Data Collection & Analysis

Basic tools and techniques for collecting and analyzing different types of data

Cait Harrigan

Department of Computer Science
University of Toronto

February 17 2022

Where does data come from?

Two broad classes of dataset that you will encounter:

Where does data come from?

Two broad classes of dataset that you will encounter:

Data "in the wild"

- Messy
- Domain expertise needed
- May require special handling (ethics approval, privacy, IP)

Where does data come from?

Two broad classes of dataset that you will encounter:

Data "in the wild"

- Messy
- Domain expertise needed
- May require special handling (ethics approval, privacy, IP)

"Domesticated" data

- Clean!
- Well studied benchmarks
- May be simulated, or otherwise not representative of the real world

Think about your problem setting

Depending on your research area, standards for how to treat data will vary widely...

Some examples:

- A statistical theory researcher using linear regression as a toy model
- A HCI researcher using survey data and eye tracking to understand how users interact with their product
- A climate researcher using meteorological data from weather stations
- A computational social science researcher scraping twitter posts

Think about your problem setting

Much in the way a not-so-complicated research question gets increasingly complicated the more you look at it, datasets tend to have hidden complexities that only reveal themselves once you've spent some time with them.

Exploratory data analysis

Spend some time poking around to understand the type of data you have, and what kinds of approaches might be appropriate.

Exploratory data analysis

Spend some time poking around to understand the type of data you have, and what kinds of approaches might be appropriate.

Identify the type of data you are working with, and standards of handling you should be aware of...

- tabular data
- images
- text / natural language
- graphs or other representations

Exploratory data analysis

Think about the data collection process. If you have access to domain experts, ask them to describe how they think about the data, and what limitations exist in how it was collected.

Exploratory data analysis

Think about the data collection process. If you have access to domain experts, ask them to describe how they think about the data, and what limitations exist in how it was collected.

- How much control do the domain experts have over the data? Is it experimental, observational, or synthetic? Is it reproducible?
- Is the collection process stable over time?
- What kinds of features are available to you? What dependencies might exist among them? Are there redundancies?
- How is data missingness treated?
- How generalizable is this data? Is it representative of the "real world"?

Data collection

When it's time to create a research proposal, you need to demonstrate to your reader that your research is both necessary and feasible. You should be able to describe the data collection process, and details on data processing that show you **understand your data's limitations.**

Data collection

When it's time to create a research proposal, you need to demonstrate to your reader that your research is both necessary and feasible. You should be able to describe the data collection process, and details on data processing that show you **understand your data's limitations.**

Some places to look for data:

- Previously published research. Critically read their description of data collection!
- Government-curated open access data is often well-documented and longitudinal
- Curated benchmark datasets (ex. MNIST, ImageNet, Kaggle datasets)
- New data generated by you or a collaborator may require a description of the validation process

See the end slide for some links to datasets you might find useful 😄

How to approach a new dataset

Start small. Try to write down some rules about the data that you can observe from just looking at a few examples.

Make a data exploration plan. Write down some questions you have about the data, and how you might answer them.

Explore. Answer the questions you wrote down! Write down results as you go for you to reference later. (This can also be important for reproducibility)!

This is an iterative process!

By answering your initial questions, you will be able to come up with the next set. You don't need to come up with The Perfect^{TM} set of questions to get started.

Example

In my research, I need to find cells in an image. These images have been prepared for me in the lab by my collaborator.

Example

In my research, I need to find cells in an image. These images have been prepared for me in the lab by my collaborator.

1. Start small: I picked one image, and chose segmentation settings that look good.

Example

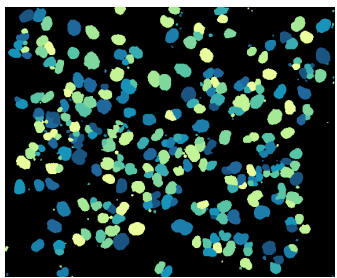
In my research, I need to find cells in an image. These images have been prepared for me in the lab by my collaborator.

1. Start small: I picked one image, and chose segmentation settings that look good.
2. Make a data exploration plan: Can I use the same segmentation settings for all the images? Are all the cells in all images approximately the same size? Do they all have similar brightness?

Example

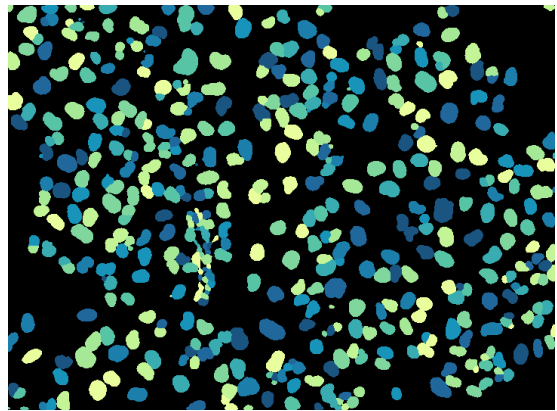
In my research, I need to find cells in an image. These images have been prepared for me in the lab by my collaborator.

1. Start small: I picked one image, and chose segmentation settings that look good.
2. Make a data exploration plan: Can I use the same segmentation settings for all the images? Are all the cells in all images approximately the same size? Do they all have similar brightness?
3. Explore:



Iterate with new questions

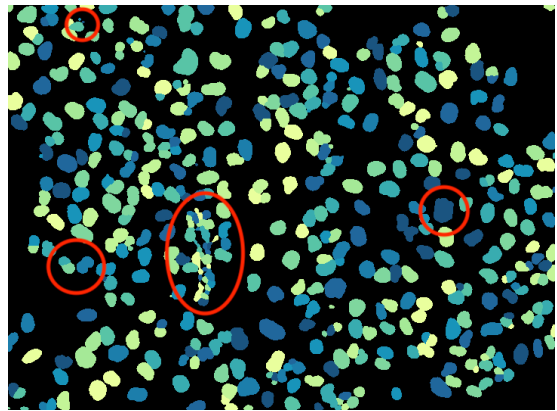
What would it look like if my segmentation failed?
What are the failure modes?



Iterate with new questions

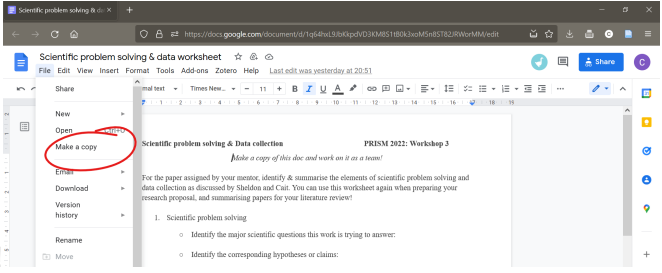
What would it look like if my segmentation failed?
What are the failure modes?

- cells that were not distinguished (doublets)
- cells that are a weird shape
- segmented objects that are not cells



Today's activity

Pick a paper. Identify summarise the elements of scientific problem solving and data collection as discussed by Sheldon and Cait. You can use this worksheet again when preparing your research proposal, and summarising papers for your literature review!



Some datasets for you 🚀

- Kaggle competition datasets
- Toronto open data
- British Columbia open data
- Ontario Data Catalogue
- Los Angeles city data
- Google Earth Engine
- Google Dataset Search
- FiveThirtyEight open data
- World Bank open data
- US Open Data Initiative DATA.GOV
- US National Historical Geographic Information System (NHGIS)
- Canada Census Data

Some more datasets for you 

Many of these websites have API to download the data. It's recommended you use these when available!

Health and Biological data

- NIH Cancer Surveillance
- NCBI Gene Expression Omnibus
- World Health Organization data
- UniProt data
- The Gene Ontology Project
- US Center for Disease Control and Prevention Data
- California Health and Human Services Open Data Portal
- Covid Data CovidTracker

Academic Publications and related

- Figshare data repository
- Zenodo data repository
- Harvard Dataverse
- Elsevier Developers API

Social Networks

- Twitter Developers API
- GitHub Developers API
- Instagram Developers API
- LinkedIn Developers API
- Zillow Developers API