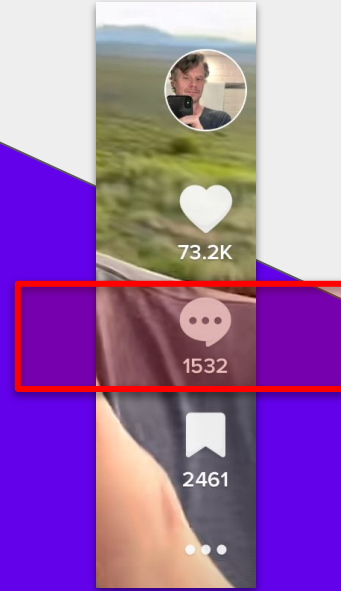


Peter Bray



I am a middle-aged
"land micro-influencer"
that was (slightly)
bullied on social media.

- But I made \$300!
- And got many more
"homestead" competitors.



(Humblebragging, I admit.)

Criticisms followed
several themes:

- Appearance defects (hair etc)
- My pronunciation (apparently I can't pronounce the letter S)
- Alleged pesticides on land

I wondered...

What ML techniques can help make
sense of 100s of critical
comments?



Comments 652

Likes 17.0K



you look like Tom Brady if he were an alcoholic
chain smoker, so still pretty good

3m Reply





But I'm not in the social analytics space anymore!

Rather, our products crawl the web to detect changes.

But we have a similar problem!

What practical solutions can accelerate insights from vast new and changing web data?

Captured: Sep 11, 6:54 am

Interim Guidance for SARS-CoV-2 Testing in Homeless Shelters | CDC
</coronavirus/2019-ncov/community/homeless-shelters/testing-guidance>

—snipped—

* As noted in the labeling for authorized over-the-counter antigen tests: Negative results should be treated as presumptive (meaning that they are preliminary results). Negative results do not rule out SARS-CoV-2 infection and should not be used as the sole basis for treatment or patient management decisions, including infection control decisions. Please see FDA guidance on the use of at-home COVID-19 antigen tests.

Diffs: [Side-by-side](#) / [Text](#) / [Code](#) / [Screenshot](#) / [Net](#) / [Version](#)

Captured: Sep 11, 6:48 am

Use and Care of Masks | CDC
</coronavirus/2019-ncov/prevent-getting-sick/masks-protect-you>

—snipped—

Updated ~~Feb. 25~~ **Sept. 9**, 2022

—snipped—

~~CDC is reviewing this page to align with updated guidance.~~

Masks can help protect you and others from COVID-19. Learn more about different types of masks and respirators and how to get the best fit.

—snipped—

* If you are at high risk for getting very sick, wear a well-fitting high-quality mask.

Diffs: [Side-by-side](#) / [Text](#) / [Code](#) / [Screenshot](#) / [Net](#) / [Version](#)

Captured: Sep 11, 6:25 am

Travel | CDC
</coronavirus/2019-ncov/travelers/when-to-delay-travel.html>

—snipped—

Updated ~~Aug. 24~~ **Sept. 8**, 2022

—snipped—

If Your COVID-19 Test is Positive

This poster is available to download and can be used as a resource for airport testing sites. The poster reminds travelers of actions they should take if their COVID-19 test is positive.

English [PDF — 409 KB, 1 page]

Diffs: [Side-by-side](#) / [Text](#) / [Code](#) / [Screenshot](#) / [Net](#) / [Version](#)

Captured: Sep 11, 5:59 am

Stay Up to Date with COVID-19 Vaccines Including Boosters | CDC
</coronavirus/2019-ncov/vaccines/fully-vaccinated.html>

—snipped—

Updated **Sept. 28**, 2022

—snipped—

* ~~A~~ People ages 6 months through 4 years should get all COVID-19 primary series doses.

* ~~A~~ People ages 5 years and older should get all primary series doses, and ~~updated COVID-19 boosters if eligible~~ the booster dose recommended for them by CDC, if eligible.

* People ages 5 years to 11 years are currently recommended to get the booster dose.

Diffs: [Side-by-side](#) / [Text](#) / [Code](#) / [Screenshot](#) / [Version](#)

Before I get into that, let's look a bit closer at Fluxguard

1 OPEN

2 CLICK ⬇ ⚙️ 🗑️

```
waitMs: 500  
selector: ".regionSelect"
```

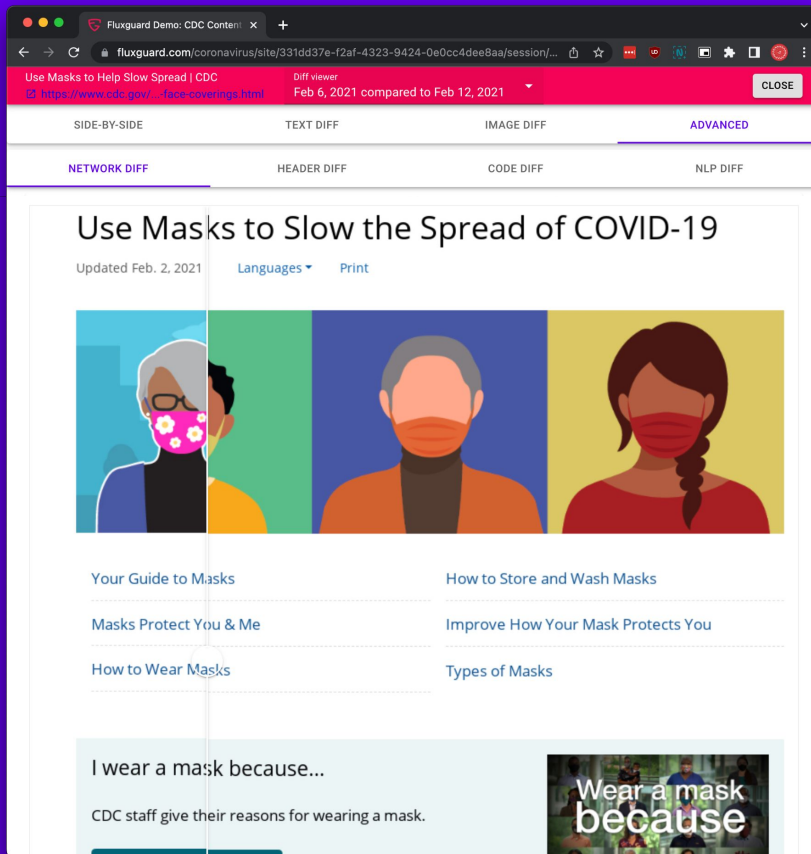
3 CLICK ⬆ ⬇ ⚙️ 🗑️

```
waitMs: 500  
selector: "#select2-result-label-4"
```

4 JAVASCRIPT_EXECUTION ⬆ ⬇ ⚙️ 🗑️

```
1 // Set age.  
2 $(".age-input").first().val("30").change();  
3  
4 // Set start date.  
5 const MyDate = new Date();  
6 MyDate.setDate(MyDate.getDate() + 1);  
7 $("#startDate").val(("0" + MyDate.getDate()).slice(-2) +  
8   '/' + ('0' + (MyDate.getMonth()+1)).slice(-2) + '/' +  
9   MyDate.getFullYear()).change();  
10  
11 // Set end date.  
12 var MyDate2 = new Date();  
13 MyDate2.setDate(MyDate2.getDate() + 10);  
14 $("#endDate").val(("0" + MyDate2.getDate()).slice(-2) +  
15   '/' + ('0' + (MyDate2.getMonth()+1)).slice(-2) + '/' +  
16   MyDate2.getFullYear()).change();
```

Multi-step/page
orchestration and
change monitoring.



Puppeteer-based
monitoring with
replay of all cookies
/ storage; network
interception to store
resources (no use
of browser cache).

Use cases:

- Multi-step form / partner validations
- Pharma regs
- Formularies
- EULAs
- Competitors

HOME + NEWS FDA RESOURCES + NLM SPL RESOURCES + APPLIC

LABEL: ALIMTA- pemetrexed disodium injection, powder, lyophilized, for solution

DRUG LABEL INFORMATION Updated August 31, 2022

If you are a consumer or patient please visit [this version](#).

DOWNLOAD DRUG LABEL INFO: [PDF](#) | [XML](#) | [PDF](#) OFFICIAL LABEL (PRINTER FRIENDLY) [PDF](#)

[VIEW ALL SECTIONS](#)

- HIGHLIGHTS OF PRESCRIBING INFORMATION**
These highlights do not include all the information needed to use ALIMTA safely and effectively. See full prescribing information for ALIMTA. ALIMTA (pemetrexed for injection), for Intravenous ...
- TABLE OF CONTENTS**
Table of Contents
- 1 INDICATIONS AND USAGE**
1.1 Non-Squamous Non-Small Cell Lung Cancer (NSCLC) ALIMTA® is indicated: in combination with pembrolizumab and platinum chemotherapy, for the initial treatment of patients with ...
- 2 DOSAGE AND ADMINISTRATION**
2.1 Recommended Dosage for Non-Squamous NSCLC - The recommended dose of ALIMTA when administered with pembrolizumab and platinum chemotherapy for the initial treatment of metastatic ...
- 3 DOSAGE FORMS AND STRENGTHS**
For injection: 100 mg or 500 mg pemetrexed as a white to light-yellow or green-yellow lyophilized powder in single-dose vials for reconstitution.
- 4 CONTRAINDICATIONS**
ALIMTA is contraindicated in patients with a history of severe hypersensitivity reaction to pemetrexed [see Adverse Reactions (6.1)].

Back to our problem... how can we correctly categorize well-defined content and changes into bespoke categories?

fluxguard.com/renderere... CLOSE

DailyMed - ALIMTA- pemetrexed disodium injection, powder...
<https://dailymed.nlm.nih...c-429c-ae04-9c88d7c55c08>

SIDE-BY-SIDE TEXT DIFF IMAGE DIFF ADVANCED

8.3 FEMALES AND MALES OF REPRODUCTIVE POTENTIAL
Based on animal data ALIMTA can cause malformations and developmental delays when administered to a pregnant woman [see Use in Specific Populations (8.1)].
[Pregnancy Testing - Verify pregnancy ...](#)

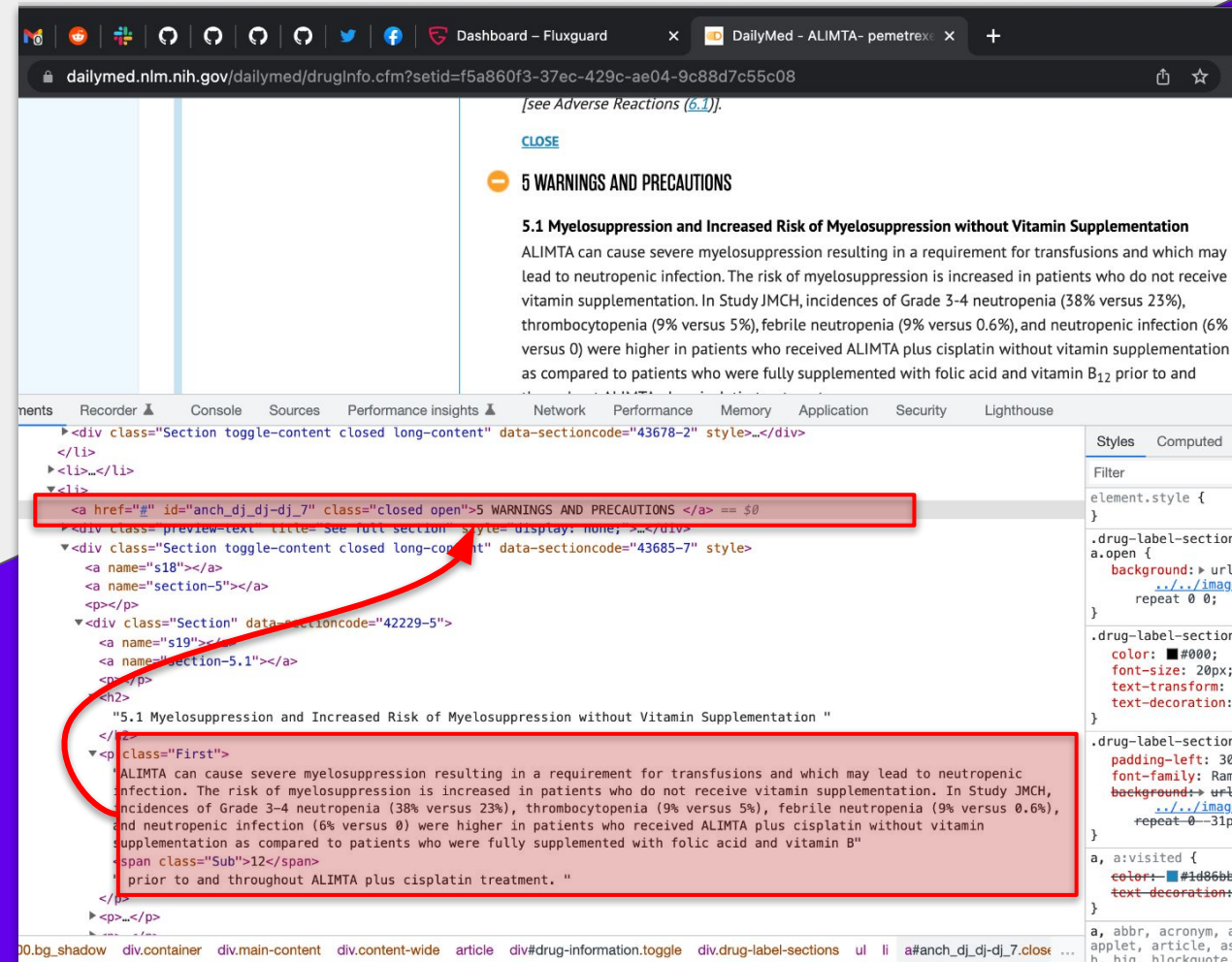
Based on animal data ALIMTA can cause malformations and developmental delays when administered to a pregnant woman [see Use in Specific Populations (8.1)].
[Pregnancy Testing](#)

Verify pregnancy status of females of reproductive potential prior to initiating Pemetrexed Injection [see Use in Specific Populations (8.1)].
[Contraception](#)

[Females](#)

Can we traverse the DOM to build rules to categorize?

- Brittle and imperfect. *Cheerio!*
- Sorta requires us to accept target site categorization / labels
- Necessitates constant tinkering to work across varied sources
- Problem also for finding specific thing on a page! What about PDFs?



The screenshot shows a web browser window with the URL `dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=f5a860f3-37ec-429c-ae04-9c88d7c55c08`. The page content displays a section titled "5 WARNINGS AND PRECAUTIONS" and a subsection "5.1 Myelosuppression and Increased Risk of Myelosuppression without Vitamin Supplementation". The browser's developer tools are open, showing the DOM tree on the left and the Styles pane on the right. A red box highlights a specific DOM element in the tree, and a red arrow points from it to the corresponding text in the page content.

DOM Tree (Left):

```
<div class="Section toggle-content closed long-content" data-sectioncode="43678-2" style="display: none;">
  <li>
    <a href="#" id="anch_dj_dj-dj_7" class="closed open">5 WARNINGS AND PRECAUTIONS </a> == $0
  </li>
</div>
```

Styles (Right):

```
element.style {
  background: url(
    repeat 0 0;
  }
.drug-label-sections
a.open {
  background: url(
    repeat 0 0;
  }
.drug-label-sections
color: #000;
font-size: 20px;
text-transform: u
text-decoration: v
}
.drug-label-sections
padding-left: 30px;
font-family: Rama
background: url(
  repeat 0 31px
}
a, a:visited {
  color: #1d06bb;
  text-decoration: u
}
```

This is a good job for Google Vertex!

- Pretty good "off-the-shelf" ML platform with great UI
- Great for bespoke modeling of well-defined page and content types
 - About us, Press Release
 - Biography, Comment
- Problem: Chicken/egg situation when it comes to training.
- Nothing easily off-the-shelf: custom training.

The screenshot shows the Google Vertex AI interface. At the top, the URL is `google.com/vertex-ai/locations/us-central1/datasets/105009957522374656;anno...`. The interface includes a search bar and a navigation menu. The main content area is titled "Item 1 of many" and shows a "Data split" dropdown set to "Default". Below this, there are tabs for "CLASSIFICATION" and "DETAILS". The "CLASSIFICATION" tab is active, showing a "Filter" section with a list of labels: "Adverse_Reactions" (selected), "Contraindications", "Drug_Interactions", "Geriatric_Use", "Overdosage", "Pediatric_Use", and "Warnings". To the right of the filter list, there is a text box containing a paragraph about ALIMTA: "ALIMTA can cause severe myelosuppression resulting in a requirement for transfusions and which may lead to neutropenic infection. The risk of myelosuppression is increased in patients who do not receive vitamin supplementation. In Study JMCH, incidences of Grade 3-4 neutropenia (38% versus 23%), thrombocytopenia (9% versus 5%), febrile neutropenia (9% versus 0.6%), and neutropenic infection (6% versus 0) were higher in patients who received ALIMTA plus cisplatin without vitamin". A red arrow points from the "Model Evaluation" text to a confusion matrix table below. The table has "True label" on the left and "Predicted label" on top. The "Predicted label" row has "adverseReactions" and "precautions". The "True label" row has "adverseReactions" and "precautions". The table shows the following values: for "adverseReactions" true label, 91% predicted as "adverseReactions" and 9% as "precautions"; for "precautions" true label, 1% predicted as "adverseReactions" and 99% as "precautions". A red arrow points from the "Deployed Filtering" text to a text box on the right. The text box contains the text: "8.3 FEMALES AND MALES OF Based on animal data ALIMTA can cause when administered to a pregnant woman Pregnancy Testing - Verify pregnancy ... Based on animal data ALIMTA can cause when administered to a pregnant woman Pregnancy Testing".

Model Evaluation

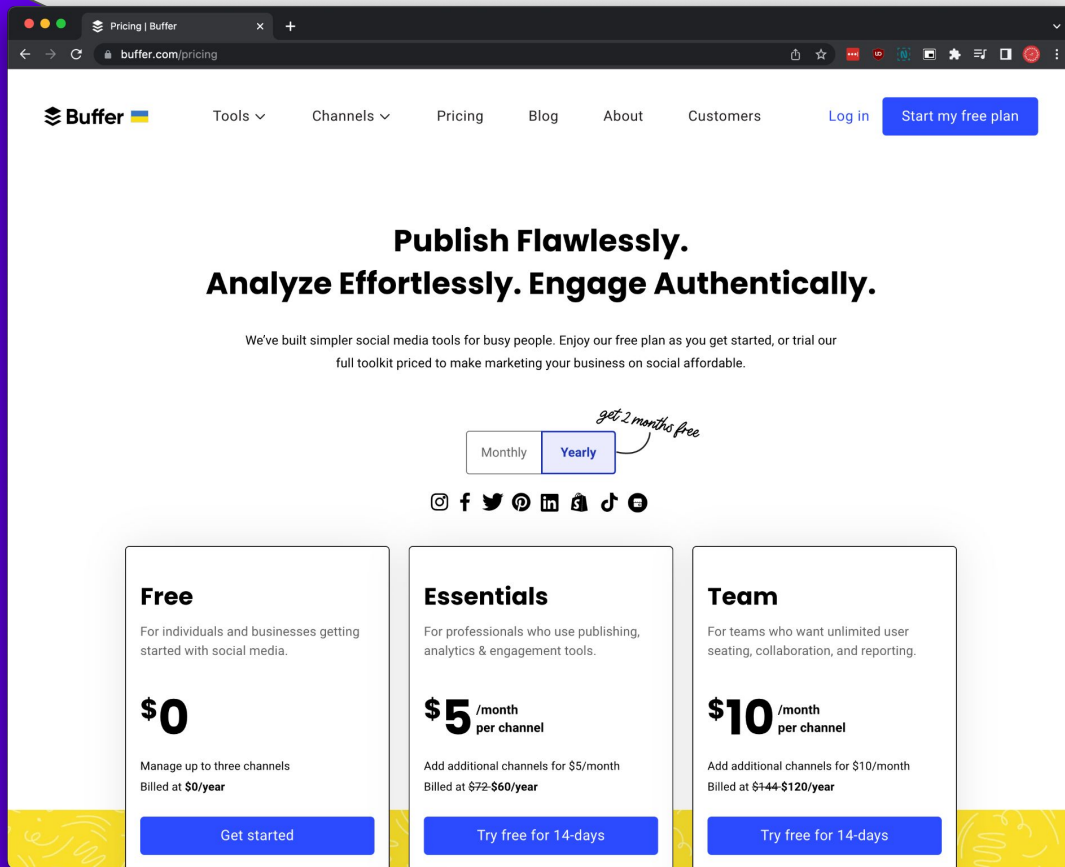
| True label | Predicted label | |
|------------------|------------------|-------------|
| | adverseReactions | precautions |
| adverseReactions | 91% | 9% |
| precautions | 1% | 99% |

Deployed Filtering

8.3 FEMALES AND MALES OF
Based on animal data ALIMTA can cause when administered to a pregnant woman
Pregnancy Testing - Verify pregnancy ...
Based on animal data ALIMTA can cause when administered to a pregnant woman
Pregnancy Testing

Other content types can be difficult to classify with "traditional" NLP.

- Home pages
- Pricing
- ...and other sparsely worded, design-rich content



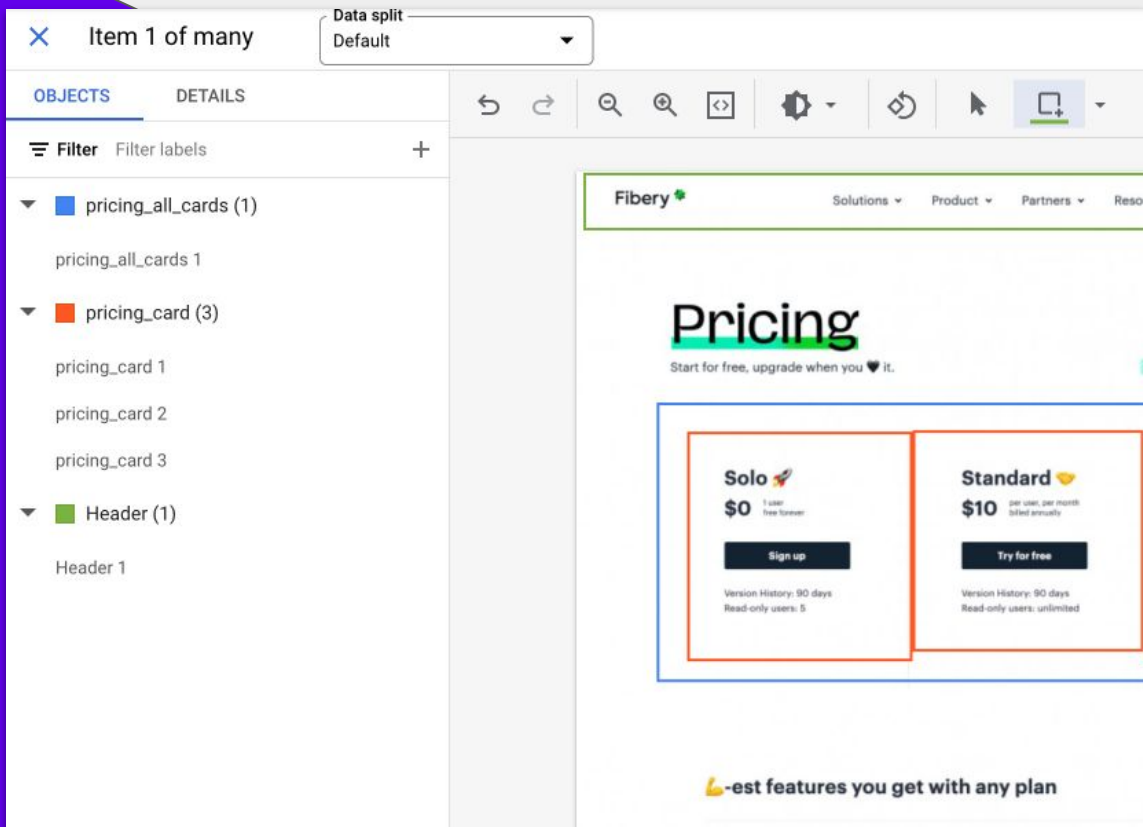
The screenshot shows the Buffer pricing page. At the top, there's a navigation bar with the Buffer logo, links for Tools, Channels, Pricing, Blog, About, and Customers, and buttons for Log in and Start my free plan. The main heading reads "Publish Flawlessly. Analyze Effortlessly. Engage Authentically." Below this, a sub-headline states: "We've built simpler social media tools for busy people. Enjoy our free plan as you get started, or trial our full toolkit priced to make marketing your business on social affordable." The pricing section features three plans: Free, Essentials, and Team. The Free plan is highlighted with a "get 2 months free" callout. The Essentials and Team plans offer a 14-day free trial. Social media icons are displayed below the plans.

| Free | Essentials | Team |
|---|---|--|
| For individuals and businesses getting started with social media. | For professionals who use publishing, analytics & engagement tools. | For teams who want unlimited user seating, collaboration, and reporting. |
| \$0 | \$5 /month per channel | \$10 /month per channel |
| Manage up to three channels Billed at \$0/year | Add additional channels for \$5/month Billed at \$72-\$60/year | Add additional channels for \$10/month Billed at \$144-\$120/year |
| Get started | Try free for 14-days | Try free for 14-days |

We can use Vertex to classify screenshots!

- Optimized for real world photos, but works pretty well
- Detect bespoke UI elements
- May require finishing with OCR + CSS selector identification.

(This is broadly the same approach that some sophisticated services provide—e.g., Diffbot).



Sometimes we want to get a little more specific!

Named Entity Recognition extracts specific pieces of well-defined categories (people, places, anatomy, financial events, etc)



| ← automl_Medical | | automl_Medical (1) | | ? |
|------------------------|-----|---------------------|--|---------|
| IMPORT | | BROWSE | | ANALYZE |
| All | 593 | Filter Filter items | | |
| Labeled | 592 | | | |
| Unlabeled | 1 | | | |
| Training | 473 | | | |
| Validation | 59 | | | |
| Test | 60 | | | |
| Filter Filter labels + | | | | |
| Text items ▼ | | | | |
| CompositeMention | 80 | | | |
| DiseaseClass | 322 | | | |
| Modifier | 410 | | | |
| SpecificDisease | 562 | | | |

| Filter | Text |
|--------------------------|---|
| <input type="checkbox"/> | 10545613 Synergistic effect of histone hyperacetyl |
| <input type="checkbox"/> | 8528199 WASP gene mutations in Wiskott-Aldrich |
| <input type="checkbox"/> | 102474 Combined genetic deficiency of C6 and C7 |
| <input type="checkbox"/> | 8533768 Evidence for linkage of bipolar disorder to |
| <input type="checkbox"/> | 313733 Hereditary C2 deficiency associated with c |
| <input type="checkbox"/> | 8101038 High residual arylsulfatase A (ARSA) activ |
| <input type="checkbox"/> | 1999339 Some Mexican glucose-6-phosphate dehy |
| <input type="checkbox"/> | 8346255 Molecular mechanisms of oncogenic mur |
| <input type="checkbox"/> | 7586656 Southern analysis reveals a large deletion |
| <input type="checkbox"/> | 2817003 Translocation t(5;11)(q13.1;p13) associat |

Text Analysis Options

We offer two different NLP models for analyzing text in several languages. Entity analysis extracts terms of various types from the text of web pages. Syntax analysis breaks down words into parts of speech, such as nouns, verbs, and adjectives. In either case, you can select which types you want to check for changes.

Entity analysis

General purpose NER

Syntax analysis

Medical NER

Financial NER

Date

Event

Both common and proper nouns ▼

Location

Both common and proper nouns ▼

Number

Organization

Both common and proper nouns ▼

Person

Both common and proper nouns ▼

Phone number

- Named Entity Recognition is good from both AWS and GCP.
- Custom models are straightforward.
- Textual context is needed for performance!

Session View: www.bbc.com - x +

fluxguard.com/site/2... CLOSE

SIDE-BY-SIDE TEXT DIFF IMAGE DIFF ADVANCED

NETWORK DIFF HEADER DIFF CODE DIFF NLP DIFF

View option
Changes only

```
ORGANIZATION: {
  COMMON: {
    CROWN_JEWELS: "1 mention",
    allies: "2 mentions",
  },
  PROPER: {
    Bohemian Rhapsody 3: "2 mentions",
    JEWELLERY: "1 mention",
    VideoCrown: "1 mention",
    WEST: "2 mentions"
  },
},
PERSON: {
  COMMON: {
    PIPER: "1 mention",
    authorities: "2 mentions",
    officials: "2 mentions",
  },
  PROPER: {
    ELIZABETH II: "9 mentions" "2 mentions",
    Hurricane Fiona 3: "1 mention",
    Hurricane Fiona 4: "1 mention",
  },
},
WORK_OF_ART: {
  COMMON: {
    PHOTOS: "1 mention",
    post: "2 mentions"
  },
  PROPER: {
    Bohemian Rhapsody 4: "2 mentions",
  },
},
}
```

Dashboard - Fluxguard x +

fluxguard.com/nlpvie... CLOSE

DailyMed - BENDEKA- bendamustine hydrochloride injection, ...
https://dailymed.nlm.nih...bd-4896-abfe-0ef4fec67ddf

SIDE-BY-SIDE TEXT DIFF IMAGE DIFF ADVANCED

NETWORK DIFF HEADER DIFF CODE DIFF NLP DIFF

View option
Changes only

Legend: More Less

Contact us for custom models and bespoke machine learning.

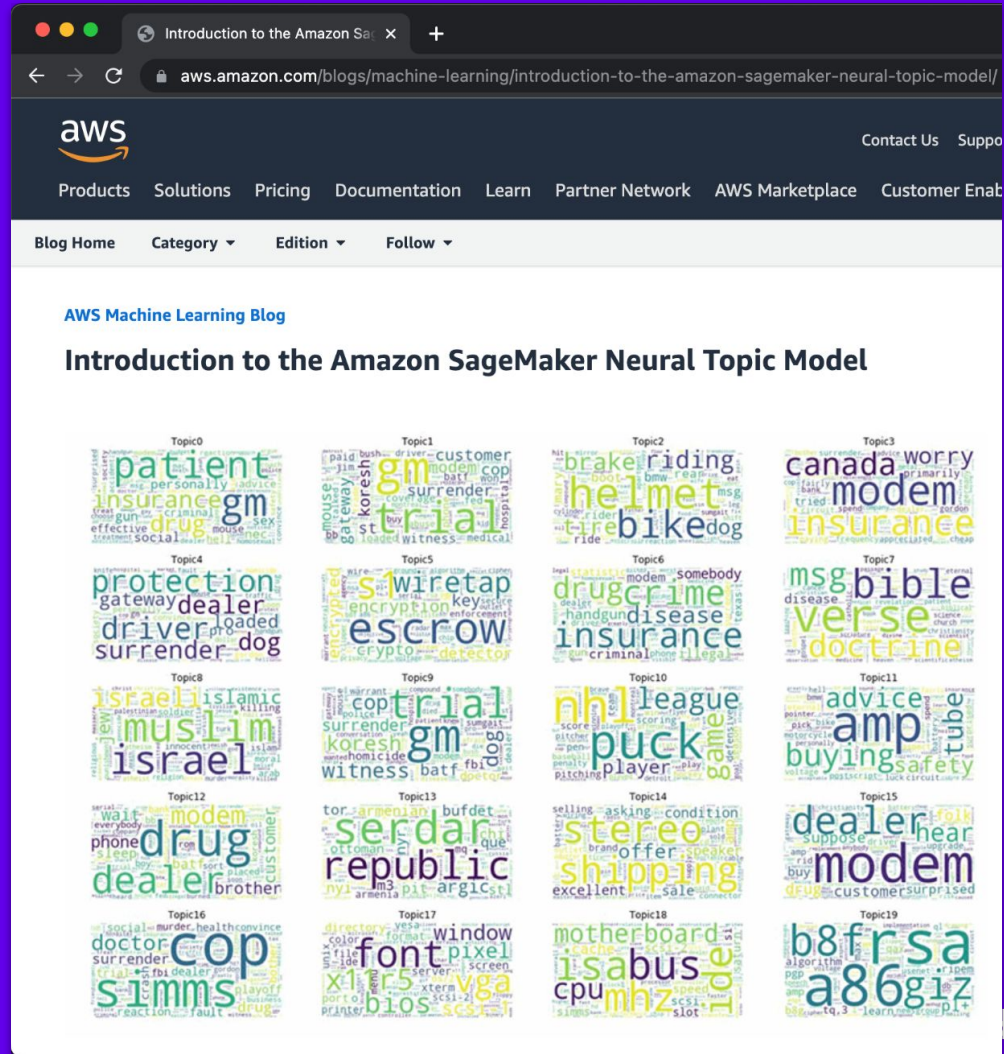
```
{
  MEDICAL_CONDITION: {
    DX_NAME: {
      chills: [ { "Text": "chills" } ],
      cytomegalovirus: [ { "Text": "cytomegalovirus" } ],
      hepatitis b: [ { "Text": "hepatitis B" } ],
      herpes zoster: [ { "Text": "herpes zoster" } ],
      mycobacterium tuberculosis: [ { "Text": "Mycobacterium tuberculosis" } ],
      pruritus: [ { "Text": "pruritus" } ],
    },
  },
  MEDICATION: {
    BRAND_NAME: {
      bendeka: [
        28: { "Text": "BENDEKA" },
        29: { "Text": "BENDEKA" },
        30: { "Text": "BENDEKA" }
      ],
    },
    GENERIC_NAME: {
      antimicrobial: [ { "Text": "antimicrobial" } ],
      bendamustine hydrochloride: [
        11: { "Text": "bendamustine hydrochloride" }
      ],
      chlorambucil: [
        2: { "Text": "chlorambucil" }
      ],
    },
  },
}
```

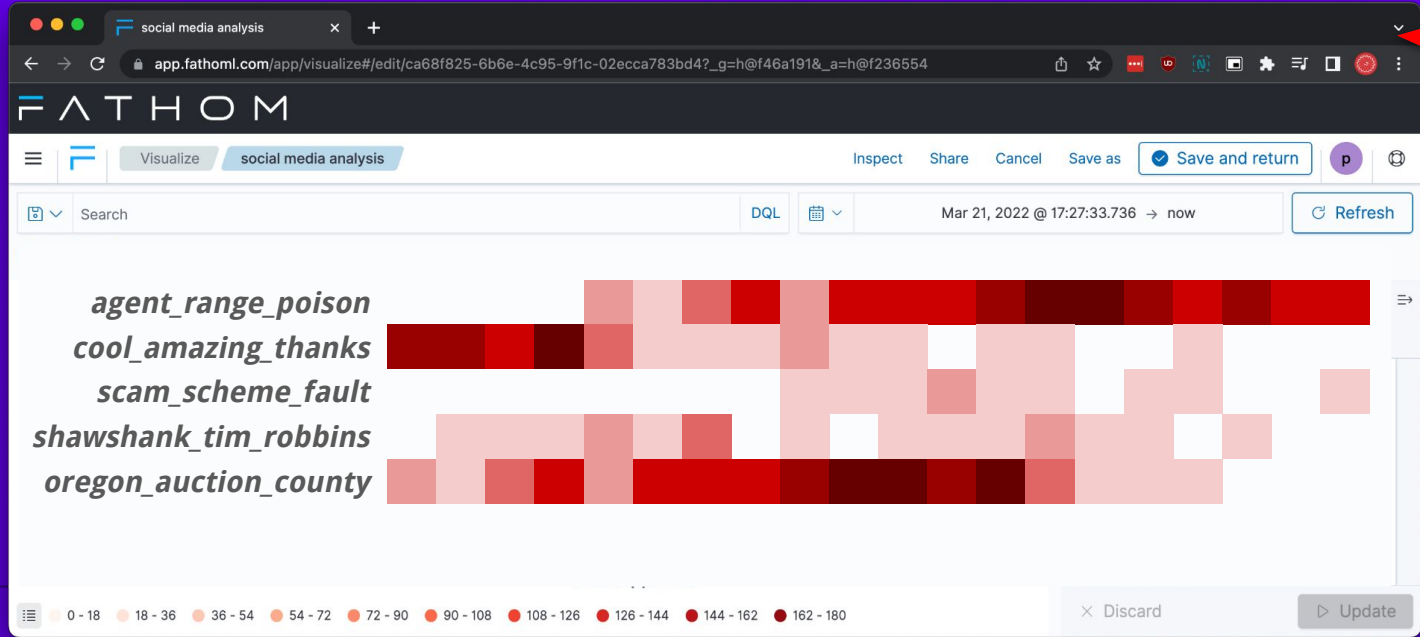
NER can power bespoke Knowledge Graphs in Elasticsearch. AWS Blog Post on how to do this: Google "[blog aws comprehend elasticsearch](#)"



All of the techniques I've highlighted may require tedious labeling. Let's consider semi-supervised topic modeling.

- No labeling
- Produces clusters when we don't know topics in advance
- Pretty easy to get started
 - Github BERTopic
 - Gensim LDA
 - AWS NTM





***Back to
my social
media
hijinks***

Topic modeling works well with entity extraction and Elasticsearch.

- Auto-classify into topics
- Extract key entities. (Sentiment analysis.)
- Visualize over time
- Interactively explore