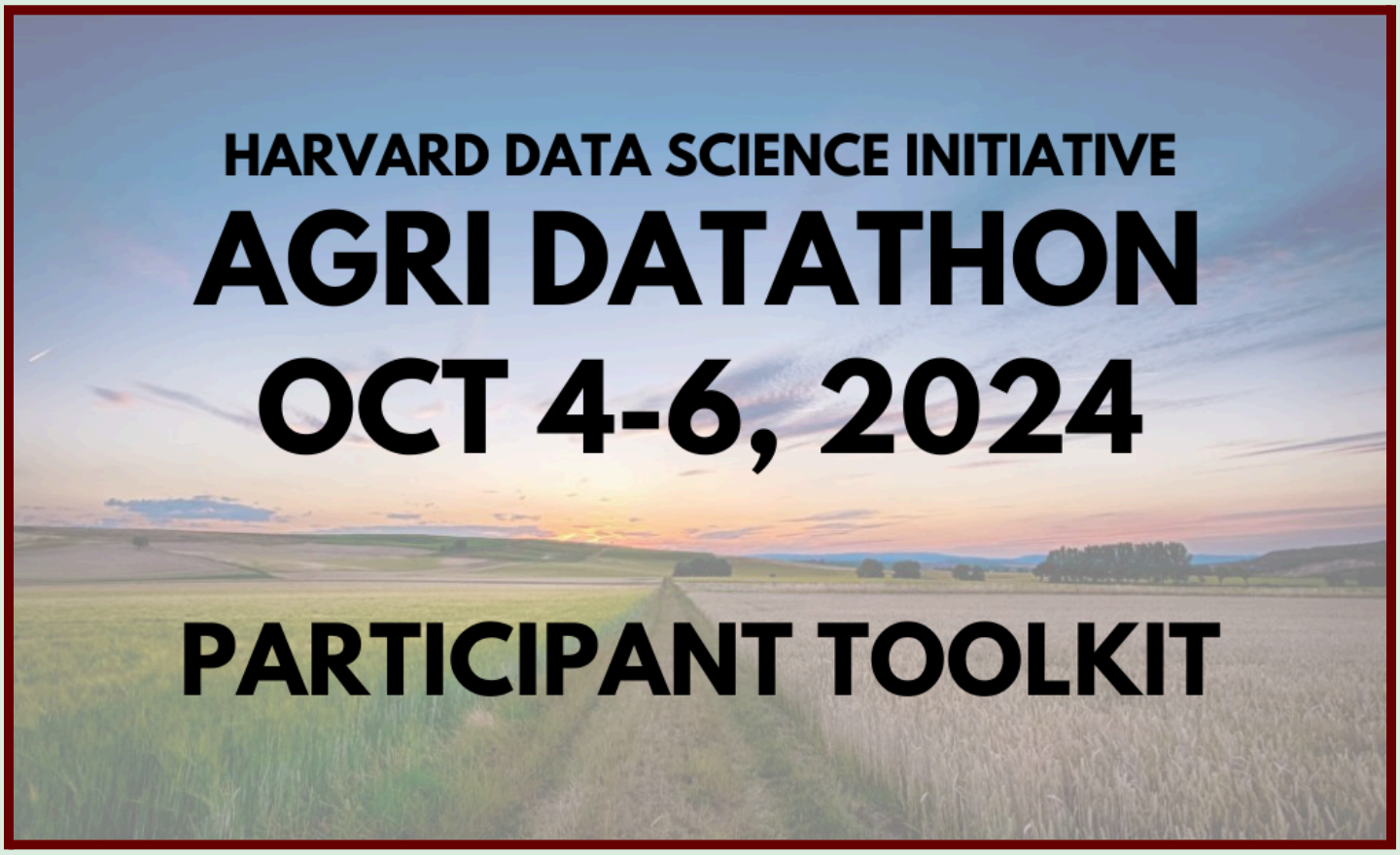


Welcome to the participant toolkit for the HDSI Agri Datathon! This is a resource that has been compiled by your organizers to help you know what to expect before, during, and after the HDSI Agri Datathon.

Have a question not answered here? Reach the organizers on Slack.

Final Submissions are now closed

CONTENTS

GENERAL INFORMATION

Eligibility

Participation in the HDSI Agri Datathon is open to current undergraduate and graduate (pre-doctoral) students with Python or C++ programming skills and a background in statistics or machine learning methodologies. Participants should be interested in the intersection of data, agriculture, food security, and/or climate. However, those without prior experience in these areas are still encouraged to participate.

Students from any college or university are eligible to participate, as long as they can attend the in-person kick-off at Harvard University on October 4th, 2024. After October 4th, the datathon will continue virtually until project submission at 11:59pm on October 6th. Note that funds for travel or overnight accommodations **are not** available.

We recommend that participants possess the following competencies:

- Data acquisition and management
- Exploratory data analysis (EDA)
- Predictive modeling or statistical learning
- Data visualization and result communication with LaTeX

Are you ready? Take this quiz to find out!

Registration

Participants should fill out the registration form here. Registration is mandatory and limited. Participants may register as a team (max 3) or as individuals. Individuals will have an opportunity to match to teams during the Oct. 4th kick-off. All participants will be notified of their acceptance within 2 weeks of registration.

Kick-Off Agenda – October 4th

In person attendance is required.

Location:
Winokur Family Hall, 1.321
Science and Engineering Complex
Harvard University
150 Western Ave
Allston, MA 02134

Enter closest to Academic Way
Google Maps: https://maps.app.goo.gl/S74138g4E1BTEk7QA
what3words: https://w3w.co/rubble.think.wallet

Getting to kick-off: Harvard Transportation

Schedule:

9:00 AM	Welcome from HDSI and USDA NASS
9:10 AM	Accessing the Agri Datathon data sources
9:40 AM	Introducing the Competition Prompts
10:00 AM	Team Formation Exercise
10:30 AM	Team Meetings - Prompt Selections
11:30 AM	**Deadline - Register Teams and Team Prompt Selection**
11:30 AM	Accessing Mentors
11:45 AM	Plenary Session Ends - Datathon Begins!

Important Dates

- Registration opens **August 22nd**
- Registration closes **September 27th**
- In-person Kickoff and Competition begins **October 4th**
- Competition continues virtually **October 5th**
- Project submission **@ 11:59 pm October 6th**
- Judging results released on or before **October 24th**

Team Structure

Each team must consist of exactly three participants. If you are registering with an incomplete team (two participants), or as an individual, don't worry! We will have a team formation exercise during the Oct. 4th kick-off. You can also check out the Slack Workspace to meet other participants before kick-off.

Permitted Resources

Teams are allowed to use texts or internet resources in the technical and methodological development of their solution. However, discussion between teams regarding their chosen methodologies is not allowed.

Use of AI tools: Teams are permitted to use generative AI tools (e.g. ChatGPT, OpenAI Codex, Google's Bard, and Meta's LLaMA) in the technical development of your submission **as long as proper citation is included.** Teams should not use generative AI tools in the creation of their submission outputs (video and written report.)

Mentors: Assistance may only be sought from designated competition mentors. Any other form of external help is against the rules.

External Data: You should be able to solve challenges with the data provided. You are welcome to use other public sources of data, however, no extra judging consideration will be given for the use of external data. Also, be aware that mentors may not be able to advise on the use of that data. If you choose to use external data, **you must cite your sources appropriately**.

Locations

The HDSI Agri Datathon kicks-off in-person on the morning of Oct. 4th, 2024 at Harvard's Science and Engineering Complex (150 Western Ave, Allston, MA). **All participants must attend in-person.**

After kick-off, teams can continue work virtually or in-person. For those who would like to continue working in-person, we recommend the following spots on campus for those without a Harvard ID:

- Smith Campus Center, 8:00am-10:00pm
- Science Center, 8:00am-8:00pm

Communication - Slack

PARTICIPANT TO-DO: Join the Slack channel

Before, during, and after the HDSI Agri Datathon, **Slack will serve as the primary platform for communication between participants and the technical lead, organizing committee, and competition**

mentors. To ensure that all participants have access to the same information, all inquiries should be directed to the designated Slack channels rather than personal email addresses. All responses are visible to all participants.

Participants can access the Slack portal here. Within the HDSI Agri Datathon, you will find several dedicated channels:

- **# general** A channel for all general inquiries related to the competition and where the winners will be announced
- **# team-formation:** A channel for participants seeking to form or join a team.
- **# mentor-meetings:** Use this channel to ask questions to mentors and organizers on Oct. 5th-6th (Saturday and Sunday.)
    - From 9:00am to 5:00pm mentors will be online to answer content-specific and technical questions.
    - From 5:00pm to 10:00pm, the Slack channel will be monitored by the Agri Datathon's technical lead, Elaine Swanson, in case of urgent challenges experienced from teams.

It is important that teams monitor the general Slack channel in case any announcements from the organizing teams are made! We encourage all participants to make use of the Slack workspace to facilitate a smooth competition experience.

DATATHON EVENT SCHEDULE

Competition Timeline
Full agendas to be released in September.

Date:	Session:	What you'll do:
Friday, Oct. 4th	In-Person Kick-off	• Team formation, if you're still looking • Register your team • Receive the prompts and select your challenge • Group work and meet with mentors in person
Saturday, Oct. 5th	Virtual or In-Person Work	• Teams can continue work in-person (Harvard locations) or virtually • Virtual meetings with mentors
Sunday Oct. 6th	Submission Day	• Teams can continue work in-person (Harvard locations) or virtually • Virtual meetings with mentors • UPLOAD YOUR SUBMISSION BY 11:59 PM
Week of Oct. 21th	Winners Announced	• Winners announced on Slack, the HDSI Newsletter,

Date:	Session:	What you'll do:
Friday, Oct. 4th	In-Person Kick-off	• Team formation, if you're still looking • Register your team • Receive the prompts and select your challenge • Group work and meet with mentors in person
Saturday, Oct. 5th	Virtual or In-Person Work	• Teams can continue work in-person (Harvard locations) or virtually • Virtual meetings with mentors
		and more!

Kickoff: The competition officially begins on Friday after all team members have completed their in-person registration and after the live welcome session has ended. The Competition Prompts will be released during the live session.

Prompt Selection: Teams will have up to one hour to select and submit their chosen prompt on a Google Sheet. Please note that after the allotted hour, no changes are allowed, the Google Sheet will be frozen. The selected prompt will be the focus of your team's work for the remainder of the competition.

Accessing Mentors

We've assembled an amazing list of experts to help guide the development of your solutions. We've listed them below for you, but note that not all mentors will be available at all times.

A.J. Kumar (In-Person)
VP of Sustainability Sciences, IndigoAg

Dan Sumner (Virtual)
Frank H. Buck, Jr. Distinguished Professor
Agricultural and Resource Economics, UC Davis

Evan Marshall (In-Person; Virtual)
GIS Coordinator, Massachusetts Department of Agricultural Resources

Elaine Swanson (In-Person; Virtual)
Technical Lead, HDSI Agri Datathon
Harvard SEAS

Hanbin Lee (Virtual)
Postdoctoral Fellow, Agricultural and Resource

Michelle Audirac (In-Person; Virtual)
Data Scientist, National Studies on Air Pollution and Health, Harvard

Mauricio Tec (In-Person; Virtual)
Postdoctoral Fellow, National Studies on Air Pollution and Health, Harvard

Oladimeji Mudele (In-Person)
Postdoctoral Fellow, Harvard T.H. Chan School of Public Health

Shane Bussmann (Virtual)
Senior Data Scientist, CIBO Technologies

Siqin (Sisi) Wang (Virtual)
Associate Professor, Spatial Sciences Institute, USC;
Visiting Scholar, Harvard Center for Geographical

Economics, UC Davis

Jeffrey Hunt (Virtual)
Mathematical Statistician, USDA

Johannes Knittel (In-Person)
Wojcicki Troper HDSI Postdoctoral Fellow, Visual Computing Group, Harvard

Karen Olsen (In-Person; Virtual)
Computer Vision Scientist, CIBO Technologies

Margaret Kosmala (In-Person; Virtual)
Principal Data Scientist, CIBO Technologies

Analysis

Virginia Harris (In-Person; Virtual)
Statistician, USDA National Agricultural Statistics Service

Wolfram Schlenker (In-Person)
Ray A. Goldberg Professor of the Global Food System, Harvard Kennedy School of Government

In-Person Mentors (Friday, 10/4)
12:00pm-5:00pm

Mentors will float through the workspace on Friday afternoon. Flag someone down if you have a question!

Virtual Mentors (Saturday-Sunday, 10/5-6)
9:00am-5:00pm, both days

Send mentors messages in Slack #mentor-meetings channel. Our moderator **@Elaine Swanson** will help direct your question to a mentor.

Too complicated to discuss by chat? Our virtual mentors will also have access to Zoom rooms - just ask **@Elaine Swanson** for a link in Slack while you and your mentor are both online!

DATA DICTIONARY

Participants in the HDSI Agri Datathon will have access to a rich array of agricultural data, drawing from comprehensive resources provided by the USDA's National Agricultural Statistics Service (NASS). This data will be made available on the first day of the competition and will be stored in Google Cloud Storage Buckets (Google Bucket).

This *could* include data that spans various dimensions of U.S. agriculture, from crop production, sales, or yields, to soil moisture and vegetation indices as derived from several years of the Census of Agriculture (from 1997 to 2022, every 5 years) and spatial data repositories from NASA satellites. The geographic areas that the data will highlight *could* be county, state, or country wide.

Participants might engage with high-resolution geospatial data such as the CropScape which offers a detailed yearly overview of cropland patterns across the continental United States and/or the CROP-CASMA application, which provides Earth observation-based data on soil moisture and vegetation conditions, essential for assessing crop health and predicting yields. Or perhaps VegScape, a platform that delivers daily, weekly, and biweekly vegetation condition indices. Beyond geospatial data, the competition could involve statistical information from the Quick Stats database, which offers official estimates on virtually every aspect of U.S. agricultural production, collected through numerous surveys and the Census of Agriculture every five years.

The NASS Developer APIs provide developers and data users with direct access to a wide range of agricultural data, including geospatial maps, crop conditions, soil moisture analytics, and statistical information from the USDA's extensive databases. Teams can practice with these APIs before the competition to gain experience in retrieving and analyzing real-world agricultural data, allowing them to better understand the datasets they'll be working with and refine their data-driven approaches. **However, all data will be packaged and staged for you.**

External Data: You should be able to solve challenges with the data provided. You are welcome to use other public sources of data, however, no extra judging consideration will be given for the use of external data. Also, be aware that mentors may not be able to advise on the use of that data. If you choose to use external data, you must cite your sources appropriately.

Data Types:

Tabular Data: Tabular data is organized in rows and columns, like a simple dataframe. Each row represents a record, while each column represents a specific attribute related to that record. For example, in the context of agricultural data, tabular data might include rows of data representing different counties, with columns capturing variables such as crop yields, acreage, or economic indicators. This type of data is often used for statistical analysis, modeling, and generating reports.

Example column labels and entries:

STATE_FIPS_CODE	COUNTY_CODE	Census Year	…	AG_LAND_CROPLAND_HARVESTED_ACRES	VEGETABLE_TOTALS_IN_THE_OPEN_PROCESSING_ACRES_HARVESTED
09	011	2017		27857	36

FIPS codes are numbers which uniquely identify geographic areas.
The number of digits in FIPS codes vary depending on the level of geography *(state, county)*. **State-level FIPS** codes have two digits, **county-level FIPS** codes have five digits of which the first two are the FIPS code of the state to which the county belongs. So we simplify the county FIPS code to three digits as long as we include the state code next to it.
Note: Several numbers in the interval **01-56** for state-level FIPS do not have states attributed with them (examples: **'03'** and **'07'**). Similar missingness in county-level FIPS.

The county level data contains what are called *"disclosure values"*. This is data that on a county level, may be specific enough to identify an individual's private information. Thus, some information on a county level is not disclosed. On the state level reporting however, this information is much less likely to identify the individual's information. Due to this, row entries will look like this in your dataframe:

STATE_FIPS_CODE	COUNTY_CODE	YEAR
1	133	1997
1	133	2002
1	133	2007
1	133	2012
1	133	2017
1	133	2022
1	999	1997
1	999	2002
1	999	2007
1	999	2012
1	999	2017
1	999	2022

Where the number **1** in the first column represents the state ***01*** – *ALABAMA* (with the leading zero dropped) for the last Alabama FIPS county code of ***01133*** – *Winston County* (with the first two digits dropped) in the second column. Then you also see the code **"999"** in the second column. **"999"** does not correspond to a county.
Rather, it is the sum of all the county information (including the disclosure values) per census year (third column). If you were to take all of the county information for a census year and sum it together, it would not necessarily add up to the number displayed in these **"999"** state census year rows. This is due to the "disclosed values" privacy effort that I described above. Be mindful of this in your analysis. There is a similar explanation for when you reach the end of the csv and you see that the state code is not a FIPS code, rather it is the national total.

STATE_FIPS_CODE	COUNTY_CODE	YEAR
99	999	2002
99	999	2007
99	999	2012
99	999	2017
99	999	2022

GeoTIFF Data: GeoTIFF is a format for storing geospatial data in raster graphics, typically used for satellite imagery, aerial photography, and other types of spatial data. GeoTIFF files contain not just the image data but also metadata that describes the spatial information, such as coordinates, projection, and resolution. In the context of the datathon, GeoTIFF files might be used to provide participants with high-resolution images of cropland, vegetation indices, or soil moisture conditions. These images can be analyzed to extract spatial patterns, assess crop conditions, and integrate with other data types for comprehensive analysis. **Please see the Google Cloud instructions `'HDSI-AGRI-TEST-PROJECT' -> 'hdsi-agri-test-bucket' -> `prompt_1`` folder for examples of `.tif files.`**

Special Characters

All values in the tabular data are numeric and textual signifiers that are different than an NA. The textual signifier (D) is the most common and arguably important. (D) refers to the "disclosure value" that is described above. There is also (Z). The textual signifier (Z) represents very small percentage values that are less than half the unit, so if a percentage is expressed to one decimal place, it'd be values less than 0.1, etc.

(D) placeholder shows up over 34,000 times in one dataset (please take note of this)

(Z) placeholder shows up over 36,000 times in one dataset (please take note of this)

Required: Adjusting Farm Income Data from Nominal to Real Dollars

Adjusting farm economic data from nominal to real dollars is crucial to account for inflation and allow for meaningful comparisons of income over time. This process ensures that changes in purchasing power and price levels are factored in, giving a more accurate representation of farm profitability and economic trends.**Teams working with economic data will be required to adjust for inflation.**

What is Nominal vs. Real Dollars?

- **Nominal dollars** are the actual dollar amounts at the time you earned or spent them, without adjusting for inflation.
- **Real dollars** are adjusted for inflation, allowing you to compare the value of money from different years. This helps to understand how much farm income is worth today compared to the past.

What is the GDP Price Index?

- The gross domestic product (**GDP) Price Index** is a tool that shows how prices of goods and services produced in the U.S. change over time. It reflects price changes for what consumers, businesses, and the government buy (but not imports).
- We use this index to adjust farm income over time so that we can compare it fairly, without the influence of inflation.

What Does "Chained" Mean?

- The GDP Price Index is a "chained" index, which means it accounts for how people adjust their spending habits when prices change. It also gets updated as new methods or data become available, ensuring the numbers stay accurate.

How to Use the GDP Price Index:

- The index is set to **100** in the base year, which is **2017** for this data.
- If you want to compare a dollar value from 1997 to 2017, you can adjust it using the index. For example, $200 in 1997 can be converted to 2017 dollars by multiplying it by **100** and dividing it by the 1997 index value of **69.337**. This equals $288 in 2017 dollars.
- If you want to adjust dollars to **2022**, divide the index for each year by the 2022 index value (**118**) and multiply by **100**. For example, the 1997 index value in 2022 dollars would be **58.8**. So, $200 in 1997 becomes $340 in 2022 dollars.

Data Source:

- The data used comes from the **U.S. Bureau of Economic Analysis (BEA)** and includes their GDP Price Index and Personal Consumption Expenditures data.

Find the `gdp.csv` file HERE

Back to Top

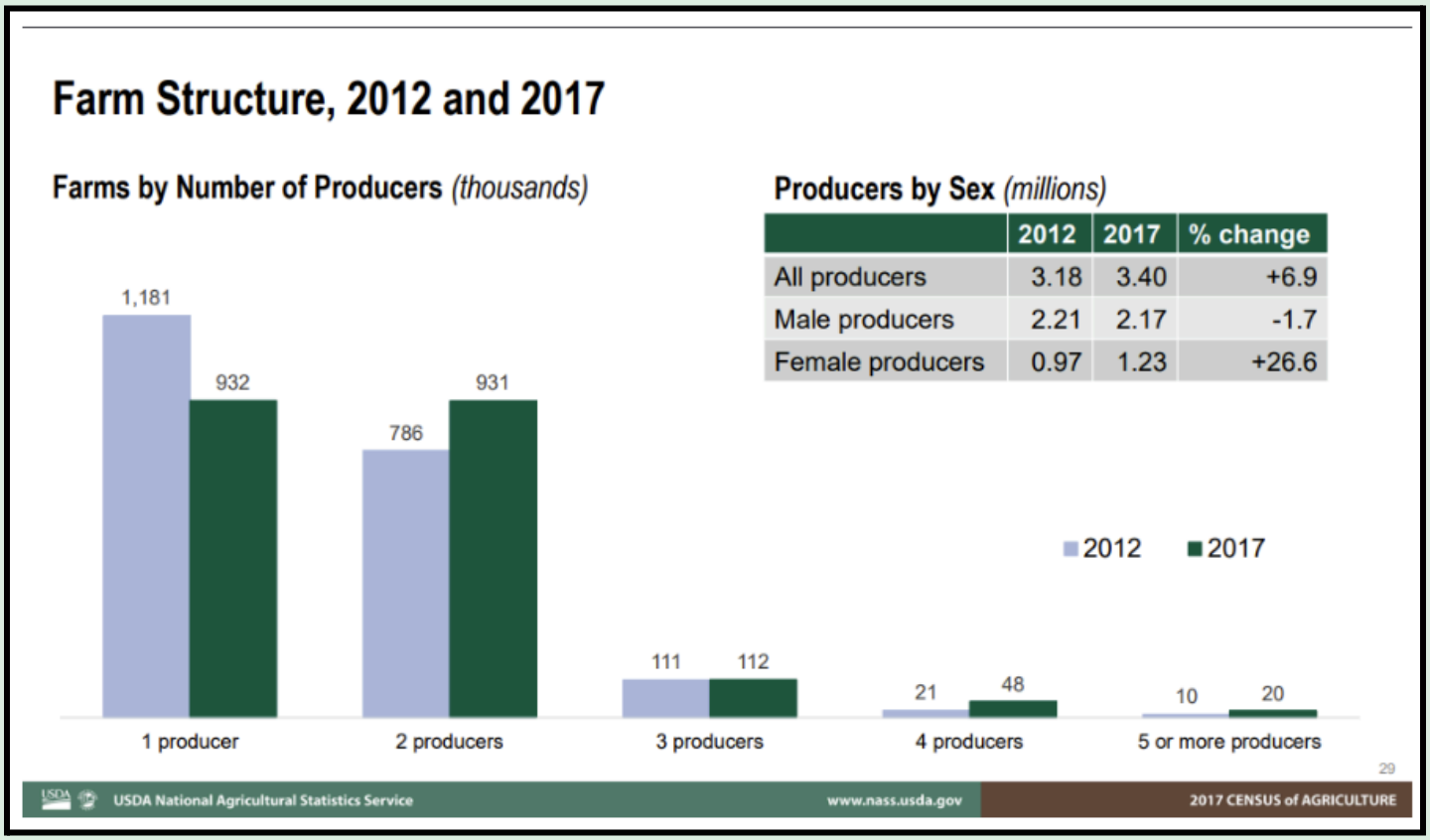
Important Note: Producer and Operator Data from the Census of Agriculture

In 1997 and previous censuses, demographic data were collected for one person per farm. From 2002 to 2012, the Census of Agriculture collected demographic data on up to three persons per farm: the principal operator and two other persons, referred to as operators in the published data. The data series was updated because it was recognized that many farming operations involved multiple people, for example, a husband and wife, two sisters, or family members from different generations, such as a father and a son. Respondents were asked to report details on up to three persons per farm who were involved in making day-to-day decisions for the farming operation. On the report form, the person listed first was designated as the principal operator.

Prior to the 2017 Census of Agriculture, data users reached out to NASS to request an update in how demographic data were collected, to better measure the participation of all persons involved in decision-making for farming operations. In particular, data users expressed concern that the role of women in farming was being undercounted in the published statistics. NASS convened an expert panel to review the method of demographic data collection and recommended updates to several items. The primary change was to collect demographic data for persons involved in making decisions for the farm (removing the "day-to-day" language from the questionnaire) and to collect detailed data for up to four persons per farm, one more than in the 2002, 2007, and 2012 Census of Agriculture.

This change in the way data were collected and published impacts data analysis. For 2002-2012 data, the metadata uses the word "operator" in the demographic data items. For 2017 and 2022, the word "producer" is used instead. Additionally, for 2002-2012, publicly available data at the county level are generally available only for the principal operator. For 2017 and 2022, data are available for all producers.

The change did not impact all demographic data in the same way. Certain producer attributes were much likelier than others to experience significant changes. As expected, the new data collection methodology increased the number of women counted, both as a share of all farm producers and in absolute numbers. While the count of all producers increased by almost 7%, the count of female producers rose by 27%, while the number of male producers decreased by 2%, similar to the change in the total number of farms, which was down 3%. The increase in the number of female producers was primarily caused by the large rise in farms reporting more than one person involved in making decisions for the farming operation.



AGRICULTURAL DATA RESOURCES

The USDA National Agricultural Statistics Service (NASS)

The USDA National Agricultural Statistics Service (NASS) is an agency within the United States Department of Agriculture (USDA) responsible for collecting, analyzing, and disseminating data related to agriculture in the United States. This agency conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. Production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm finances, chemical use, and changes in the demographics of U.S. producers are only a few examples. This information supports informed decision-making by farmers, policymakers, researchers, and the public.

NASS is committed to providing timely, accurate, and useful statistics in service to U.S. agriculture. To uphold our continuing commitment, NASS will:

- Report the facts on American agriculture, facts needed by people working in and depending upon U.S. agriculture.
- Provide objective and unbiased statistics on a preannounced schedule that is fair and impartial to all market participants.
- Conduct the Census of Agriculture every five years, providing the only source of consistent, comparable, and detailed agricultural data for every county in America.

- Serve the needs of our data users and customers at a local level through our network of State field offices and our cooperative relationship with universities and State Departments of Agriculture.
- Safeguard the privacy of farmers, ranchers, and other data providers, with a guarantee that confidentiality and data security continue to be our top priorities.

Read more:
- Understanding Agricultural Statistics
- Talking About NASS

Agriculture Outlook Forum
The Agricultural Outlook Forum (AOF), is USDA's oldest and largest annual gathering.

USDA's Agricultural Outlook Forum began in 1923 to distribute and interpret national forecasts to farmers in the field. The goal was to provide the information developed through economic forecasting to farmers so they had the tools to read market signals and avoid producing beyond demand. Since then, the Forum has developed into a unique platform where key stakeholders from the agricultural sector in the United States and around the world come together every year to discuss current and emerging topics and trends in the sector.

Read more:
- 2024 AOF Program and Slides
- A Snapshot of U.S. Agriculture: Highlights from the 2022 Census of Agriculture; Bryan Combs, Chief, Environmental, Economics, & Demographic Branch, NASS, Washington, DC
- Know Your Farmers: Why the Ag Census Matters; Roger Cryan, Chief Economist, American Farm Bureau Federation, Washington, DC

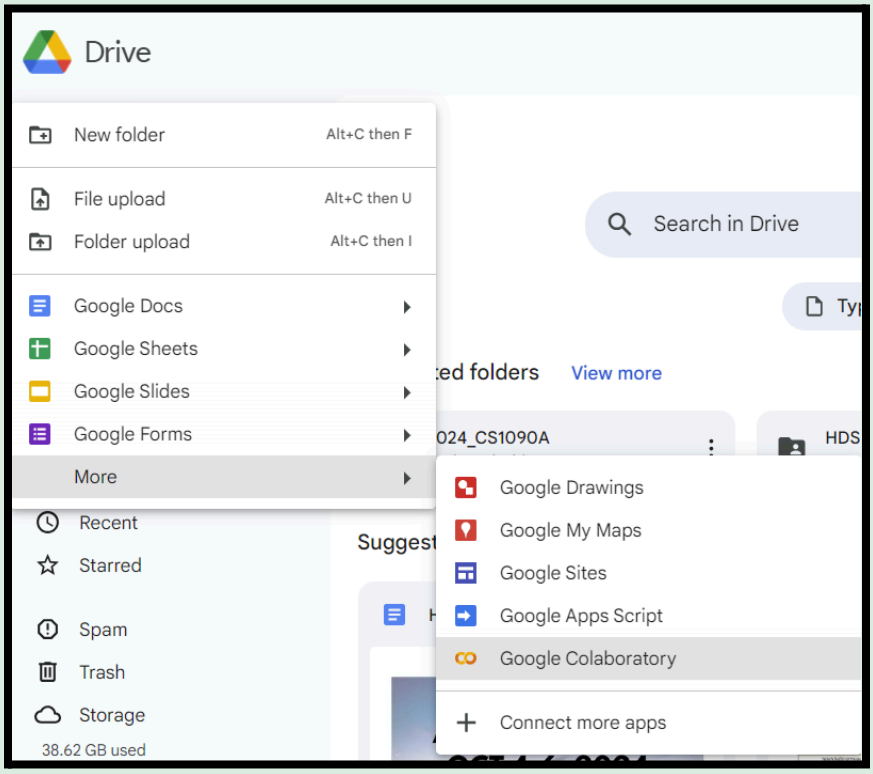
Agriculture - Harvard Libraries Research Guide
A guide to agricultural information, with emphasis on the United States.

TECHNOLOGY AND LOGISTICS

Google Drive
We recommend you create a shared Google Drive or a shared Google Folder with your team to store any important data, results, or visualizations along the way. Another resource here. You won't be able to push data into the Google Bucket, only read + download from it.

Google Colab
We encourage you to prepare with Google Colab in advance to ensure a smooth and efficient workflow during the competition. We have created a **test project and test bucket** with artificial data. Your team can practice pulling this data in a shared Google Colab Notebook. Please see the photo below on how to open Google Colab in your Google Drive. I recommend doing this in a Shared Google Folder with you and your team. If you

don't see the option for **Google Colaboratory** you will have to download it. Click through the "+ Connect more apps" and search for **Colaboratory**. This is free and it only takes a few seconds.



Google Colab primarily supports Python, but it also allows you to run C++ code within the notebook. Your team has flexibility in your choice of which programming language to use.

Using Python and C++ in Google Colab

- Python Support: Google Colab is designed to run Python code seamlessly. You can use it to execute Python scripts, import libraries, and perform data analysis, machine learning, etc.
- C++ Support: Colab also supports running C++ code directly within the notebook. To execute C++ code, you can use the following approach:

1. Creating and Running C++ Code:
   You can create a code cell in Colab and use the `%%writefile` magic command to write C++ code to a file. Then, compile and run the code using the terminal commands provided by Colab.

Example:

```
%%writefile quick_example.cpp
#include <iostream>
using namespace std;

int main() {
    cout << "I'm using C++ in Google Colab!" << endl;
    return 0;
}
```

After writing the C++ code to a file, you can compile it using the `!g++` command and run the compiled program:

```
!g++ quick_example.cpp -o quick_example
!./quick_example
```

2. Combining Python and C++:
   You can integrate Python and C++ code within the same Colab notebook, using Python for tasks like data processing and visualization, and C++ for performance-critical sections of your project. You are allowed to use this hybrid approach for the competition to leverage the strengths of both languages.

Learn more: Google Colab Tutorial

Instructions for Importing Data from Google Cloud Storage into Google Colab:
As a participant in the competition, you'll need to access blobs stored in Google Cloud Storage. The term "blob" is short for *Binary Large Object*, which is a general term used in computing to describe a large piece of data, such as an image, video, document, or any other binary data. Your team will follow these steps to import the data into your Google Colab environment.

1. Install Required Libraries

```
## Start by installing the necessary Google Cloud libraries in your Google Colab env
!pip install --upgrade --quiet gcsfs google-cloud-storage
```

2. Authenticate Your Google Account

```
## Authenticate your Google account to access Google Cloud services. This will allow you to interact with
Google Cloud Storage using the credentials associated with your account.
## Also import your Google Drive to access your team's Shared Google Drive as well as Operating System
from google.colab import auth
from google.colab import drive
import os
auth.authenticate_user()
drive.mount('/content/drive', force_remount=True)

## If your team is using a Shared Drive in a Google Workspace (usually through a University Gmail account)
directory_path = '/content/drive/Shared drives/YOUR_SHARED_DRIVE_NAME/'

## if your team is using a Shared Folder in a non-Google Workspace (for a free, individual account)
directory_path = '/content/drive/MyDrive/YOUR_SHARED_FOLDER_NAME/'
```

3. Set Up the Google Cloud Storage Client with the "HDSI-AGRI-TEST-PROJECT" ID

```
## Initialize the Google Cloud Storage client.
from google.cloud import storage
client = storage.Client(project='HDSI-AGRI-TEST-PROJECT')
```

4. Access the Specified Bucket and See What is Inside

```
## All the data you need will be located in a specific prompt folder located in a Google bucket.
## I have included some code for you all to start exploring file types

bucket_name = 'hdsi-agri-test-bucket'

## Access the specified bucket
bucket = client.bucket(bucket_name)

## List all blobs (files and folders) in the bucket
blobs = list(bucket.list_blobs())

## Initialize variables to count and store folder names and file types
folders = set()
file_types = set()
blob_count = 0

## Loop through all blobs to gather folder names and file types
for blob in blobs:
    blob_name = blob.name

    ## Check if it's a folder (by convention, ends with '/')
    if blob_name.endswith('/'):
        folders.add(blob_name)
    else:
        ## Capture file type
        file_extension = blob_name.split('.')[-1] if '.' in blob_name else 'Unknown'
        file_types.add(file_extension)

    blob_count += 1

## list the folders
print("Folders in the bucket:")
for folder in folders:
    print(f" - {folder}")

## Show the file types of the first 5 files (skip folders)
print("\nFirst 5 file types:")
counter = 0
for blob in blobs:
    if not blob.name.endswith('/') and counter < 5:
        file_name = blob.name
        file_extension = file_name.split('.')[-1] if '.' in file_name else 'Unknown'
        print(f"File name: {file_name}, File type: {file_extension}")
        counter += 1
```

```
## Total count of blobs
print(f"\nTotal number of blobs in the bucket: {blob_count}")

## Print the unique file types found
print(f"\nFile types in the bucket: {', '.join(file_types)}")
```

LaTex/Overleaf

To make the most of your competition time, we strongly encourage participants to familiarize themselves with LaTeX formatting before the competition begins. Understanding the basics of LaTeX will help you produce a polished final document and avoid unnecessary technical difficulties during the competition.

We recommend using Overleaf.com for preparing your submission. Overleaf is an intuitive platform that simplifies the process of creating LaTeX documents, allowing you to focus on content of your write-up rather than formatting challenges. Once your document is complete, it can be easily downloaded as a PDF for submission with the links in the appropriate place.

Learn more: Intro to LaTex/Overleaf

Recording Your Video Presentation

Participants may use any option they are familiar with to record their video presentation. However, here are two recommended options:

Zoom Recording: Start a Zoom meeting with all team members' cameras on. With one member sharing the slides that contain a description of your work, you can walk through the slides together. This session can be recorded directly using Zoom's recording feature.

Open Broadcaster Software: Another recommendation is using OBS (Open Broadcaster Software), a free and open-source screen recording tool, as an option for recording your video presentation. OBS allows for high-quality recording of your screen, enabling you to capture your slides and team member's videos, seamlessly.

PREPARING YOUR SUBMISSION

Deadline: Final submissions are due by Sunday, October 6th, at 11:59 pm. Submissions must be in the form of a PDF document, which will be **submitted to the designated Google Form** at the top of this page. It will be available until a few moments after the deadline. Your team will not be able to view submissions of other teams.

Final Deliverable Format:

The submission must be a 3-page document, **including figures, tables, and text (excluding references),** formatted in LaTeX. Submissions must follow the template posted below. There are very detailed notes in the *README.md* in the GitHub Repository below.

Access the LaTex submission template on a GitHub Repo here.

The final PDF must include:

1. Your team's text write-up of your work with all required sections completed. (3 pages)
2. A link to a YouTube video summarizing your team's work. (6 minutes)
3. A link to your team's Google Colab Notebook.

Link Accessibility: It is very important that both links are set to be accessible to anyone. Failure to do so will result in your explanation and code being inaccessible to judges, which will impact the evaluation of your submission. You are allowed to have someone outside of your team verify that the links are sharable and your document is complete before submission.

Code Execution, Validation, and Organization

- Include TOC (table of contents)
- Ensure that your entire code, whether Python, C++, or both, runs smoothly from start to finish without errors or warnings.
- Test your notebook thoroughly to ensure that all dependencies are met and that the code can be executed in the Colab environment without requiring additional configuration from judges.
- Please make sure that your Google Colab Notebook has clear descriptions and comments throughout. You will be doing your official write-up with text in LaTeX, but judges should be able to follow and understand what your team did and why.
- Include an illustration/flow chart of your model. This can also be used for your presentation.

Video Presentation Tips (Dr. Pavlos Protopapas' 10 commandments)

1. Instead of reading from a script, practice your presentation at least three times to develop a natural flow and pay attention to your intonation.
2. Avoid creating slides that do not contribute to the overall message of your presentation. Make every slide count, and be sure to talk about each slide.
3. Check the audio quality and make sure there is no background noise.
4. Be sure to include slide numbers in your presentation.
5. Keep the font sizes consistent throughout your presentation, and use no more than three different fonts. Highlight only essential keywords and avoid unnecessary capitalization.
6. Edit and smooth the transitions between slides, and consider the aesthetics, such as using colors.
7. Avoid using slides that are too text-heavy. Split the information onto multiple slides or ensure a balance between text and white space.
8. Introduce your teammates and use "we" instead of "I" to show that the presentation is a team effort.
9. Try to stay within the given time constraints for the presentation.
10. Be excited about your project! Engage with the audience and convey the highlights of your efforts.

JUDGING

Judges are subject matter experts across a wide range of backgrounds in data science, remote sensing, agricultural economics, or similar disciplines. The judging process will be conducted **virtually between October 7-21, 2024**.

Teams will be evaluated according to the following criteria:

1. **Clarity and Organization** - Is the presentation/report well-structured and easy to follow?
2. **Relevance to Problem Statement** - How well does the submission address the specific problem statement or challenge?
3. **Methodology and Analysis** - Are the methods and analytical approaches sound, thorough, and well-explained?
4. **Innovation and Creativity** - How original is the approach or solution?
5. **Data Visualization** - Are data visualizations clear, effective, and well-integrated into the presentation/report?

Cash prizes

Teams will be awarded the following prizes:

- $900 for first place
- $600 for second place
- $300 for third place
- Honorable mentions for Best Presentation and Best Visualization

Equal shares of any cash prize will be paid to each participant on a winning team. Note that prize payments are treated as income – individual recipients of prize money will be responsible for the tax implications of their winnings.

Winners will be announced in the Slack channel, the HDSI Agri Datathon website, the HDSI newsletter, and more!

FREQUENTLY ASKED QUESTIONS

We will continue to post questions to the toolkit, however join our Slack to get up-to-the-minute answers to all your burning questions!

Is there a cost to attend?

No, the datathon is free for participating undergraduate and graduate students.

Are travel/meal/accommodation stipends available?
We currently do not have stipends available for participant expenses. Any changes will be announced via the HDSI newsletter.

Can I participate virtually?
We require all participants to attend our in-person kick-off, on campus at Harvard University. After kick-off, your team can decide to work virtually up through submission.

What if I don't have a team?
Register as an individual! There will be time for team formation during kick-off.

What if my team is incomplete?
You should still register as a team! Just note in the form that you're seeking a third team member. There will be time for team formation during kick-off.

What happens if I win a cash prize?
Equal shares of any cash prize will be paid to each participant on a winning team. Note that prize payments are treated as income – individual recipients of prize money will be responsible for the tax implications of their winnings.

I'm not ready to register. How can I receive updates about the datathon?
Sign up for the HDSI newsletter. We'll be sending regular reminders to our community there.

Is this limited to Harvard?
Nope! Any undergraduate or graduate student can participate, as long as you can attend the in-person kick off on Oct. 4th.