13: AI Application Design Process

———

Juho Kim

Administrative Notes

- **Assignment #3: "Evaluating Long-Tail AI Tasks"**

  - Shorter than #1 and #2

  - To be released later this week

  - There will be another short, final assignment on multimodal interaction.

- **Milestone #2: Prototype**

  - Due: Nov 22 (Wed)

Previously on CS492G...

DATA → MODEL → INTERACTION

Today's Learning Objectives

After today's class, you should be able to…

- Understand distinct stages of desinging an AI-powered application.
- Identify major design / engineering considerations in each of the design stages.

Reflection on the last in-class activity

Future of Crowd Work

Crowdsourcing is useful in all stages of AI design.

DATA	MODEL	INTERACTION

- Data generation

- Evaluating and debugging models

- Hybrid intelligence systems
- Behavioral studies

Future of Crowd Work

"Can we foresee a future crowd workplace in which we would want our children to participate?"

- Worker considerations
  - Motivation, feedback, pay
- Requester considerations
  - Coordination, task decomposition, quality control

Kittur, Aniket, et al. "The future of crowd work." *Proceedings of the 2013 conference on Computer supported cooperative work*. 2013.

How to Stop Silicon Valley from Building a New Global Underclass

GHOST WORK

Mary L. Gray and Siddharth Suri
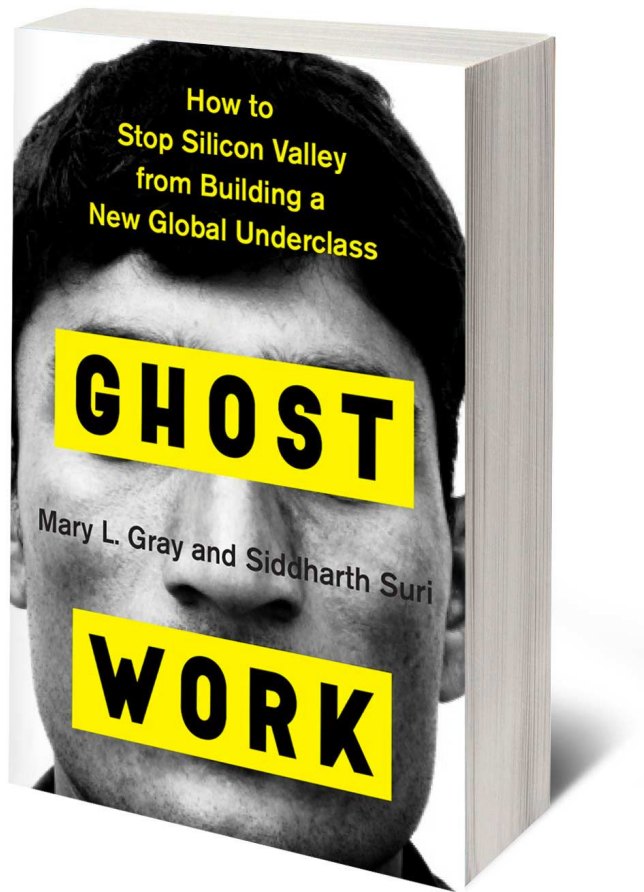
"These people doing 'ghost work' make the internet seem smart."

- Generating training data
- Flagging bad content
- Manually fixing AI errors

~8% of Americans have contributed to "ghost economy".

https://ghostwork.info/ghost-work/

Future of Crowd Work: Design Goals

- Create career ladders
  - motivation, job design, reputation, hierarchy.
- Improve task design through better communication
  - QA, job design, task assignment, realtime crowd work, synchronous collaboration, platform.
- Facilitate learning
  - quality assurance, AIs guiding crowds (and vice versa), task assignment, reputation and credentials, platform.

AI Design Process

What's unique in designing AI-powered systems?

- **Data matters**
  - Big data preferred, Cost & time associated in collecting and handling data, Bias and fairness issues
- **Model matters**
  - Probabilistic, Hard to predict, Possibly a black box, Transparency, Interpretability, Resources
- **Interaction matters**
  - Mental model, Explanation & feedback, Intelligent UI, Helping users in the task, Human-AI collaboration

Our Three-Stage Framework

DATA	MODEL	INTERACTION

DATA
- Collecting, cleaning, processing, labeling, verifying data
- Crowdsouring, Logs, User input, …

MODEL
- Train + Test + Parameter tuning
- Apply to a new problem for prediction.
- Picking the right model is challenging.

INTERACTION
- Model results (and explanations) are shown to the user through UI.
- User feedback can be fed back to the system.

Steps in applying ML to practice

7 Steps of Machine Learning

- Gathering Data
- Preparing that Data
- Choosing a Model
- Training
- Evaluation
- Hyperparameter Tuning
- Prediction

https://www.youtube.com/watch?v=nKW8Ndu7Mjw&vl=en
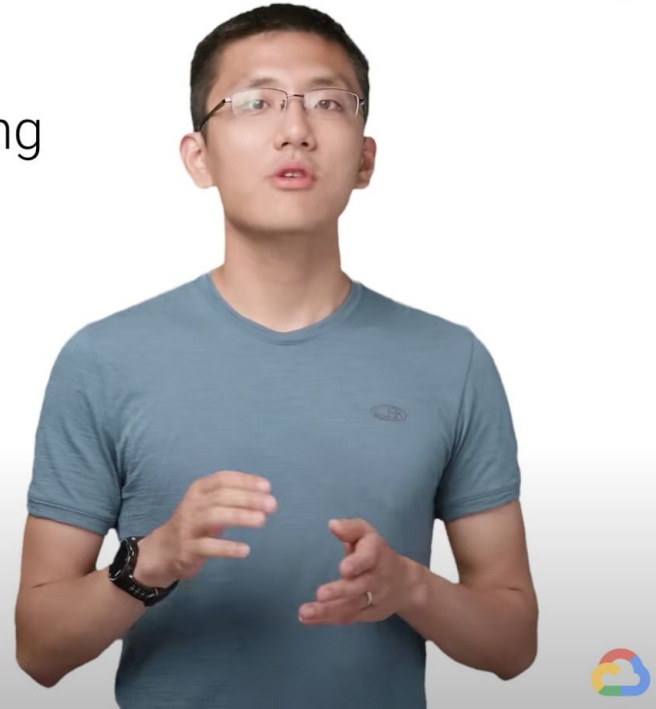
AI Design Process

-
  - *Gather data*
  - *Prepare the data*
  - *Choose a model*
  - *Train the model*
  - *Evaluate the model*
  - *Tune hyperparameters*
  - *Predict*
-

DATA

MODEL

(Human-Centered) AI Design Process

- **Set a goal & identify user needs.**
- *Gather data*
- *Prepare the data*
- *Choose a model*
- *Train the model*
- *Evaluate the model*
- *Tune hyperparameters*
- *Predict*
- **Design human-AI interaction**
- **Apply to UI**
- **Evaluate & Improve**

PLANNING

DATA

MODEL

INTERACTION

Step 1. Set a goal & identify user needs.

- User-centered design process is helpful here.
- Without this stage, it's likely to yield technology that kind of works but nobody uses.
- Netflix: *"The user either finds something of interest [within the first 60 or 90 seconds] or the risk of the user abandoning our service increases substantially,"*
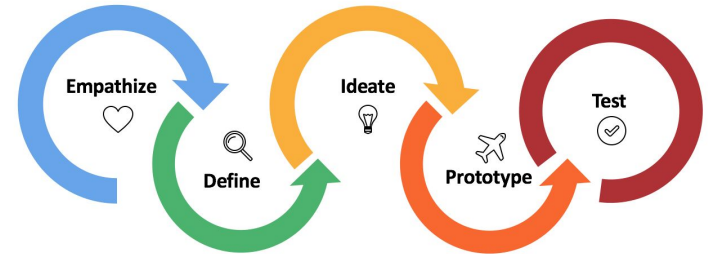
Gomez-Uribe, Carlos A., and Neil Hunt. "The netflix recommender system: Algorithms, business value, and innovation."
ACM Transactions on Management Information Systems (TMIS) 6.4 (2015): 1-19.
http://www.spring2innovation.com/2019/04/design-thinking-vs-user-centred-design/

Step 2. Gather data

- You need to know the required quantity and quality levels of the data to be collected.
- Features of the data need to be identified.
- User privacy / logging overhead / cost of collection
- Netflix
  - your interactions with our service (such as your viewing history and how you rated other titles)
  - other members with similar tastes and preferences on our service
  - information about the titles, such as their genre, categories, actors, release year, etc.
  - the time of day you watch
  - the devices you are watching Netflix on
  - how long you watch

Step 3. Prepare the data

- Processing, Labeling, Filtering
- Exploratory Data Analysis: get a sense of your data
- Segment into training/test data
- Crowdsourcing / dedicated data workers
- Netflix: 1,300 clip groups based on preference, 2,000 preference groups
  - "Netflix has hired real life humans to categorize every bit of TV shows and movies and apply tags to each of them in order to create hyperspecific micro genres such as "Visually-striking nostalgic dramas" or "Understated romantic road trip movies".

https://uxplanet.org/netflix-binging-on-the-algorithm-a3a74a6c1f59

Step 4. Choose a model

- There are tons of models available.
- Common to test multiple models and compare their performance.
- Data matters: type, size, features... of the data affect which model suits better.

Step 5. Train the model

- Iteratively improving the model performance through trial and error.
- Make predictions with a set of weights and biases, compare against the ground truth, and make improvements.

https://www.youtube.com/watch?v=nKW8Ndu7Mjw&vl=en
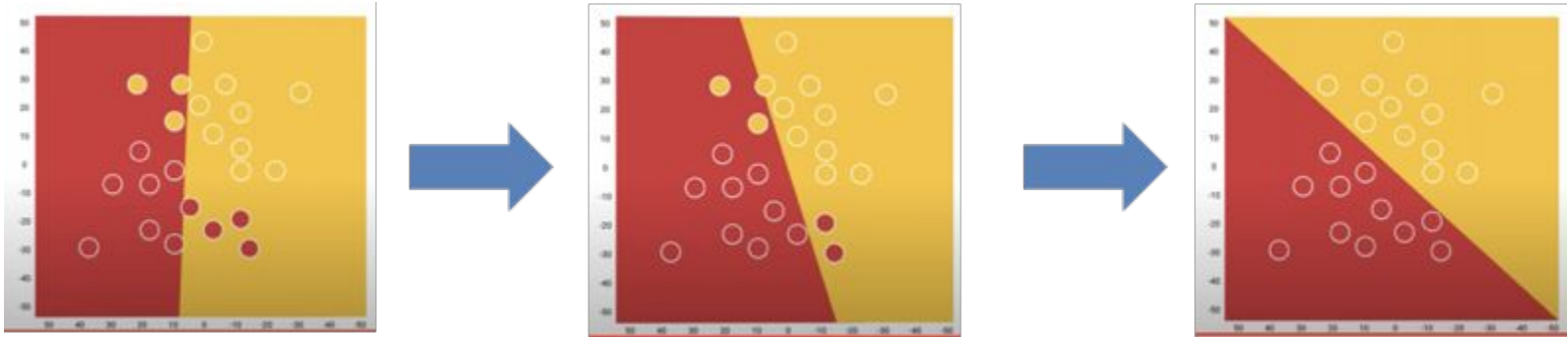
Step 6. Evaluate the model

- Apply the trained model to unseen data (test data) and see how well it performs.
- Overfitting: model too optimized for the training data
- Occam's razor: *"entities should not be multiplied unnecessarily which is interpreted as requiring that the simplest of competing theories be preferred to the more complex"* [Merriam-Webster]
- Model: abstract, simple, general solution to explain a complex phenomenon.
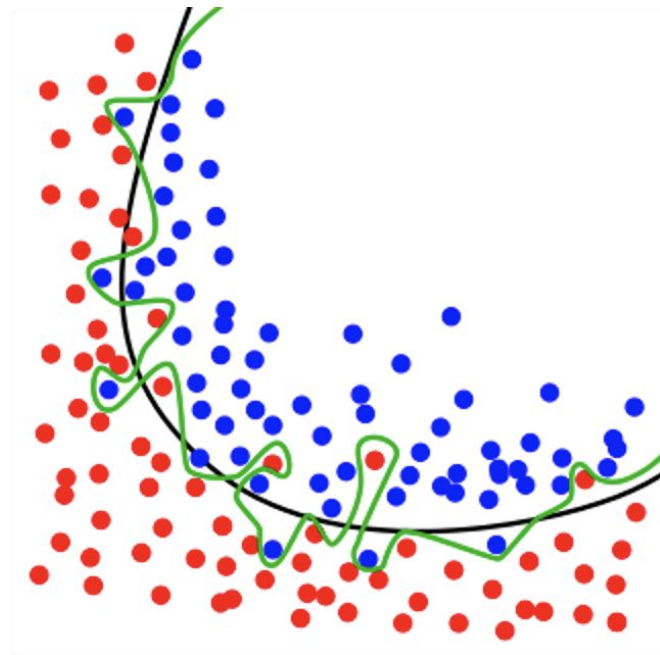
Step 7. Tune hyperparameters

- Make further performance improvements by changing various factors that comprise the model.
  - e.g., number of iterations, learning rate
- Experimental process: Often relies on experience and best practices

Step 8. Predict

- Now apply the model to a real target scenario so that it is used to make a (hopefully accurate) prediction.
  - e.g., Netflix recommends clips a user is likely to watch.
- This is where AI truly shines (if done well): relying on model prediction vs manual judgment or hard-coded rules
  ⇒ more efficient and scalable
- Practical issues
  - Content: does the model fare well against new content?
  - Users: does the model fare well against new users?

Step 9. Design human-AI interaction

- Now come back to the user and think how the model can be used to truly address the user's needs.
- UX of AI matters.
    - User motivation
    - Mental model
    - Risk
    - Trust
    - FAccT
    - Interpretability
    - …
- "How much AI is okay?"

Step 10. Apply to UI

● Combine the human-AI interaction model and the AI model, and integrate them into a UI.

Step 10. Apply to UI

- Thinking about how AI can enhance UX is important.
- Thinking how AI can damage UX is (even more) important.

https://netflixtechblog.com/selecting-the-best-artwork-for-videos-through-a-b-testing-f6155c4595f6
https://uxplanet.org/netflix-binging-on-the-algorithm-a3a74a6c1f59

Step 11. Evaluate and iterate

- Use various metrics to measure the success through a live deployment: beyond just the model performance.
- Often combined with A/B testing for more informed decision making.
- Metrics in practice: click-through rate, time to completion, diversity of options presented, explainability, ...
- Establish a process for iterative improvement.

Neil Hunt, 2014 (former Netflix CPO)

Step 11. Evaluate and iterate

Quantifying Netflix Benefits

A good choice leads to a complete viewing
A poor choice leads to abandonment, and risk of cancel

10% "better" choices → +500M/month good outcomes
If 1% of those avoids a cancellation → $500M/year

Our measurement thresholds:
    0.1% retention improvement ($5..50M/year)
    0.1% more viewing per time period

Neil Hunt, 2014 (former Netflix CPO)

(Human-Centered) AI Design Process: THOUGHTS?

- **Set a goal & identify user needs.**
- *Gather data*
- *Prepare the data*
- *Choose a model*
- *Train the model*
- *Evaluate the model*
- *Tune hyperparameters*
- *Predict*
- **Design human-AI interaction**
- **Apply to UI**
- **Evaluate & Improve**

PLANNING

DATA

MODEL

INTERACTION

What parts of it has changed with genAI/LLM?

- Why do you think these changes occurred?
- What do these changes mean?

- **Set a goal & identify user needs.**
- *Gather data*
- *Prepare the data*
- *Choose a model*
- *Train the model*
- *Evaluate the model*
- *Tune hyperparameters*
- *Predict*
- **Design human-AI interaction**
- **Apply to UI**
- **Evaluate & Improve**

- UX methods, needfinding
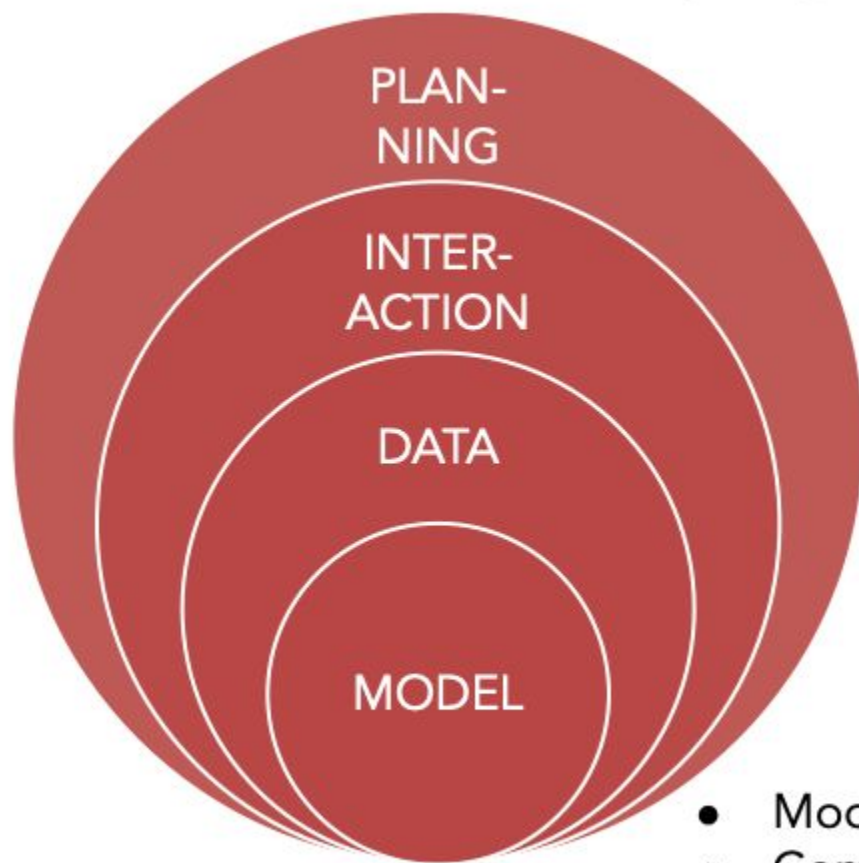
- Model performance to interaction performance
- Mental model, Explanation & feedback, Intelligent UI, Human-AI collaboration

- Data pipeline
- Collecting, labeling, processing cost
- Bias, fairness

- Model performance
- Computational cost
- Interpretability, transparency

PLAN-NING

INTER-ACTION

DATA

MODEL

MODEL-CENTRIC DESIGN PROCESS

Here's a cool model.
Let's build a better performing model.

Here's a cool model.
What can we build with it?
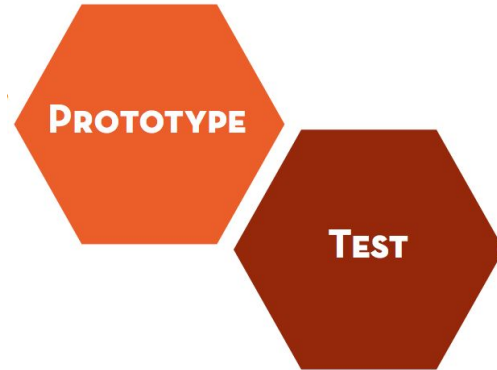
ITERATIVE USER-CENTERED DESIGN PROCESS

Interaction post-hoc ▯
guide the design & research process

Combine model- & data-centric approaches

Empathize

Define

Ideate

Prototype

Test

Iteration

Image Credit: Stanford d.school
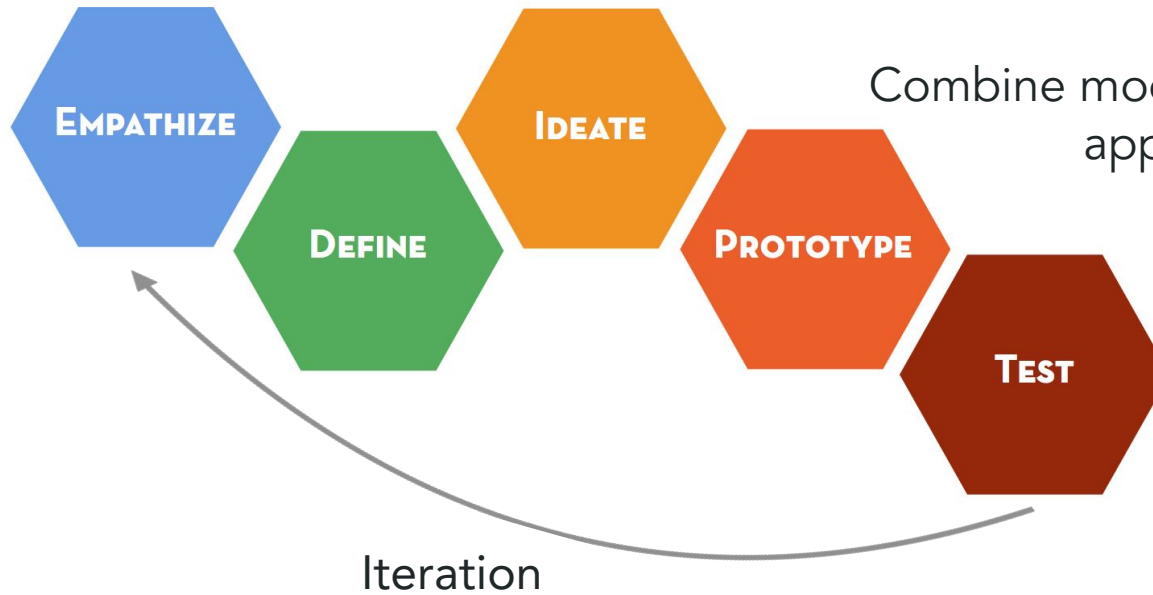
ACTIVITY: KAIST course recommendation system

- As designers of such a system, what are some guiding questions in each of the stages in the design process we should ask ourselves?
- Teams of 3-4, 10 mins

yellkey.com/any

Issues in the AI design process

- There exists a large technical debt in ML systems.
- Data handling consumes too much time / cost.
- Communicating about data / model is difficult.
- Not much attention is given to human-AI interaction.
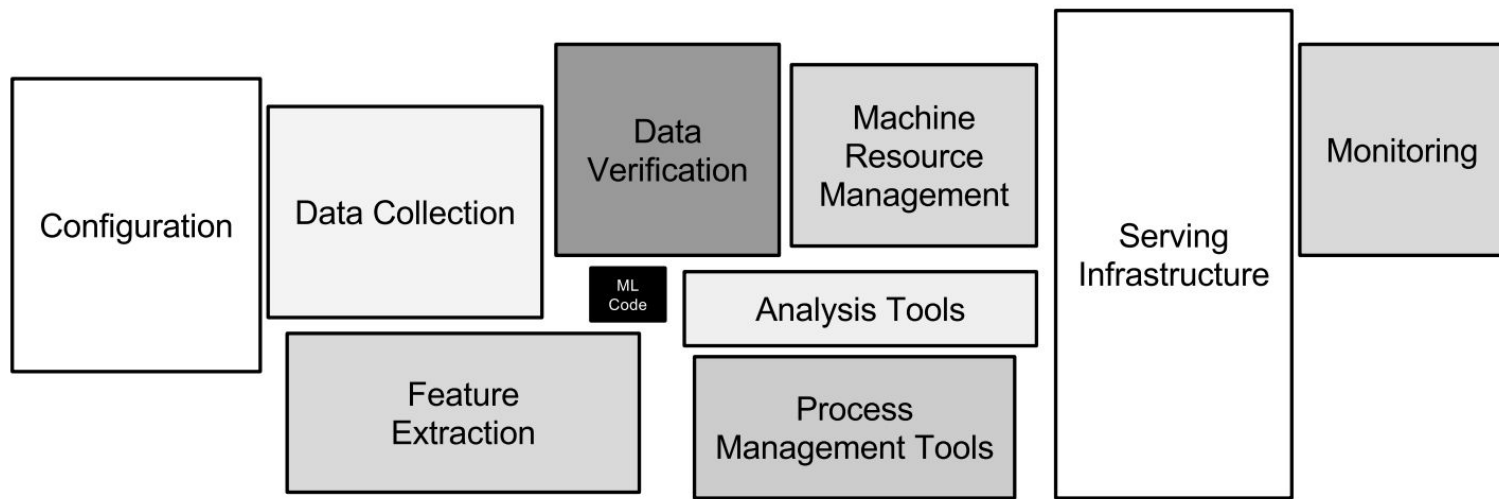
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Sculley, David, et al. "Hidden technical debt in machine learning systems." Advances in neural information processing systems. 2015.

Data work consumes most time.

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%
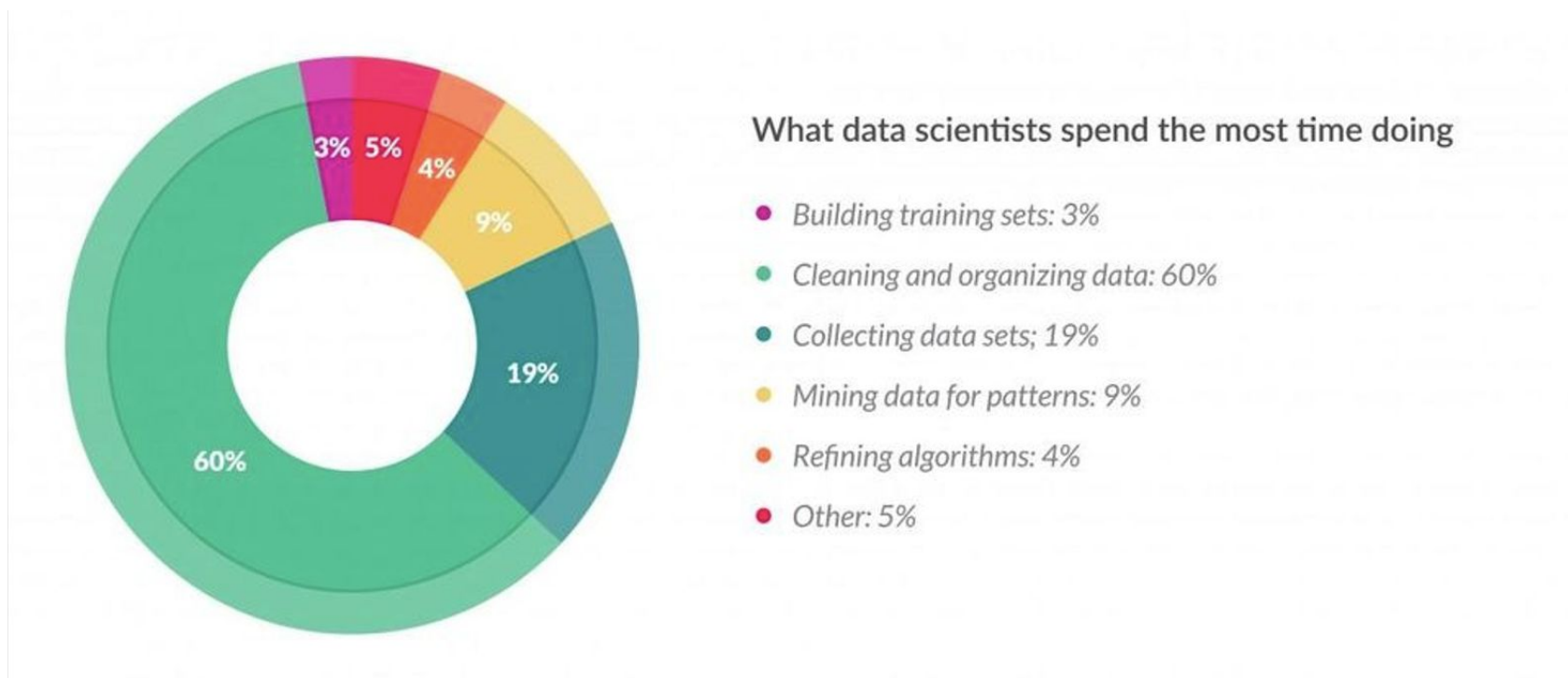
Model Cards

- Short document about a trained ML model, with its intended use and performance characteristics.
  - Goal: To help users decide whether and how to apply the model to their context.
- A structured communication medium (like a spec sheet for hardware devices and electrical components) to be shared across different stakeholders.
- Also, data sheets for data

Gebru, Timnit, et al. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018).
Mitchell, Margaret, et al. "Model cards for model reporting." Proceedings of the conference on fairness, accountability, and transparency. 2019.
https://modelcards.withgoogle.com/about

ACTIVITY: Let's crowdsource model card generation.

- Let's fill in the missing information to complete a model card for an image cropping AI.
- Groups of 3-4, 20 mins

www.yellkey.com/special