



Newspaper Navigator: Hosting the Dataset and Deploying the Search Application

Chris Adams & Benjamin Charles Germain Lee

LIBRARY
OF CONGRESS
LABS

news-navigator.labs.loc.gov

PAUL G.
ALLEN
SCHOOL
W

LEADS HIS LEAGUE ON INITIAL SACK



Hal Chase.
Enough For Everybody Except



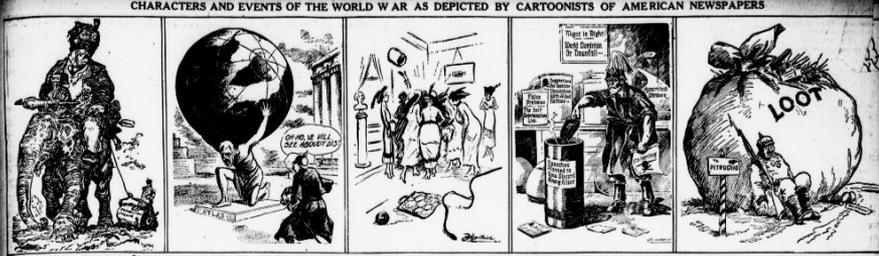
FOR UNCLE SAM
GOLF WAR THREATENED



REV. E. W. HARTTHORNE, D. D.
Nashville, Tenn., and ardent
Congress worker.



Red



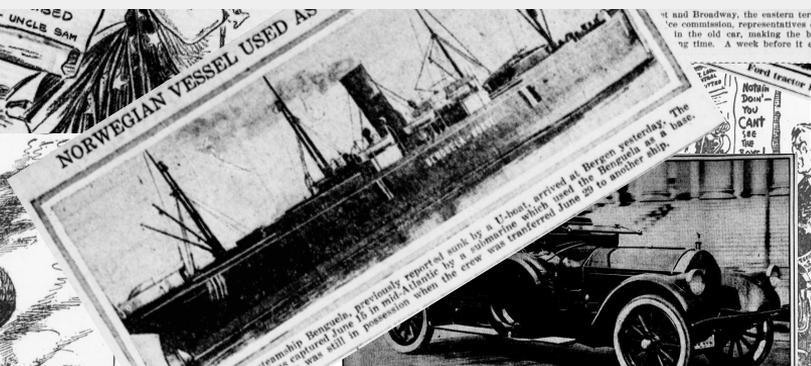
IN TAIN FOR THE ERCE... STERLING HIS STUFF... BROUGHT... THE GOOD NEWS ABOUT... THE BROWNING BAN... THE BROWNING CLUB... NEW YORK'S LAST HORSE CAR QUITS

Can we "throw open the treasure chest" by training a machine learning algorithm to process all of Chronicling America?



SCIENCE	600
INTELLIGENCE	750
MISCELLANEOUS	100
PHYSICS & CHEMISTRY	500
PHYSIOLOGY & MEDICINE	800
PSYCHOLOGY	100
NEW BOOKS	2843
NEW EDITIONS	1211
TOTAL FROM 1916	10,445

AMERICAN BOOK OUTPUT FOR 1917
to graphically exhibited by Fred E. Woodman.



NORWEGIAN VESSEL USED AS
steamship Benguela, previously removed sunk by a U.S. boat, arrived at Bergen yesterday. The captor was in possession when the Benguela was captured June 15 in mid-Atlantic by a submarine which used the Benguela as a base. The crew was transferred June 20 to another ship.

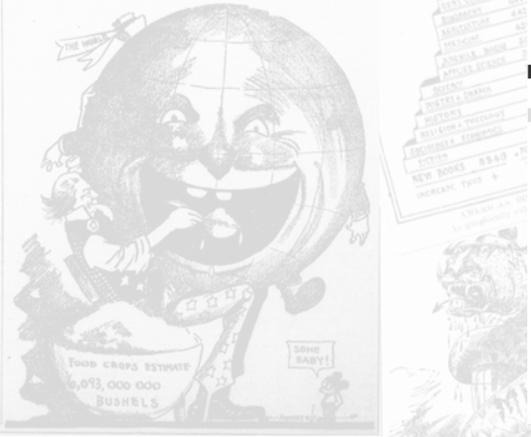


THE NATION'S GOING DRY.
WHITE—Dry. BLACK—Wet.
Number of prohibition states, 20. The District of Columbia is dry by act of congress. Alaska is dry by a vote of the people ratified by congress. The Porto Rican citizenship bill, passed by the Sixty-fourth congress, provides for prohibition in the island subject to a referendum. Including the dry territory in wet states, more than 87 per cent of the area of the United States and more than 90 per cent of the population are under prohibition.
Eight states are in submission campaigns. Two of these—Iowa and Utah—already have statutory prohibition and in November will vote to make it constitutional. New Mexico also will vote on the question next November; others in November, 1918. Ohio is in a wet-and-dry fight through the initiative.

Chronicling America



Enough For Everybody Except Germany



Search My Collection Train My AI Navigators Chronicling America LC Labs Data Archaeology About

Select an AI navigator: **baseball players** + New AI Navigator

Save (+) The AI navigator will try to find photos like the ones selected below!

(-) The AI navigator will try to avoid photos like the ones selected below!

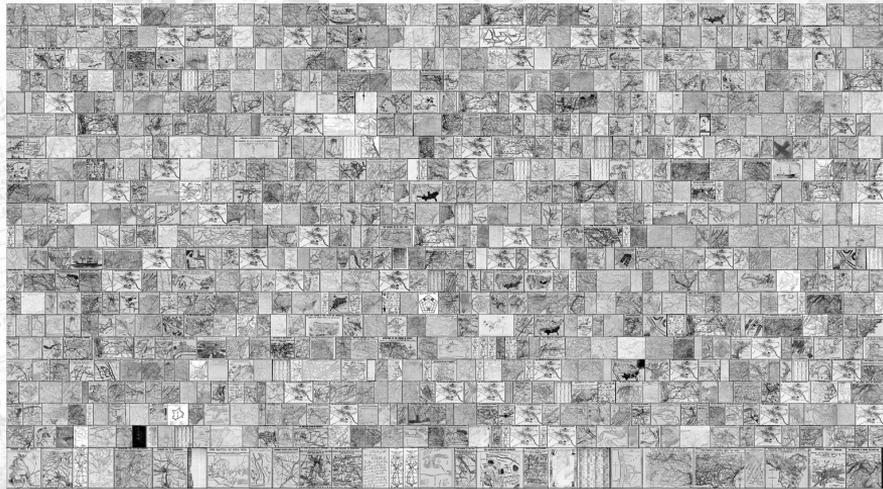
baseball players

Below, your AI navigator has found visually similar photos based on the selections on the left. You can re-train your AI navigator by adjusting your selections on the left and making new selections below.

...the last time it was a few years ago. However, the system requires the participation, representative of the in the old one, looking the system to finish. A week before it started



I. The Newspaper Navigator Dataset

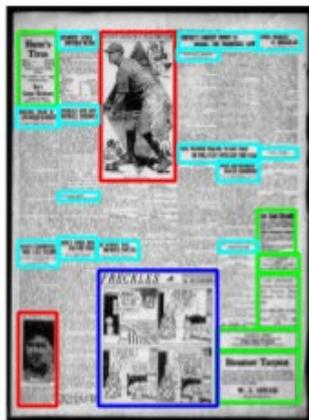


news-navigator.labs.loc.gov

Download Image,
METS/ALTO OCR



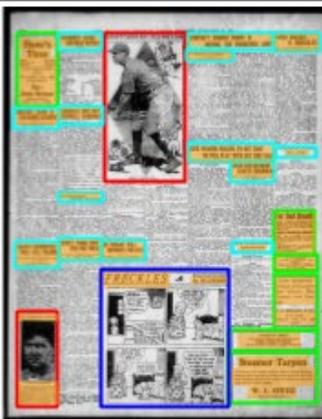
Run Visual Content
Recognition



Crop and Save
Visual Content



Generate
Image
Embeddings



Extract OCR within
Predicted Bounding
Boxes

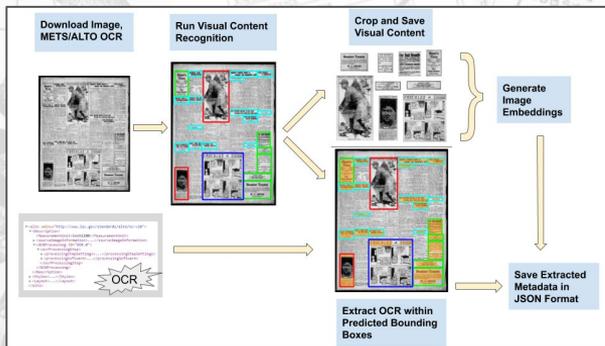
Save Extracted
Metadata in
JSON Format

```
<?xml version="1.0" encoding="UTF-8" ?>
<alio xmlns="http://www.loc.gov/standards/alto/ns-v2#"
  <description>
    <measurementUnit>inch1200//measurementUnit</>
    <sourceImageInformation>.../sourceImageInformation</>
    <ocrProcessingID>OCR-0</>
    <ocrProcessingStep>
      <processingStepSettings>.../processingStepSettings</>
      <processingSoftware>.../processingSoftware</>
    </ocrProcessingStep>
  </alioProcessing>
  <description>
    <style>.../style>
  </description>
</alio>
```

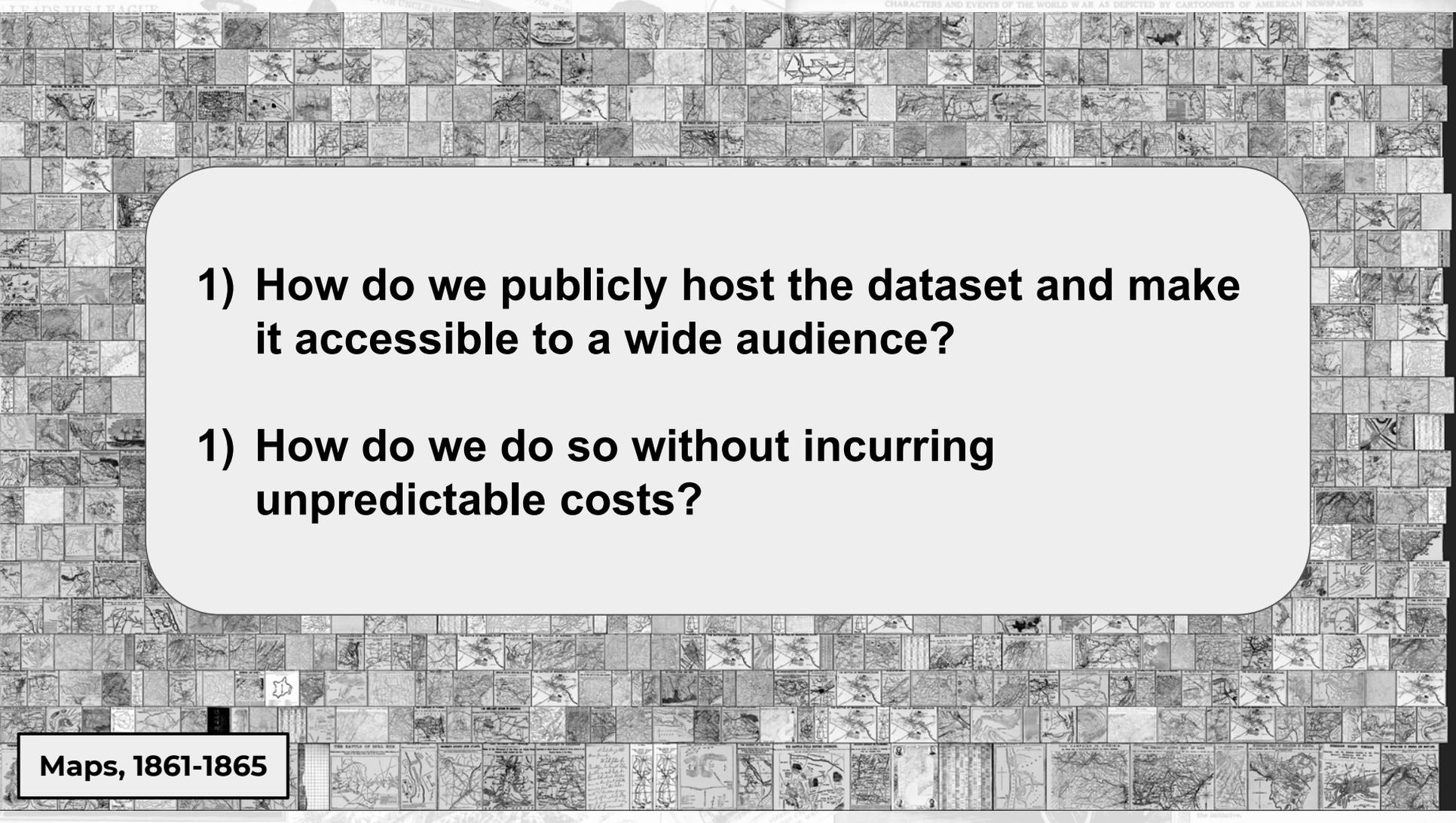
OCR

Running the Pipeline

- 16,368,041 *Chronicling America* pages processed (99.998%)
 - ~100 TB of image & XML data
- ~19 days on 2 g4dn.12xlarge Amazon AWS EC2 instances
 - 96 vCPUs & 8 NVIDIA T4 GPUs



x 16.3 million pages



1) How do we publicly host the dataset and make it accessible to a wide audience?

1) How do we do so without incurring unpredictable costs?

Maps, 1861-1865

news-navigator.labs.loc.gov

- 2 primary forms of access:
 - For computation: S3 requests under the “requester pays” model
 - For public use: prepared datasets (no coding or S3 necessary)

LIBRARY OF CONGRESS LABS

About Work Events LC for Robots

Experiments

Newspaper Navigator

Explore the visual and textual content within the Chronising America digitized newspaper collection in new ways using machine learning!

Overview

Welcome to the Newspaper Navigator dataset! This dataset consists of extracted visual content for 16,358,041 historic newspaper pages in Chronising America. The visual content was identified using an object detection model trained on annotations of World War I-era Chronising America pages, including annotations made by volunteers as part of the Beyond Words crowdsourcing project. The resulting visual content recognition model detects the following types of visual content:

1. Photograph
2. Illustration
3. Map
4. Comic/Caricoon
5. Editorial Cartoon
6. headline
7. Advertisement

The dataset also includes text corresponding to the visual content, identified by extracting the Optical Character Recognition, or OCR, within each predicted bounding box. For example, if the visual content recognition model predicted a bounding box around a headline, the corresponding textual content provides a machine-readable version of the headline. Likewise, for a photograph, illustration, or map, this textual representation often contains the title and caption.

II. The Newspaper Navigator Search Application

The screenshot shows the web interface of the 'news-navigator' application. At the top, there is a navigation bar with links for 'Search', 'My Collection', 'Train My AI Navigators', 'Chronicle America', 'LC Labs', 'Data Archaeology', and 'About'. Below this, the main content area is divided into several sections:

- Select an AI navigator:** A dropdown menu currently shows 'baseball players'. Below it is a '+ New AI Navigator' button.
- Save (+) The AI navigator will try to find photos like the ones selected below!**: A grid of 12 baseball player photos, each with a small '+' icon in the bottom right corner, indicating they are selected for training.
- Train Your AI Navigator!**: A button to initiate the training process.
- (-) The AI navigator will try to avoid photos like the ones selected below!**: A grid of 12 photos, mostly of men in military-style uniforms, each with a small '-' icon in the bottom right corner, indicating they are selected to be avoided.
- baseball players**: A search input field containing the text 'baseball players'. Below it, a grid of 24 baseball player photos is displayed, representing the results of the search.

news-navigator.labs.loc.gov/search

Newspaper Navigator Search Application

- A search user interface for 1.5 million photos from the Newspaper Navigator dataset
- Empowers users to train open facets (“AI navigators”) to retrieve visually similar content
- Exposes the machine learning training to the user
- Training & predicting on all 1.5 million photos is nearly instantaneous

The screenshot displays the Newspaper Navigator Search Application interface. At the top, there are navigation tabs: "Search", "My Collection", "Train My AI Navigators", "Chronicle America", "LC Labs", "Data Archaeology", and "About".

The main content area is titled "baseball players". Below the title, there is a section for "Save" with the text: "(+) The AI navigator will try to find photos like the ones selected below!". This section shows a grid of 10 baseball player photos, with several highlighted in green.

Below this, there is a section for "New AI Navigator" with the text: "(-) The AI navigator will try to avoid photos like the ones selected below!". This section shows a grid of 10 photos, with several highlighted in blue.

On the right side of the interface, there is a section titled "baseball players" with the text: "Below, your AI navigator has found visually similar photos based on the selections on the left. You can re-train your AI navigator by adjusting your selections on the left and making new selections below." This section shows a grid of 20 baseball player photos.

At the bottom right, there is a map of the United States with the text: "THE NATION'S GOING DRY. WHITE—Yes. BLACK—No." and "THE FUTURE OF COLUMBIA IS DRY BY ACT OF CONGRESS. Alaska to dry by a vote of the people invited by congress. The Pure Food and Drug Bill, passed by the sixty-fourth congress, provides for prohibition in the broad subject to a referendum. Including the dry territory in wet states, more than 80 per cent of the area of the United States and more than 90 per cent of the population are under prohibition. Eight states are in prohibition companies. Two of these—Vermont and Utah—already have statutory prohibition and in November will vote to make it constitutional. New Mexico also will vote on the question next November; others in November, 1915. This is a nationwide fight through the initiative."

Search My Collection

Train My AI Navigators

Chronicling America

LC Labs

Data Archaeology

About

Select an AI navigator:

baseball players

+ New AI Navigator

Save

(+) The AI navigator will try to find photos like the ones selected below!



Train Your AI Navigator!

(-) The AI navigator will try to avoid photos like the ones selected below!



baseball players

Below, your AI navigator has found visually similar photos based on the selections on the left. You can re-train your AI navigator by adjusting your selections on the left and making new selections below.



<https://news-navigator.labs.loc.gov/search>

BLACK—West of Columbia is dry by...

Challenges Surrounding the Search Application's Deployment

- How do we create a sustainable solution for deploying the search application with low maintenance requirements?
- How do we handle issues surrounding load at launch without over-paying the rest of the time?
- How do we ensure that the interactive machine learning component is fast and is not a challenge for upkeep?

Search My Collection Train My AI Navigators Chroniling America IC Labs Data Archaeology About

Select an AI navigator:

baseball players

Save

(+) The AI navigator will try to find photos like the ones selected below!

baseball players

baseball players

(-) The AI navigator will try to avoid photos like the ones selected below!

baseball players

baseball players

Solution: Containerization & AWS Fargate!

- Containers are designed to be ephemeral, with no state to preserve except where it's explicitly configured
- Automated container build process (see GitHub repo), making it easy to ship security updates and for new people to use the project
- We deployed the container on AWS ECS Fargate: no host servers or configuration to manage, auto-scaling policy adds and removes instances based on load or failures, health checks avoid accidentally deploying a broken build

The screenshot displays the 'Train My AI Navigators' web interface. At the top, there are navigation links: 'Search', 'My Collection', 'Train My AI Navigators', 'Chroniling America', 'LC Labs', 'Data Archaeology', and 'About'. The main content area is titled 'baseball players'. Below the title, there is a section for 'Save' with instructions: '(+) The AI navigator will try to find photos like the ones selected below!'. This section shows a grid of images with several baseball players highlighted in green. Below this is a section for 'New AI Navigator' with instructions: '(-) The AI navigator will try to avoid photos like the ones selected below!'. This section shows a grid of images with several men in suits highlighted in blue. The interface also includes a search bar and a 'New AI Navigator' button.

18 Months since Deployment...

- Running continuously for a year and a half with virtually no issues!
- Supported tens of thousands of user sessions to date
- Engagement from scholars, teachers, students, genealogists, and members of the American public
- More usage statistics are forthcoming in a *Signal* blog post with LC Labs!

The screenshot displays the 'AI Navigator' interface. At the top, there are navigation tabs: 'Search', 'My Collection', 'Train My AI Navigators', 'Chroniling America', 'LC Labs', 'Data Archaeology', and 'About'. The main content area is titled 'baseball players' and includes a 'Save' button. Below the title, there are two sections: one for 'Select an AI navigator:' and another for 'New AI Navigator:'. The 'Select an AI navigator:' section shows a grid of images with green boxes around them, indicating they are selected. The 'New AI Navigator:' section shows a grid of images with blue boxes around them, indicating they are to be avoided. The interface also includes a search bar and a 'Train My AI Navigators' button.

Newspaper Navigator Resources

- Dataset: <https://news-navigator.labs.loc.gov/>
- Search Application: <https://news-navigator.labs.loc.gov/search>
- GitHub repository: <https://github.com/LibraryOfCongress/newspaper-navigator>

Acknowledgments

We would like to thank:

- Jaime Mears, Eileen Jakeway, Meghan Ferriter, Laurie Allen, Abbey Potter, Leah Weinryb Grohsgal, Kate Zwaard, Deb Thomas, Nathan Yarasavage, Robin Butterhof, Tong Wang, John Foley, Megan Halsband, Brian Foo, Mark Sweeney, the entire NDNP staff at the Library of Congress, and everyone who has made *Chronicling America + Beyond Words* possible
- Dan Weld at the University of Washington

Acknowledgments

- This research is based upon work supported by the **National Science Foundation Graduate Research Fellowship** under **Grant DGE-1762114** and the **Library of Congress Innovator in Residence** position.

LIBRARY
OF CONGRESS
LABS



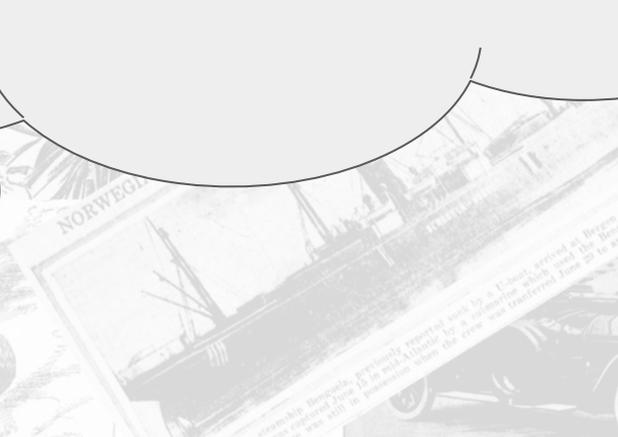
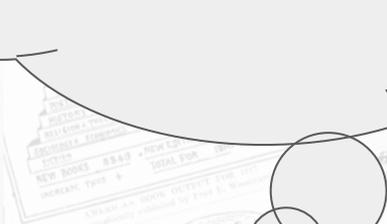
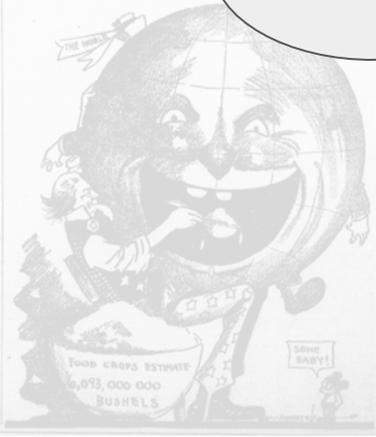
LEADS HIS LEAGUE ON INITIAL SACK

CHARACTERS AND EVENTS OF THE WORLD WAR AS DEPICTED BY CARTOONISTS OF AMERICAN NEWSPAPERS



Questions?

Hal Chase.
Enough For Everybody.



THE NATION'S GOING DRY.
WHITE—Dry. BLACK—Wet.
Number of prohibition states, 23. The District of Columbia is dry by act of congress. Alaska is dry by a vote of the people passed by congress. The Four States comprising bill, passed by the sixty-fourth congress, provide for prohibition in the island subject to a referendum. Including the dry territory in wet states, more than 32 per cent of the area of the United States and more than 50 per cent of the population are under prohibition.
Eight states are in submission campaigns. Two of these—Iowa and Utah—already have statutory prohibition and in November will vote to make it constitutional. New Mexico will vote on the question next November; others in November, 1915. This is a wet-and-dry fight through the initiative.