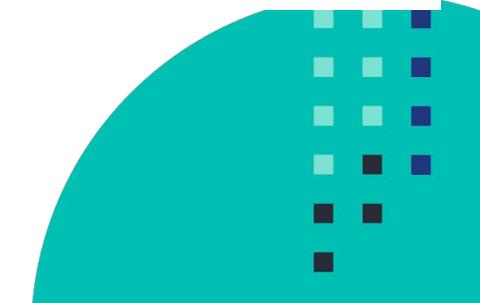




Vancouver Meetup: Beginner's Crash Course to Elastic Stack Series

Part 2: Understanding the relevance of your search with Elasticsearch & Kibana

Lisa Jung
Developer Advocate @Elastic



Connect with the Elastic Community

Find your local User Group:



- <https://community.elastic.co/>

Virtual User Group:



- <https://community.elastic.co/amer-virtual/>

Connect with the Elastic Community

Community Slack Workspace:



https://join.slack.com/t/elasticstack/shared_invite/zt-an1h0etg-04Fl2hA9vvASBkYPe~QZmw



Vancouver Meetup: Beginner's Crash Course to Elastic Stack Series

Part 2: Understanding the relevance of your search with Elasticsearch & Kibana

Lisa Jung
Developer Advocate @Elastic



Beginner's crash course to Elastic Stack Series

- **Part 1: Intro to Elasticsearch and Kibana**
 - use case of Elasticsearch and Kibana
 - the basic architecture of Elasticsearch
 - perform CRUD(Create, Read, Update, and Delete) operations with Elasticsearch and Kibana

Missed the first workshop? No worries!

- **Part 1: Intro to Elasticsearch and Kibana**
 - Repo: <https://ela.st/vancouver-workshop-1>

The Elastic Stack

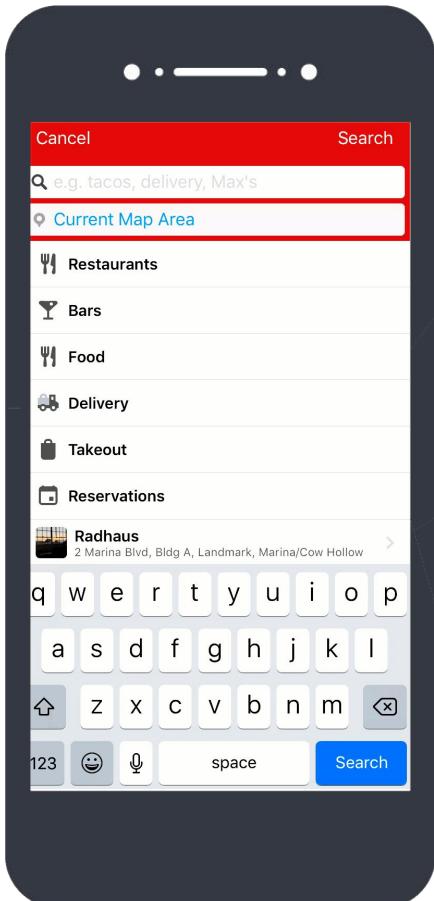
Reliably and securely take data from any source, in any format, then search, analyze, and visualize it in real time.



Elasticsearch

Store | Search | Analyze



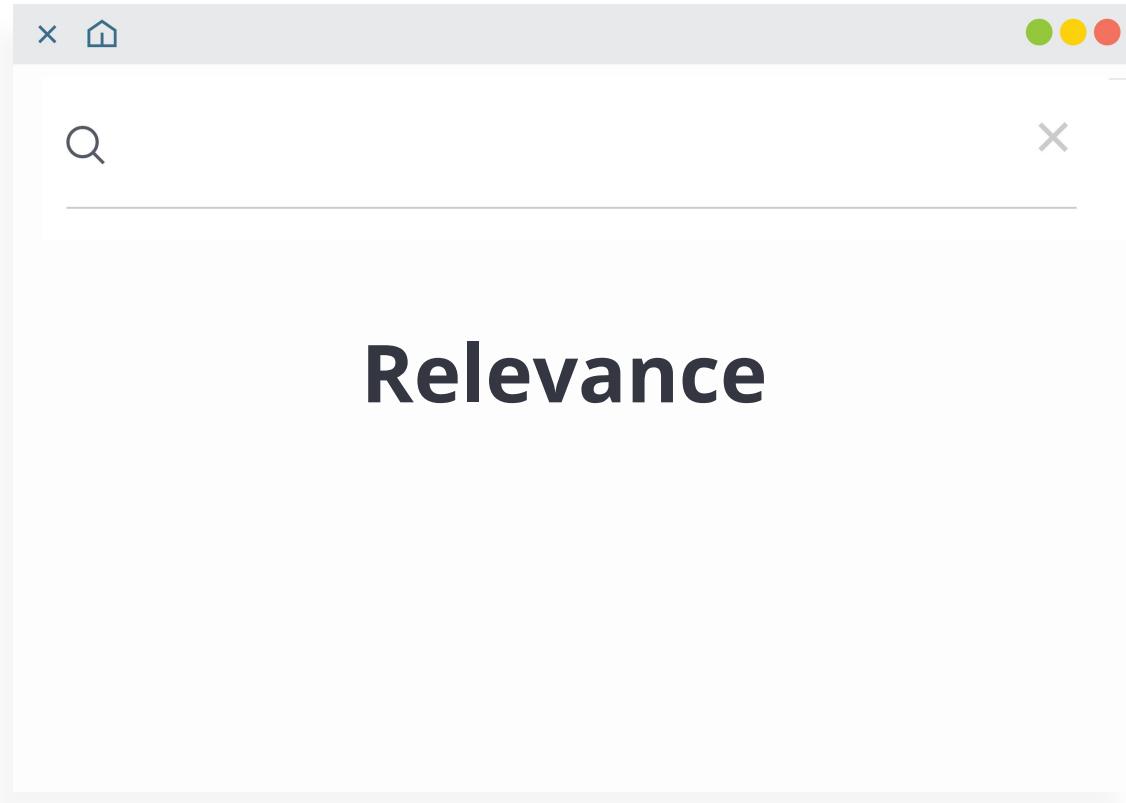


The image shows the Instacart mobile application interface. At the top, there is a navigation bar with a "X" button, a house icon, and three colored dots (green, yellow, red) on the right. The main header features the Instacart logo and a "Stores" dropdown. On the right side of the header are links for "Delivery in 94086", "Account", "Help", and a green "Cart" button with a red notification badge showing the number "4". The background of the screen is a photograph of fresh produce including avocados and kale. In the center, there is a Safeway logo with the word "Safeway" below it and links to "View pricing policy" and "More info". A search bar at the bottom has the prefix "can|". A dropdown menu is open, listing categories under the heading "Canned Goods": "Department", "Canned Fruit & Applesauce", "Canned & Jarred Vegetables", and "Canned Meals & Beans". Below this, there is a "Coupon saving" section with a "Shop Coupons" button, and a "Save Now" button next to a Kraft logo. At the bottom of the screen, there is a message "Based on your cart" and a "View more" link.

Speed, Scale, **Relevance**

Elastic is a search company.

We focus on value to users by producing fast results that operate at scale and are relevant. This is our DNA. We believe search is an experience. It is what defines us, and makes us unique.



How do we measure relevance?

- **Precision**
- **Recall**

Elasticsearch

Store | Search | Analyze



I store data as documents!

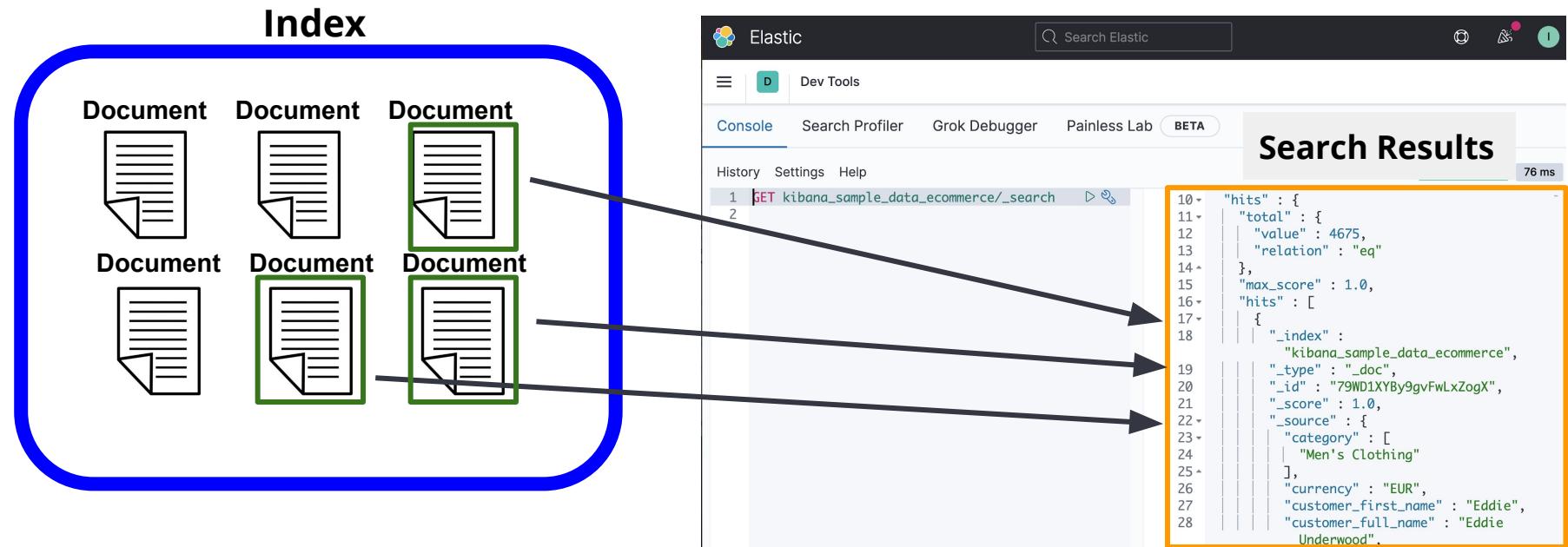
Index

Document Document Document
Document Document Document



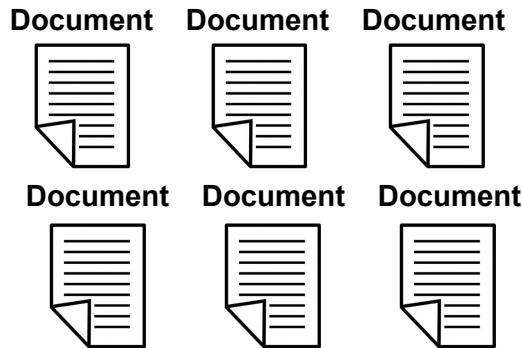
Documents with similar traits are grouped into an index!

When search query is sent, Elasticsearch retrieves relevant documents and presents the documents as search results.

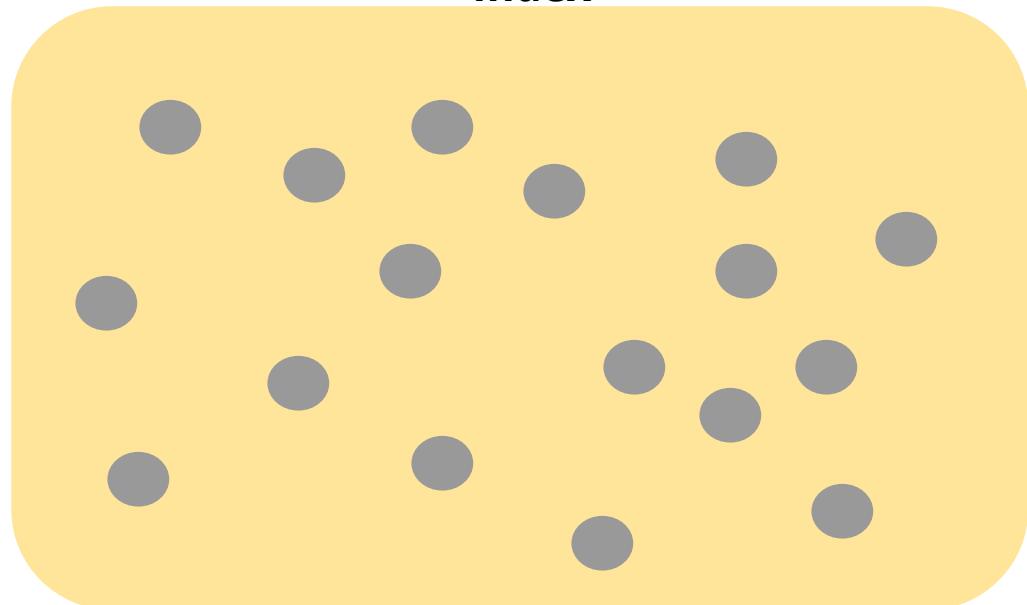


These two diagrams depict the same thing!

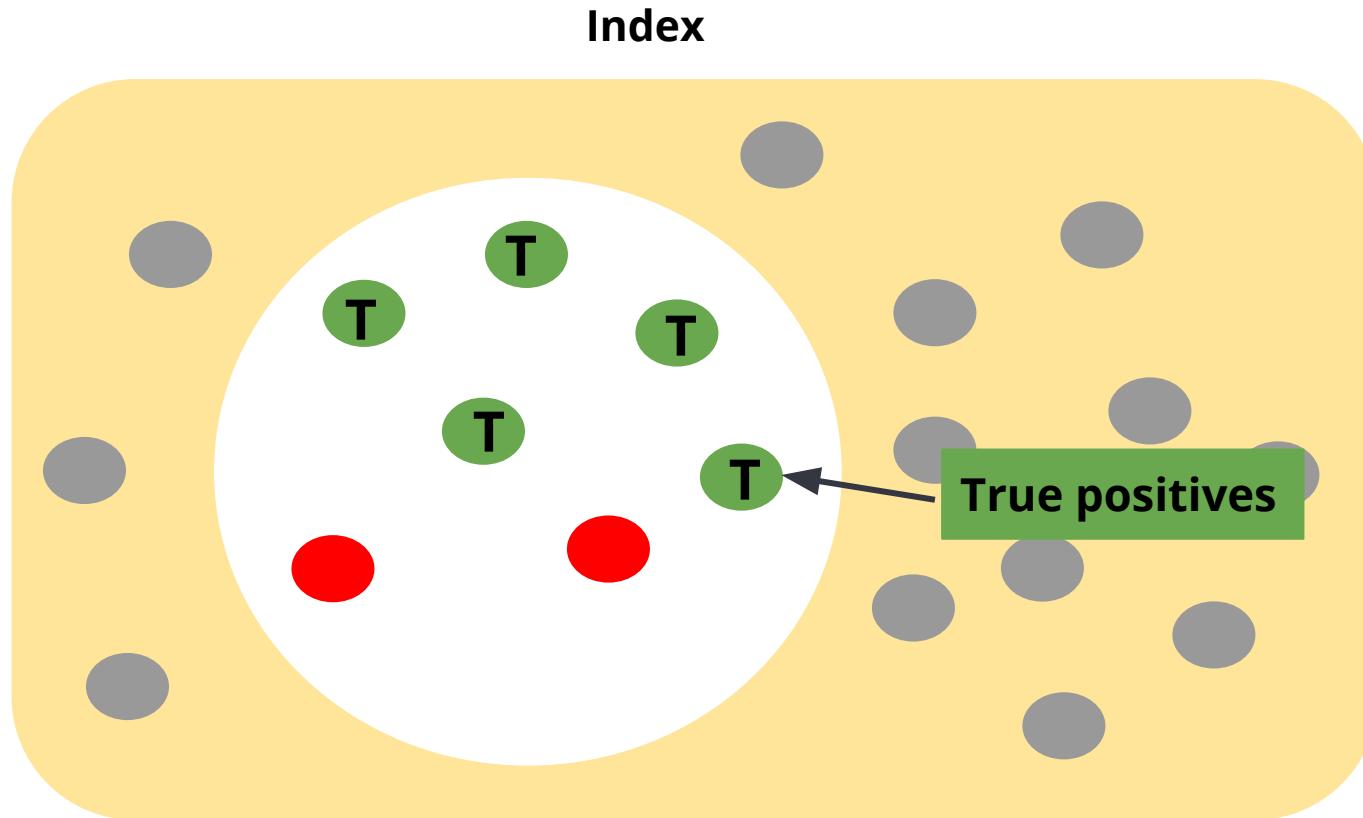
Index



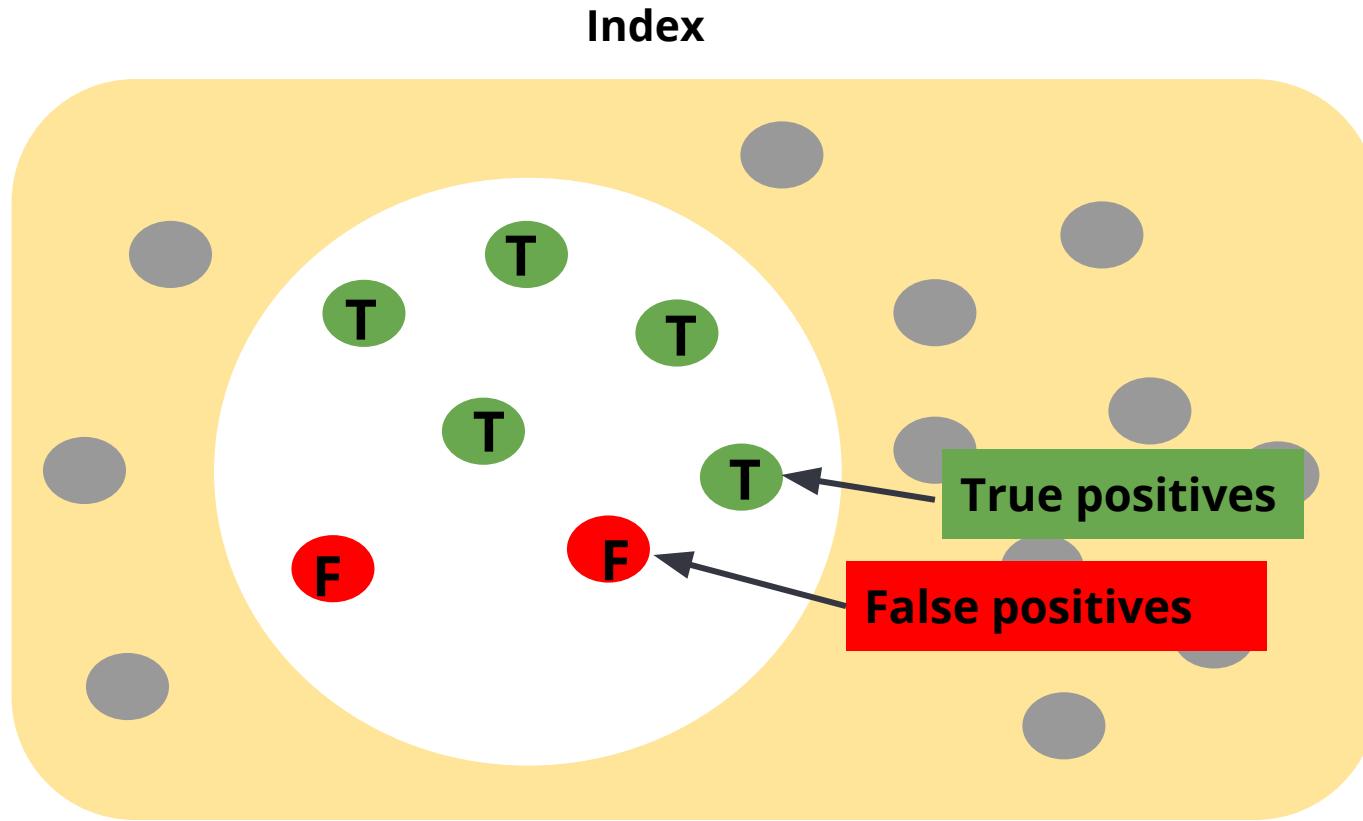
Index



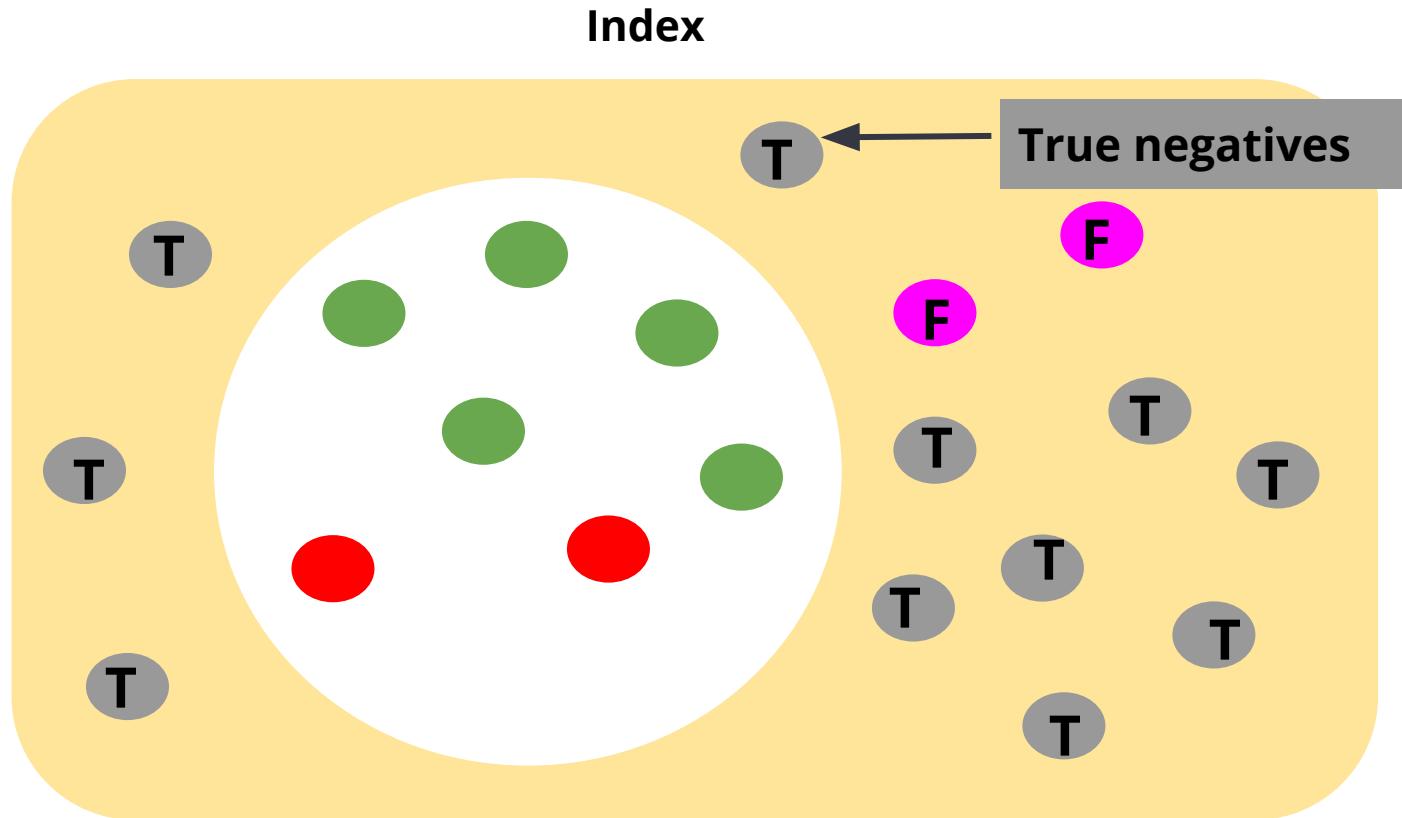
True positives are relevant documents that are returned to the user.



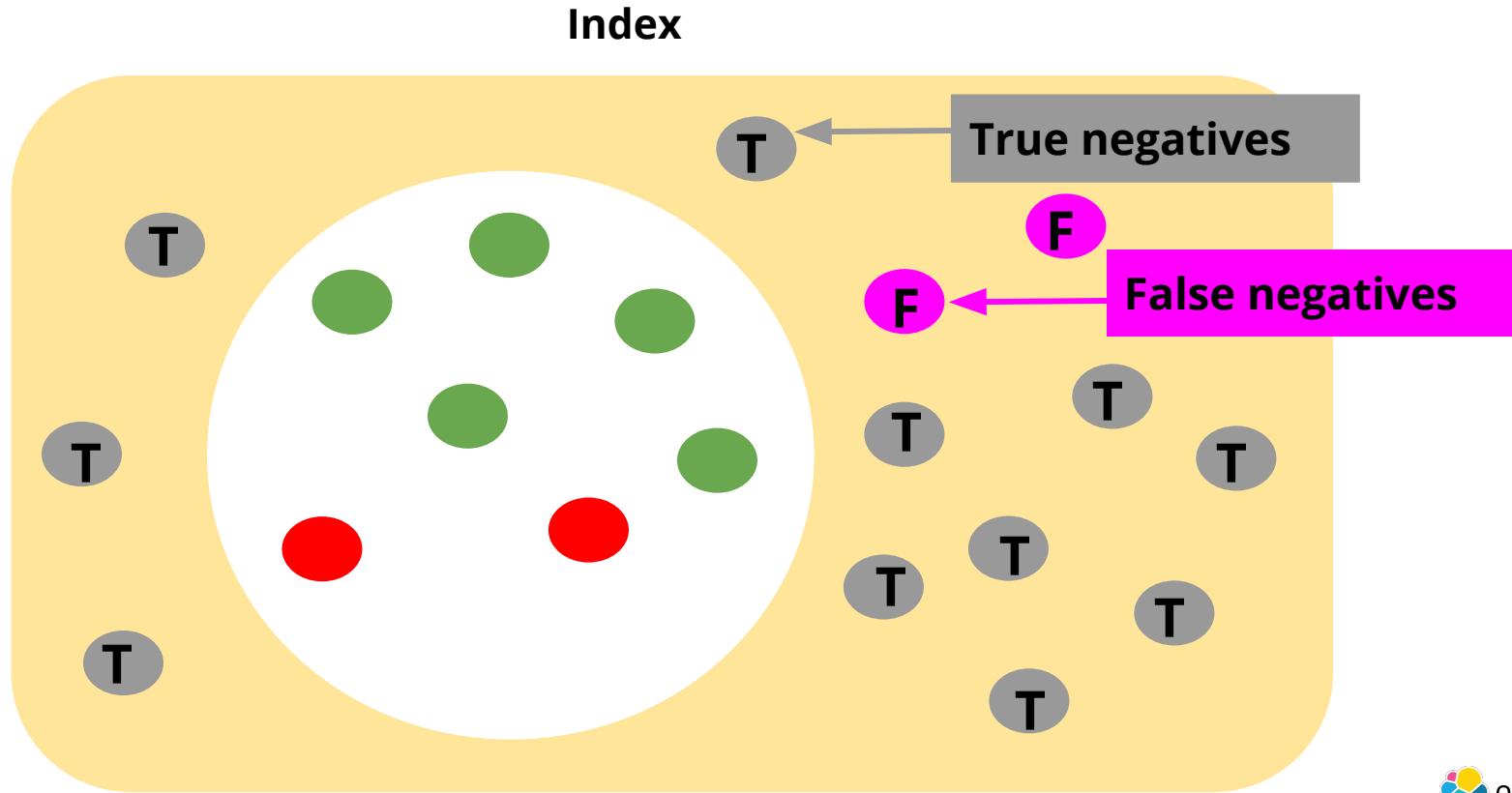
False positives are irrelevant documents that are returned to the user.



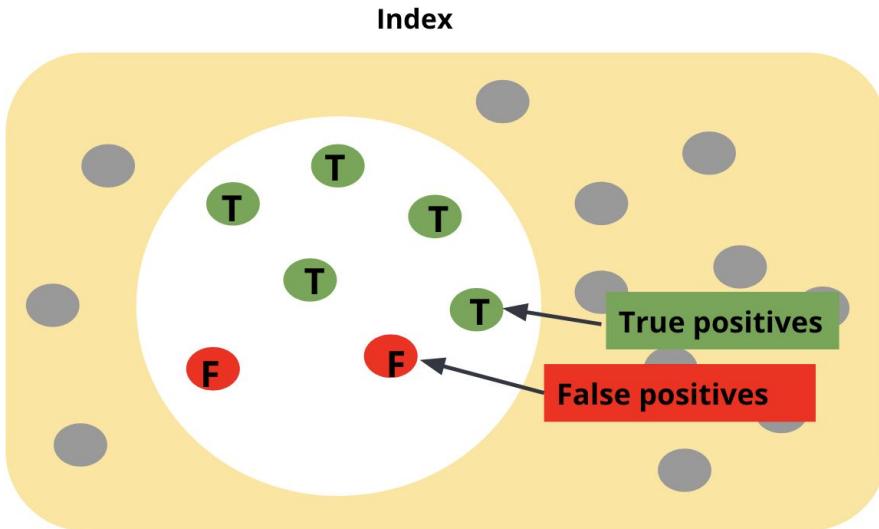
True negatives are irrelevant documents that are not returned to the user.



False negatives are relevant documents that were not returned to the user.



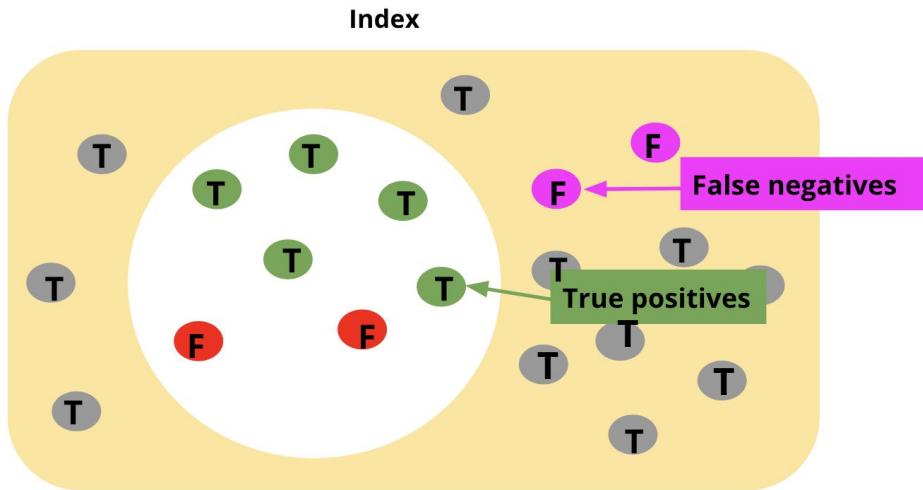
What is precision?



$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

What portion of the retrieved data is actually relevant to the search query?

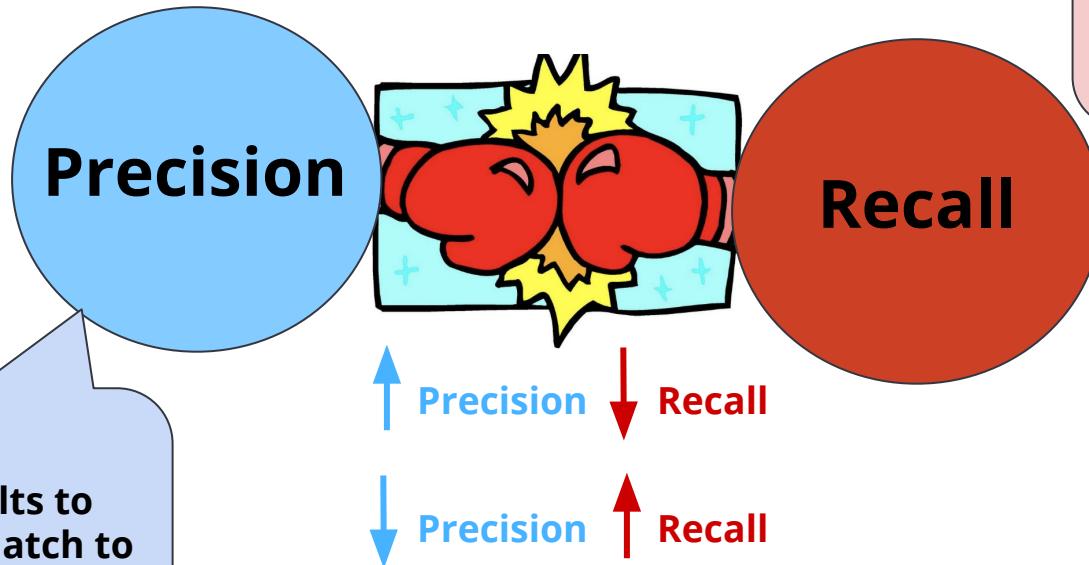
What is recall?



$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

What portion of relevant data is being returned as search results?

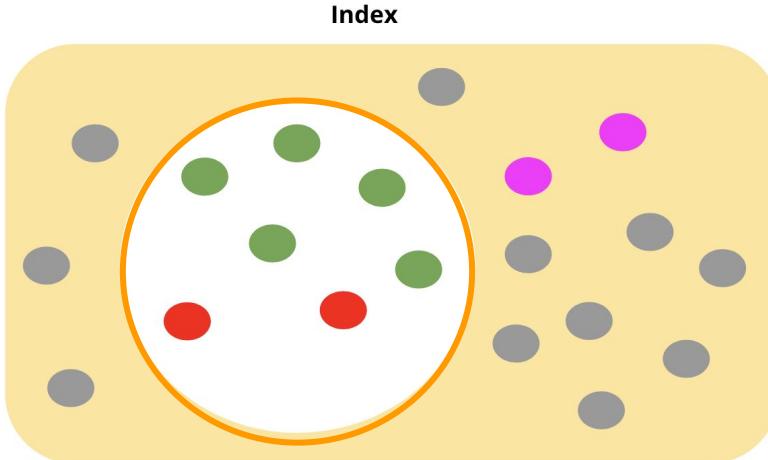
Precision and Recall are inversely related



I want all the retrieved results to be a perfect match to the query, even if it means returning less documents.

I want to retrieve more results even if documents may not be a perfect match to the query.

Precision and recall determine which documents are included in the search results.



Elastic Dev Tools

Console Search Profiler Grok Debugger Painless Lab BETA

History Settings Help

```
1 GET kibana_sample_data_ecommerce/_search
2
3
4
5
6
7
8
9
10 "hits" : {
11   "total" : {
12     "value" : 4675,
13     "relation" : "eq"
14   },
15   "max_score" : 1.0,
16   "hits" : [
17     {
18       "_index" :
19         "kibana_sample_data_ecommerce",
20       "_type" : ".doc",
21       "_id" : "79WD1XYBy9gvFwLxZogX",
22       "_score" : 1.0,
23       "_source" : {
24         "category" : [
25           "Men's Clothing"
26         ],
27         "currency" : "EUR",
28         "customer_first_name" : "Eddie",
29         "customer_full_name" : "Eddie Underwood"
30       }
31     }
32   ]
33 }
```

Precision and recall do not determine which of the returned documents are more relevant compared to the other!

Ranking refers to ordering of the results (from most relevant results at the top, to least relevant at the bottom).

The screenshot shows a search interface with a light gray header bar containing a close button (X), a home icon, and three colored dots (green, yellow, red). Below the header is a search bar with a magnifying glass icon and the text "How to form good habits". To the right of the search bar is another close button (X). The main content area is divided into two columns by a vertical line. The left column is labeled "Most Relevant" and contains three ellipsis ("...") entries. The right column is labeled "(Highest Score)". The right column is labeled "Less Relevant" and contains three ellipsis ("...") entries. The left column is labeled "Least Relevant" and contains three ellipsis ("...") entries. The right column is labeled "(Lowest Score)".

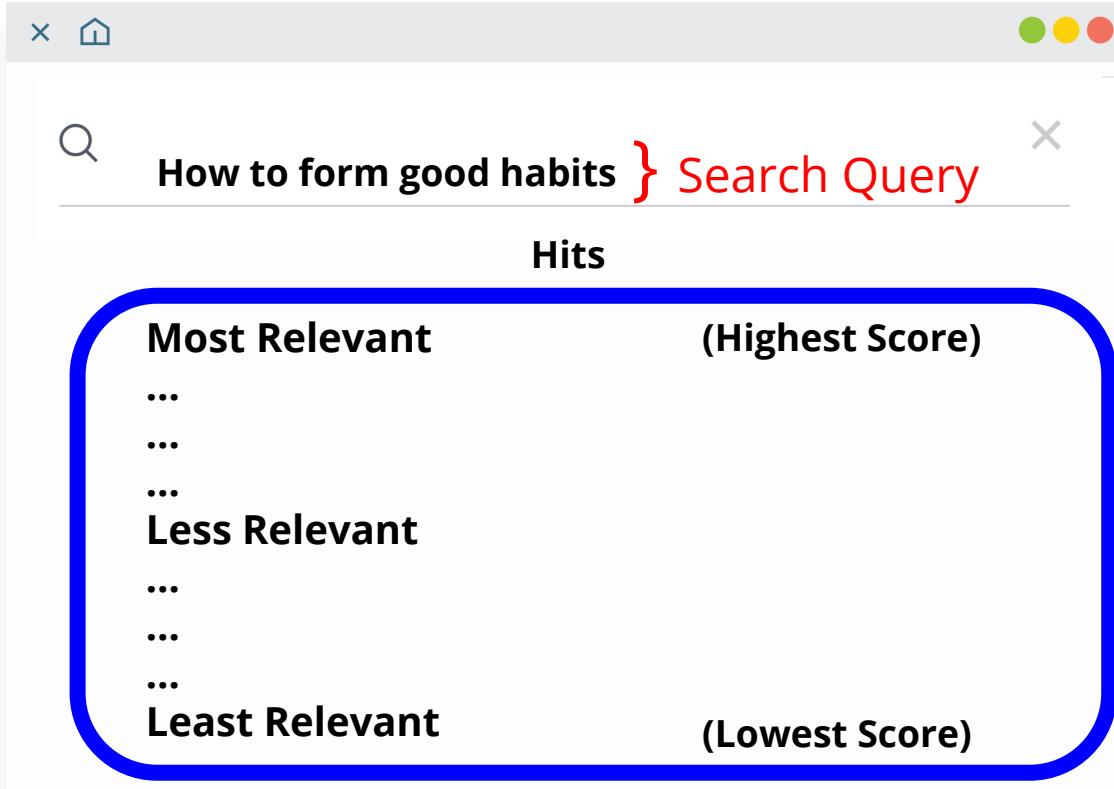
What is score?

- The score is a value that represents how relevant a document is to that specific query
- A score is computed for each document that is a hit

What is score?

- Term Frequency(TF)
- Inverse Document Frequency(IDF)

What is score?



Term Frequency(TF) determines how many times each search term appears in a document.

The screenshot shows a search interface with a navigation bar and a search bar. The search bar contains the query "How to form good habits". The word "habits" is highlighted with a red box. Below the search bar, the results are displayed as JSON documents. The first result, "Atomic Habits", has a red box around its description, which includes the text "No matter your goals, Atomic Habits offers a proven framework for improving every day. James clear, ... habits...habits ...habits". The text "TF= 4" is also highlighted with a red box. The second result, "The Mental Toughness Handbook", has a blue box around its description, which includes the text "Imagine boldly facing any challenge that comes your way... 5 daily habits you must embrace to strengthen your mind and harden your resolve. Why willpower and motivation are unreliable...". The text "TF= 1" is also highlighted with a blue box.

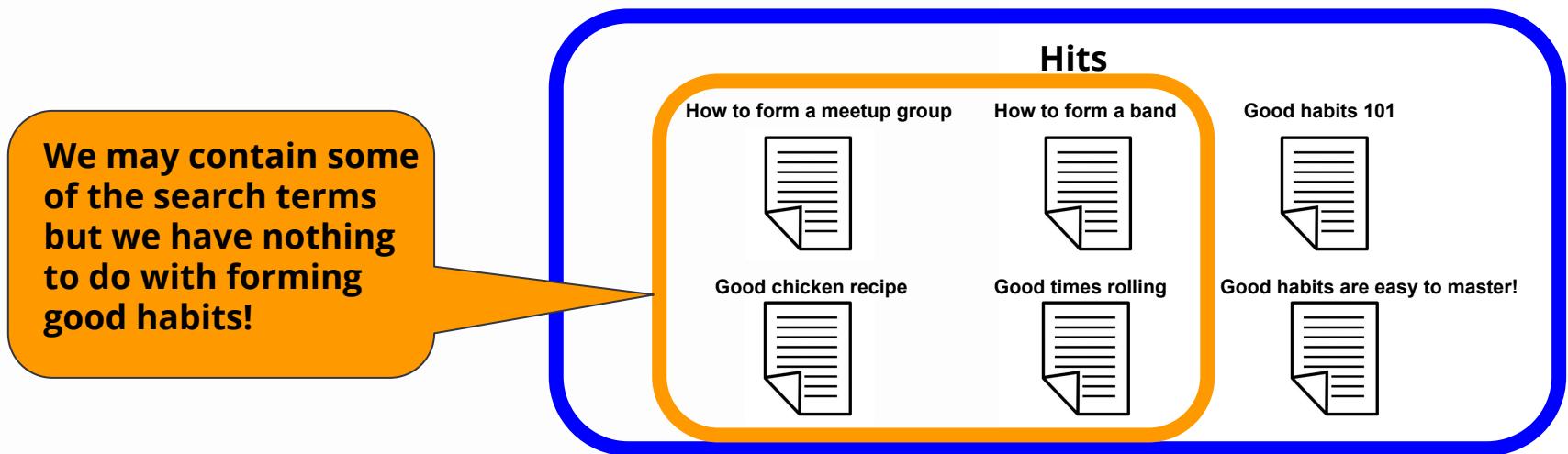
```
{  
  "title": "Atomic Habits",  
  "author": "James Clear",  
  "category": "self-help",  
  "description": "No matter your goals, Atomic Habits offers a proven framework for improving every day. James clear, ... habits...habits ...habits" TF= 4  
}  
  
{  
  "title": "The Mental Toughness Handbook",  
  "author": "Damon Zahariades",  
  "category": "self-help",  
  "description": "Imagine boldly facing any challenge that comes your way... 5 daily habits you must embrace to strengthen your mind and harden your resolve. Why willpower and motivation are unreliable..." TF= 1  
}
```

If search terms are found in high frequency in a document, the document is considered more relevant to the search query.

What is Inverse Document Frequency(IDF)?



IDF diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely!



Fine tuning precision or recall using Elasticsearch and Kibana



Click on the link to the workshop repo.

<https://ela.st/vancouver-workshop-2>

Scroll down to the Resources section & click on Instructions for downloading Elasticsearch & Kibana

Beginner's Crash Course to the Elastic Stack Series

[Vancouver] Part 2: Understanding the relevance of your search with Elasticsearch and Kibana

Welcome to the Beginner's Crash Course to the Elastic Stack!

This repo contains all resources shared during Part 2: Understanding the relevance of your search with Elasticsearch and Kibana.

Resources

[Free Elastic Cloud Trial](#)

[Instructions for downloading Elasticsearch and Kibana](#)

[Presentation](#)

[Dataset from Kaggle used for tutorial](#)

[Elastic Vancouver Chapter](#) Want to attend live workshops? Join the Elastic Vancouver Chapter to get the deets!

Search for information

There are two main ways to search in Elasticsearch:

1. Queries
2. Aggregations



Click on Upload a file option

The screenshot shows the Elastic Home page. At the top, there's a navigation bar with the Elastic logo, a search bar, and user icons. Below the navigation is a main content area titled "Home". This area features three cards on the left: "Enterprise Search" (green), "Observability" (pink), and "Security" (dark grey). On the right, there's a large card for "Kibana" (blue) with a sub-section for "Visualize & analyze". At the bottom, there's a section titled "Ingest your data" with three options: "Add data", "Add Elastic Agent", and "Upload a file". The "Upload a file" option is highlighted with a red box. A "Try our sample data" link is located above the "Upload a file" box.

Elastic

Search Elastic

Home

Add data Manage Dev tools

Enterprise Search
Search everything →

Build a powerful search experience.
Connect your users to relevant data.
Unify your team content.

Observability
Centralize & monitor →

Monitor infrastructure metrics.
Trace application requests.
Measure SLAs and react to issues.

Security
SIEM & Endpoint Security →

Prevent threats autonomously.
Detect and respond.
Investigate incidents.

Kibana
Visualize & analyze →

Analyze data in dashboards.
Search and find insights.
Design pixel-perfect presentations.
Plot geographic data.
Model, predict, and detect.
Reveal patterns and relationships.

Ingest your data

Add data
Ingest data from popular apps and services.

Add Elastic Agent
Add and manage your fleet of Elastic Agents and integrations.

Upload a file
Import your own CSV, NDJSON, or log file.

Try our sample data

Download and unzip News Category Dataset from Kaggle

The screenshot shows the Kaggle website displaying the 'News Category Dataset'. The dataset was created by Rishabh Misra and updated 2 years ago (Version 2). It contains 802,372 records and is available in JSON format (80.03 MB). The dataset is licensed under CC0: Public Domain and is tagged with news, nlp, classification, deep learning, and linguistics. The 'Data Explorer' section shows the file 'News_Category_Dataset_v2.json' (80.03 MB) with an 'About this file' summary.

Dataset

News Category Dataset

Identify the type of news based on headlines and short descriptions

Rishabh Misra • updated 2 years ago (Version 2)

Data Tasks Notebooks (64) Discussion (3) Activity Metadata Download (25 MB) New Notebook

Usability 10.0 License CC0: Public Domain Tags news, nlp, classification, deep learning, linguistics

Description

Context

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from [HuffPost](#). The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

Content

Each news headline has a corresponding category. Categories and corresponding article counts are as follows:

- POLITICS: 32739
- ENTERTAINMENT: 17200
- TECHNOLOGY: 15000
- SCIENCE: 13000
- ARTS: 10000
- SOCIETY: 9000
- SPORTS: 8000
- WORLD: 7000
- OPINION: 6000
- SCIENCE & TECHNOLOGY: 5000
- ARTS & ENTERTAINMENT: 4000
- TECH & SCIENCE: 3000
- WORLD & POLITICS: 2000
- OPINION & SCIENCE: 1000

Data Explorer

80.03 MB

{ } News_Category_Dataset_v2...

< **News_Category_Dataset_v2.json** (80.03 MB)

About this file

The file contains 802,372 records. Each json record contains following attributes:

category : Category article belongs to
headline : Headline of the article
authors : Person authored the article

Drag and drop a file you want to upload.

Elastic

Machine Learning / Data Visualizer / File

Overview Anomaly Detection Data Frame Analytics Data Visualizer Settings

Visualize data from a log file EXPERIMENTAL

The File Data Visualizer helps you understand the fields and metrics in a log file. Upload your file, analyze its data, and then choose whether to import the data into an Elasticsearch index.

The File Data Visualizer supports these file formats:

- Delimited text files, such as CSV and TSV
- Newline-delimited JSON
- Log files with a common format for the timestamp

You can upload files up to 100 MB.

This feature is experimental. Got feedback? Please create an issue in [GitHub](#).

Select or drag and drop a file

Kibana will give you an analysis of the first 1000 lines of your data and give you a summary of your dataset.

Elastic

Machine Learning / Data Visualizer / File

Overview Anomaly Detection Data Frame Analytics Data Visualizer Settings

News_Category_Dataset_v2.json

File contents

First 1,000 lines

```
1 {"category": "CRIME", "headline": "There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV", "authors": "Melissa Jeltsen", "link": "https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89", "short_description": "She left her husband. He killed their children. Just another day in America." "date": "2018-05-26"}  
2 {"category": "ENTERTAINMENT", "headline": "Will Smith Joins Diplo And Nicky Jam For The 2018 World Cup's Official Song", "authors": "Andy McDonald", "link": "https://www.huffingtonpost.com/entry/will-smith-joins-diplo-and-nicky-jam-for-the-official-2018-world-cup-song_us_5b09726fe4b0fdb2aa541201", "short_description": "Of course it ", "date": "2018-05-26"}  
3 {"category": "ENTERTAINMENT", "headline": "Hugh Grant Marries For The First Time At Age 57", "authors": "Ron Dicker", "link": "https://www.huffingtonpost.com/entry/hugh-marries_us_5b09212c4b0568a880b98c", "short_description": "The actor and his longtime girlfriend Anna Eberstein tied the knot in a civil ceremony.", "date": "2018-05-26"}  
4 {"category": "ENTERTAINMENT", "headline": "Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork", "authors": "Ron Dicker", "link": "https://www.huffingt /entry/jim-carrey-adam-schiff-democrats_us_5b0950e8e4b0fdb2aa53e675", "short_description": "The actor gives Dems an ass-kicking for not fighting hard enough against ", "date": "2018-05-26"}  
5 {"category": "ENTERTAINMENT", "headline": "Julianne Margulies Uses Donald Trump Poop Bags To Pick Up After Her Dog", "authors": "Ron Dicker", "link": "https://www.huffir /entry/julianne-margulies-takes-down-his-trump-poop-bags-to-pick-up-after-her-dog_us_5b0950e8e4b0fdb2aa53e678", "short_description": "The "Nightline" actress told us in a recently published interview that she uses Trump's dog bags to pick up after her dog because they're better than regular ones." "date": "2018-05-26"}
```

Summary

Number of lines analyzed 1000

Format ndjson

Time field date

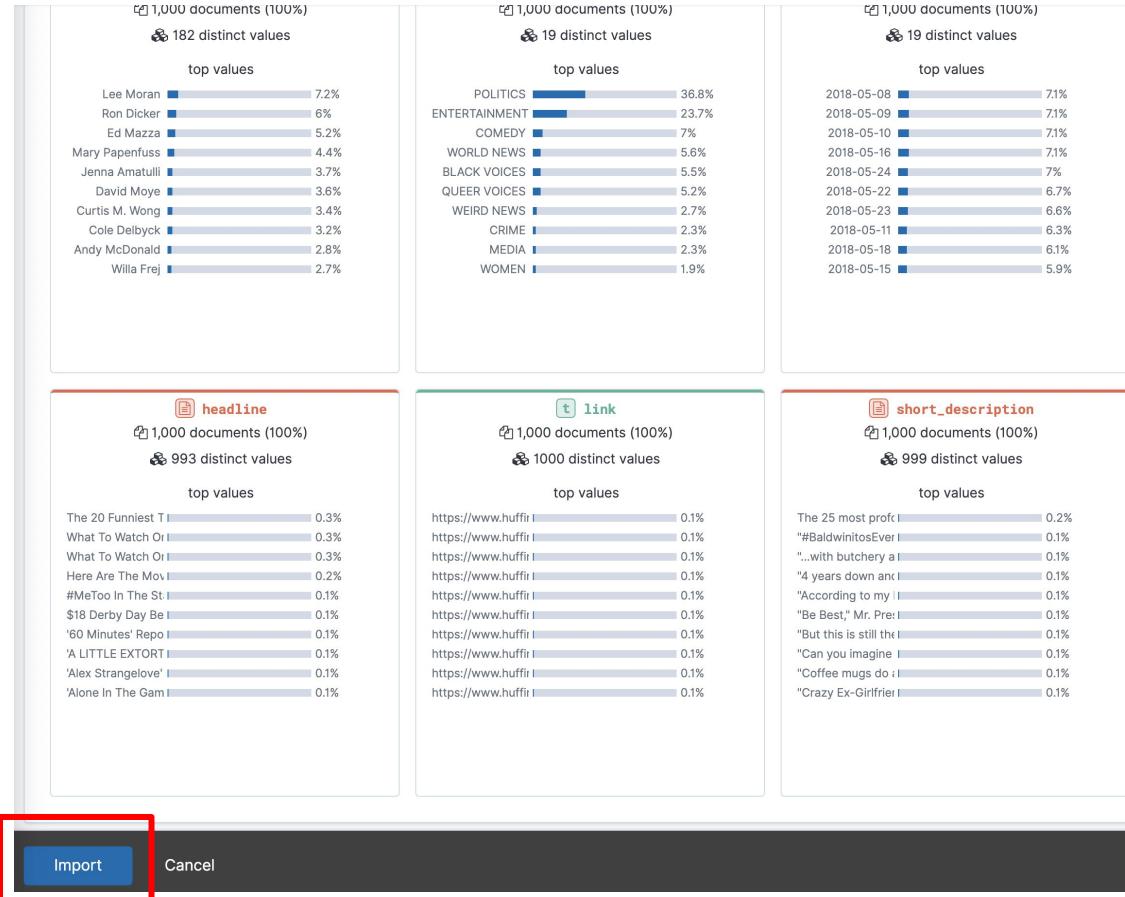
Time format ISO8601

Override settings Analysis explanation

Field section displays fields identified, high level statistics, and top occurring values



Click on import button



Name your index and click on import.

The screenshot shows the Elastic Data Visualizer interface. At the top, there is a navigation bar with the Elastic logo, a search bar labeled "Search Elastic", and several icons. Below the navigation bar, the breadcrumb navigation shows "Machine Learning / Data Visualizer / File". The main menu includes "Overview", "Anomaly Detection", "Data Frame Analytics", "Data Visualizer" (which is underlined, indicating it is the active tab), and "Settings".

The main content area is titled "News_Category_Dataset_v2.json". It contains a form for importing data:

- A title "Import data" with an "EXPERIMENTAL" badge.
- A "Simple" tab selected, with an "Advanced" tab option.
- An "Index name" field containing "news_headlines".
- A checked checkbox for "Create index pattern".
- A blue "Import" button, which is highlighted with a red rectangular border.

Then Elasticsearch will take of the rest!

The screenshot shows the Elasticsearch Data Visualizer interface. At the top, there's a navigation bar with icons for Elastic, a search bar, and various settings. Below the navigation, a breadcrumb trail shows 'Machine Learning / Data Visualizer / File'. The main menu includes 'Overview', 'Anomaly Detection', 'Data Frame Analytics', 'Data Visualizer' (which is underlined, indicating it's the active tab), and 'Settings'.

The current page title is 'News_Category_Dataset_v2.json'. In the main content area, there's a section titled 'Import data' with an 'EXPERIMENTAL' badge. It has two tabs: 'Simple' (which is selected) and 'Advanced'. The 'Index name' field contains 'news_headlines'. A checked checkbox labeled 'Create index pattern' is present. A 'Reset' button is at the bottom of this section.

Below this, a timeline shows five steps: 'File processed', 'Index created', 'Ingest pipeline created', 'Data uploaded', and 'Index pattern created', each marked with a checkmark icon. Underneath the timeline, a summary table provides details about the import:

✓ Import complete	
Index	news_articles
Index pattern	news_articles
Ingest pipeline	news_articles-pipeline
Documents ingested	200853

At the bottom of the page are 'Back' and 'Cancel' buttons.

Click on menu icon, and open Dev Tools.

The screenshot shows the Elastic Stack interface with the following details:

- Header:** Elastic
- Search Bar:** Search Elastic
- Top Navigation:** Home, Data Visualizer (selected), Settings
- Left Sidebar:**
 - Recently viewed:** No recently viewed items
 - Observability:** Overview, Logs, Metrics, APM, Uptime, User Experience
 - Security:** Overview, Detections, Hosts, Network, Timelines, Cases, Administration
 - Management:** Dev Tools (selected), Fleet, Stack Monitoring, Stack Management
- Main Content:** A timeline showing four events:
 - Index created
 - Ingest pipeline created
 - Data uploaded
 - Index pattern created
- Bottom:** A footer bar with several icons.

Click on dismiss and delete the default query.

The screenshot shows the Elasticsearch Dev Tools interface. On the left, there's a navigation bar with 'Console' selected, along with 'Search Profiler', 'Grok Debugger', and 'Painless Lab'. Below the navigation is a toolbar with 'History', 'Settings', and 'Help'. A search bar at the top right contains the placeholder 'Search Elastic'. On the far right, there are three small icons: a gear, a magnifying glass, and a red circle with a number '1'.

The main area is split into two panes. The left pane is an editor where a user has typed the following Elasticsearch query:

```
GET _search
{
  "query": {
    "match_all": {}
  }
}
```

This query is highlighted with a blue rectangular box. The right pane is a response viewer showing the results of the query. At the top of the right pane, a modal window titled 'Welcome to Console' is displayed. The modal contains the following text:

Quick intro to the UI
The Console UI is split into two panes: an editor pane (left) and a response pane (right). Use the editor to type requests and submit them to Elasticsearch. The results will be displayed in the response pane on the right side.

Console understands requests in a compact format, similar to cURL:

```
1 # index a doc
2 PUT index/_doc/1
3 {
4   "body": "here"
5 }
6
7 # and get it ...
8 GET index/_doc/1
```

While typing a request, Console will make suggestions which you can then accept by hitting Enter/Tab. These suggestions are made based on the request structure as well as your indices and types.

A few quick tips, while I have your attention

- Submit requests to ES using the green triangle button.
- Use the wrench menu for other useful things.
- You can paste requests in cURL format

At the bottom right of the modal, there is a blue 'Dismiss' button.

Fine tuning precision or recall using Elasticsearch and Kibana



Questions?



Join us for the part 3 of the workshop series!



Beginner's Crash Course to Elastic Stack Part 3

[Austin](#)

Wed, Feb 24, 12:00 PM (CST)

13 RSVP'ed



RSVP for this event now!

[RSVP](#)

Availability ends February 24th

[RSVP](#)





Lisa Jung

Developer Advocate @Elastic

Discussion forum: <https://discuss.elastic.co/>

Blog: <https://dev.to/lisahjung>

Twitter: @LisaHJung

