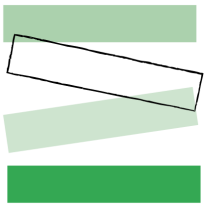
Data Collection + Evaluation

Chapter worksheet



Instructions

Use the exercises below as needed throughout your product's development.

Exercises

1. Get to know your data [~1 hour]

Decide what data you need, whether or not it already exists, and understand the sources.

2. Speak with a domain expert [~1 hour]

Use these questions as a starting point to speak with an expert in the domain.

3. Data collection considerations matrix [~1 hour]

Examine the goals of your collection effort and how you will know when you have the right data.

4. Data Labelers + Task Design [~3 hours]

Work with your labelers to ensure they have the right tools for this critical work.

5. Write data disaster/diligence headlines [~1 hour]

Avoid data disasters before they happen with this brainstorming activity.

1. Map user needs to data requirements

The first task your team has to complete is to identify the type and scope of data needed to train an ML model that can meet your users' needs.

Use the template below for each unique user need your ML model will impact.

Example: building a recipe recommendation service that suggests new dishes to cook.

User needs & data needs	
Users	*Home chefs*
User action (core value prop)	*Cook a new dish using the recipe based on recommendation*
ML system output	*Recommendations for new recipes*
ML system learning	*Patterns of behavior around choosing recipe recommendations*
Training dataset needed	*Set of recipes user has previously found, used, and liked*
Key **features** needed in dataset	*Ingredient cost* *Cuisine type* *Allergens* *Dietary restrictions*
Key **labels** needed in dataset	*Home cook's accept / reject of recommended recipe* *Home cook's feedback as to why suggestion rejected (user-generated label)* *Recipe ratings from other users*

Data **formatting**	*dish_name (all lower case)*
Real world data considerations	*Does your recipe dataset…* *...account for speciality holiday dishes?* *...reflect dietary and allergen concerns?* *...account for different cooking equipment?*
Data source key user questions	*“How does the app know what I like?”* *"Where do these recipes come from?"*

Synthesize your core data needs with the template below.

Our product/service uses:

- _____**{ data source }**_____
- _____**{ data source }**_____
- _____**{ data source }**_____

to provide _____**{ user type }**_____ with____**{ core value prop }**_____.

Critical labels for our data include:

- _____**{ data label }**_____
- _____**{ data label }**_____
- _____**{ data label }**_____

We’re aware of how the real world (e.g. time of year, changing trends) can impact the data used in our model.

To reflect the dynamism of the real world we made sure our data includes:

- **_____{ real world data consideration }_____**
- **_____{ real world data consideration }_____**
- **_____{ real world data consideration }_____**

2. Speak with a domain expert.

Once your team has the user-data needs template complete, identify **domain experts** who can give you feedback on your initial data hypotheses.

A domain expert is someone with a specialization in your ML model's subject area (not necessarily a ML expert) and can give you insights into the real-world implications of your data.

Questions for domain experts

- What data are important in your domain for <target use case>?
    - What makes data usable vs. unusable in your domain?
- How are data collected in your domain <target use case>? (e.g. in person, on paper, over the phone, online, a mix?)
    - Do you have recommendations for data collection and/or labeling organizations?
- What problems occur with the data (e.g. reporting, representation, capturing, updating)?
- Are there any environmental and/or temporal circumstances that impact data collection (e.g. type of sensor used, time of day/year)?

- How easy or difficult is it to reuse data in your domain?
- What are the top 3-5 things people should be aware of when it comes to working with data in this domain?

3. Data Collection Weighted Matrix[1]

Once your team knows what data will be required to train your model based on your answers to in the user + data needs template from exercise 1 and you've consulted with domain experts, you'll need to determine if you can get those data from:

- An existing dataset
- A new dataset

Use the weighted matrix below with your team to gain consensus on your data collection plan *(example matrix filled in below for a team with 6 people voting)*:

1. Have each team member vote for which dataset type is the best option for each row
    - The dataset criteria are suggested, you can change the criteria based on your team needs, but we strongly recommend always including 'fit for use case' and 'maintainability'
2. Multiply the number of votes for each option by the associated weight
3. Total the weighted number of votes per dataset option to give direction to your data collection plan.

[1] This exercise is adopted from the weighted matrix exercise featured in Martin, Bella, and Bruce M. Hanington. **Universal Methods of Design**: 100 **Ways** to Research Complex Problems, Develop Innovative Ideas, and **Design** Effective Solutions. Beverly, MA: Rockport Publishers, 2012.

Dataset options → Data criteria ↓	Weight	**Existing dataset (no transformations)**	**Existing dataset (with transformations)**	**New dataset + Existing dataset**	**New dataset**
Fit for use case *Is this data appropriate for your users and use case? Consider PII and Protected Characteristics: in some regions it's illegal to use them to make certain predictions.* *Are there any risks of the dataset excluding certain user groups?* *Have you used the Facets tool or some other tool/technique to evaluate the dataset for bias?*	3	(1x3)	(1x3)	(1x3)	(3x3)
Legality / Compliance *What data standards are in place for compliance, licensing, documentation?* *See if you the dataset has a Data Card (or whether your team would need to create one)*	3	(2x3)	(1x3)	(1x3)	(2x3)
Maintainability *Does your team have a plan for maintaining the data post launch?* *How will data stay up to date over time?*	2	(1x2)	(1x2)	(1x2)	(2x2)
Data collection effort *How will the data be collected?* *How will your team ensure ethical data collection practices?*	2	(1x2)	(3x2)	(1x2)	(1x2)
Cost *What are the costs of choosing the most expedient data vs. the best data?*	1	(1x1)	(3x1)	(1x1)	(1x1)
Total		14	17	11	22

4. Data Labelers + Task Design

If your feature uses supervised learning and you are using a new dataset, you need to understand the people who will be teaching or evaluating your model, also known as "raters", (or “oracles”, “labelers”, or “analysts”).

Labelers can be:

- Employees at a labeling company
- Volunteers
- Your own team members
- Or a combination of all of the above!

Use the questions below to get to understand potential mental model mismatches between your labelers vs. your users.

4.1 Who are your labelers?

- What are the particular perspectives or biases that labelers may be bringing to this task that could impact the quality of the labels?

Consider what contextual knowledge would be important for a person labeling data for:

- An AI music recommendation system
- An AI predicting likelihood of depression
- An AI for recommending job candidates

- How will you compensate labelers fairly for their work?

Consult with your domain expert for advice.

4.2 Task Instructions checklist

Help your labelers master a task by creating easy to use instructions.

DRAFT AND PILOT

- ☐ Draft instructions and budget time to get feedback from labelers on any aspects of the instructions that are unclear. *If you have already made instructions, don't worry! You can ask for feedback at any point.*

BITE-SIZE

- ☐ Break down instructions into manageable chunks by using bullets for steps, data items, or rules.
  - ☐ In house labeling teams and 3rd party companies may have the benefit of doing in person/remote trainings, but that doesn't mean instructions shouldn't be broken down into easily referenceable chunks

EXAMPLES/IMAGES

- ☐ Add at least 3 positive, negative, and ambiguous examples to illustrate expectations.
- ☐ If you are advertising a task on an open crowd platform, use images to capture worker interest in your task.

EXPLANATIONS

- ☐ Explain the overall goal of the effort to provide context and get labeler investment.
- ☐ Explain criteria for acceptance, and clearly state what errors would trigger a rejection of the task. Allow for a feedback mechanism for labelers to flag ambiguous cases.

ACCESSIBILITY

- ☐ Highlight if the task is fully accessible or requires specific abilities to complete.

4.3 Task design and usability

In case you missed it - read the article First: Raters to understand how different types of labeling impact the design of labeling tools.

- ☐ **Do the task yourself!**
  - ☐ Catch and correct any usability issues prior to testing with labelers.
- ☐ **Observe people completing your task**
  - ☐ Can labelers complete key tasks quickly and without errors?
  - ☐ Note: make it clear you are evaluating the task and not the individual's performance.
- ☐ **Plan for unsures**
  - ☐ Is your labeling UI forcing labelers to label prematurely or in error?
  - ☐ How are you thinking about inter-rater reliability?
    - ☐ Will labelers be able to periodically indicate their level of confidence for a given task submission? (This technique can help reduce the need for multiple ratings)
    - ☐ Can the data be labeled in more than one way?
- ☐ **Welcome feedback on your task/tool**
  - ☐ What incentives are there for labelers who speak up about discrepancies or interesting insights beyond the scope of the task?
- ☐ **Provide feedback to labelers in a timely manner**
  - ☐ How will labelers know they are doing a good job and that their feedback is valued?

Additionally, you can use / modify the following questionnaire to evaluate the usability of your task:

Please evaluate the usability of the task you are working on.

	Agree	Disagree	Not applicable	Comments
1. The goal of the task is clear				
2. The task instructions are comprehensive				
3. The task instructions are easy to reference				
4. The task was easy to learn				
5. The steps to complete the task are in a logical sequence				
6. The task shortcuts are useful				
7. The task shortcuts are logical				
8. It is easy to ask questions and get answers about the task				
9. The time to complete the task is appropriate				

5. Data disaster/diligence headlines

Write data disaster (and diligence) headlines to spot problematic data issues before they happen. Use these headlines to identify any data concerns to follow up with your engineering partners.

Data privacy	Customers of **{product}** upset to learn it uses **{sensitive data}**.
	{Product} champions essential data by limiting use of **{sensitive data}**.
Guiding questions ● How do you get access to the data? Do you have permission? ● What anonymization and/or aggregation techniques does your product use?	

Data exclusion	Uproar over **{product}'s** lack of **{data type}** that excludes **{user group}.**
	Praise for **{product}'s** inclusion of **{data type}** that benefits **{user group}.**
Guiding questions ● What is the downstream, real-world effect of this model's performance? ● What data is missing that would adversely impact certain user groups?	

Data ethics	Calls to boycott **{product}** over unfair treatment of **{humans involved in data collection/labeling}.**
	{Product} sets the bar for **{humans doing data collection/labeling}** by **{action taken to compensate fairly}**.
Guiding questions ● Who are the humans involved collecting and/or labeling your data? ● How are you compensating them for this critical work?	

Data transferability	**{Product}** cancelled over faulty **{data}** used from **{inappropriate source}**.
	{Product} innovates in leveraging **{data}** from **{source}** by **{action taken to transform data}.**
Guiding questions ● What risks are present for using data not originally intended for your use case?	

Data fragility	**{Product}** down as team struggles to fix **{key data input sources}.**
	{Product} outperforms competitors thanks to including **{data}** that accounts for **{real world consideration}.**
Guiding questions ● Does your data reflect the real world? e.g. for image based systems does it include off center/blurry images?	