Module 2A

Find Existing Data

Learning Objectives

Discover available data sources
Balance priorities to choose the right data set
Find substitute or proxy data source

SAVI

Prep

30 MIN

Handouts
Print 4 copies of Handout 2A-1
Print 15 copies of course overview

Online Resource
SAVI.org Data Sources page

Data
None

Technology
A computer with a projector
One workstation for each group

Setup
Tables and chairs for three groups of three to four each. One workstation for each group.

How to read this guide

- Each section of the lesson plan has a card or two guiding you through what to do or say.
- The lesson plan has lecture sections and group work sections.
- Cards also show learning objectives and the time alotted for that card.

Gray boxes are tips for the teacher. They might give a useful example to share, or tell you how to illustrate a concept.

Blue boxes are points where the class contributes to the conversation. You might ask for examples of a concept or solutions to a problem.

Lecture Objective: Recap Module 1A: Goals, outcomes, outputs **3 MIN**

- In Module 1A we learned that goals are broad statements that answers the question, "What does your organization hope to accomplish?" Goals are general, conceptual, and abstract.

  Example: "We're increasing the academic success of our youth by making quality child care accessible, fostering early reading skills and supporting students through graduation."

- We learned that outcomes are are meaningful, measurable change related to the goal. They are what you have accomplished.

  Example: "Participating students scored 37% higher on reading comprehension tests than their peers

- We learned that outputs are what we plan to do to accomplish that outcome. They are specific, measurable, concrete activities. They should specify what you plan to do, and they should allow you go back and see if you did it or not.

  Example: "650 students received tutoring at 39 locations"

Lecture Objective: Recap Module 1A: Measurable questions **2 MIN**

- We learned to take broad questions and "operationalize" them, or make them more measurable.
- We do this by narrowing their scope in terms of time and geography, and normalizing them so you can compare across different times and geographies.

Example: "How many crimes have taken place in Indianapolis?"

This becomes: "How many violent crimes per 1,000 people took place in the Near Eastside in 2016?"

Group Work — Objective: Decide what goal, outcome, and questions they will working on in class. — 5 MIN

- "Now that we have arrived at measurable questions, we are going to work in groups to figure out what data sources could help us answer those questions."
- "We will learn what sources are available, how to choose between sources, and how to find substitute data."
- Ask participants to meet in groups of 3-4 and decide which group members' goals, outcomes, and questions they will be working on.
- Provide any participant that did not come to the class with any goals, outcomes, and questions in mind with one of the sets available in handout 2A-1.

If possible, try to steer a group to a difficult question that will be a good example for proxy data.

Lecture Objective: Learn about types of data, geographic scope, and geographic resolution **5 MIN**

- Often we can jump straight to the data point we think we need, but it's better to step back and understand our problem, understand available datasets, and make the best decision about which data to use.
- There are a few key dimensions to consider when looking at a dataset.
- Type of Data: Really consider what kind of information is in this dataset. There might be spatial information, chronological information, categorical information.

Ask the class to think of datasets that would have spatial information. It might be necessary to explain that even if the data is in a spreadsheet, it can still have spatial information, like addresses, zip code, or county name.

- Geography of data: If there is any spatial information, what kind of geography is it about? Geography has two dimensions: scope and resolution.

Have the class imagine they survey every person in their neighborhood. What would be the geographic scope of this data? (Neighborhood) What would be the geographic resolution? (Each house, or point data) Now what if we only surveyed enough people to estimate data for each block, not each house?

Lecture	Objective: Learn about data formats, frequency, and cost	5 MIN

- The format of the data determines what you can do with it. If you find a report or an online dashboard, it will probably be easier to understand the results quickly, but you may not be able to customize the "query."
- If the data can be downloaded as a spreadsheet or a text file (like CSV), it will take a little more work to summarize, but you might have more options for how to summarize and cross-tabulate the data.
- Frequency: Make sure you understand how often the data is published and what time range that publication covers. We will talk more about this soon when we learn more about the American Community Survey.
- Cost: This one is pretty clear. Some data is free, some data is not. Sometimes, paid subscription data can have valuable advantages, but there is so much public data out there, you are very likely to find something valuable in public data.

Ask the class to think of any other dimensions that would be important in assessing a dataset. Have we missed anything?

Lecture Objective: Learn about point, line, and polygon data. **5 MIN**

For each spatial data type, ask the class for examples first, and then advance the slide to show a couple examples.

- After going through each data type, look at a point data example:

Let's think back to that neighborhood survey we just talked about. If you surveyed the resident of every single house in the neighborhood, what kind of spatial data would you have?

- Now look at a polygon example:

What if you only surveyed a sample of the neighborhood, maybe a few people on each block?

- Now look at a line example:

What if instead of surveying residents of houses, your neighborhood survey was a team of volunteers collecting data about the condition of streets in the neighborhood. What kind of spatial data would that be?

Lecture Objective: Learn more about point and aggregate data **5 MIN**

Discuss the benefits and challenges of point data.

Ask participants if the can think about what could the benefits and challenges of using point data be. For example, in our neighborhood survey, what if we asked every household in the neighborhood about their income? What would be the benefits and challenges of information at this level of detail? (Answer: the benefits are the detailed spatial information you can analyze, the challenge is privacy concerns.)

Discuss the benefits and challenges of aggregate data.

Ask participants if the can think about what could the benefits and challenges of using aggregated data be. In our neighborhood survey, if data is aggregated to the block level, we could probably publish our results without ticking off our neighbors. But, we would not be able to do certain kinds of analysis, for example, do people of a certain income tend to live next door to each other. We don't have high enough geographic resolution.

Discuss the advantages of the aggregation and processing that SAVI does. Mention that SAVI stands for "Social Assets and Vulnerability Indicators" and explain that assets are point data while vulnerabilities are aggregated data.

Lecture Objective: Learn about census geographies and the ACS **5 MIN**

- **To describe census geographies, it's helpful to remind the class that the census goes door to door and surveys every household. So inititally, this is point data.**
- **You can draw an example on the white board of a few houses next to each other.**
- **They aggregate this data to blocks to avoid privacy issues. Draw the block around the group of houses.**
- **Now draw multiple blocks to show how they combine to form block groups.**
- **Do the same for census tracts, townships, and counties, drawing each consecutive boundary and showing how it contains the smaller boundary.**

Now that we have learned about the census and its geographies, let's discuss the pros and cons of the census. The good thing is it is very nearly complete. It attempts to survey every individual. (Discuss the fact that it can miss some residents.) The downside is such an effort is only possible every decade. To resolve this, the Census Bureau developed the American Community Survey. The pros are the data is available every year. The downside is the ACS is a survey, so the data is only an estimate.

Lecture Objective: Understand the tradeoffs in data attributes **5 MIN**

- Each of these attributes can relate to these others. If you are looking for spatial data, that might impact the format of the data. If you're looking for data that is more "real-time" rather than delayed by a year or two, that might cost money.

Ask the class what other tradeoffs they can think of. This is a good point to think about proxy data. While you might go into your data search with a particular data point in mind, you may discover that this data is behind a pay wall, it is something you would have to collect yourself, or it would take a lot of work to process. (Time is money.) Can you find another data point that meets all your goals better? The idea is to think about data goals more broadly. You don't just need an answer to a question, you need information that achieves your program goals, your financial needs, and your technical facility.

- SAVI is free, but it is reliable, easy to access, and has high data quality. We try to make the calculation easier for you by reducing some of the trade-offs. Tons of census and ACS data that would be time consuming to process for your particular geography is quick to find in SAVI. Many confidential administrative records (like births and deaths) are aggregated and published through SAVI.

Group Work Objective: Identify datasets that could be use to answer their questions. **30 MIN**

GROUPS WORK TOGETHER (15 MIN)

- The first 15 minutes will be for participants to meet on their groups and decide which datasets from SAVI.org/support-training/data-sources might be better suited to answer their measurable question.
- Reiterate how the information available on the data matrix columns might affect their decision.
- If a group has identified the possible datasets faster than expected (or if the other groups need a little more time), encourage that group to go online and navigate through the selected dataset.

CLASS OFFERS FEEDBACK (15 MIN)

- Use the last 15 minutes to hear which dataset each group has selected and the reasoning behind their selection. Invite the class to offer feedback and questions. Make sure feedback is led by the rest of the class.
- If a group had some trouble finding a proper dataset, this is an opportunity to discuss proxy data. If no one had trouble finding a good dataset, ask the class for an example of a time when they could not find the data they needed.
- In either case, help find a proxy data solution for that scenario. This should be class-led. Ask the class for ideas of what data might make a good substitute.

Lecture	Objective: Overview on descriptive statistics and metadata	15 MIN

- Provide some basic description about mean (average), median, and mode, and how they differ from one another. Use the example of median income from the example households in the slide deck.
    - Mean: the average value.
    - Median: the middle value.
    - Mode: the most frequent.
- Explain what metadata is. Why is it important? Where can it be found? How can it be saved/storaged? Why should be the metadata downloaded and saved at the same time as the data?

Lecture Objective: SAVI Tools walk through **15 MIN**

Using the descriptions on slides 42 and following, talk about the advantages and disadvantages of each SAVI tool.

The important concept here is that there are many SAVI tools, and they each differ in terms of geographic scope, geographic detail, available indicators, and customizability.

Class Work Objective: Identify which SAVI tools datasets could be use to answer their questions. **30 MIN**

- Together as a class, walk through each of the three groups' measurable questions in a SAVI tool.
- Spend about 10 minutes on each group.
- Begin by asking what measurable question they worked on in the last activity and what dataset they found that may help answer that question.
- Look at the Use SAVI page on SAVI.org to remind the class about the attributes of each tool. Discuss as a class which tool might be best.
- Try to answer the question using the selected tool.
- Through this exercise, the class should see that: 1) they may need to modify their question to find the right data or to find it in SAVI and 2) once they find information in SAVI this can lead to more and more exploratory questions.
- At the end of this exercise, remind the class that we are talking about quantitative data here, but qualitative data is also very important. To really understand a place, we probably all know that you need a lot more contextual information and qualitative information.

Any questions?