session 1

how to start (almost) any project

ml.school

1 problem understanding

2 problem framing

3 data strategy

4 labeling strategy

5 initial prototyping

This is not a $9.99 machine learning course.
This is a program about **fundamental principles** and **timeless ideas** for building machine learning systems.

These sessions will help you **unlearn what you think machine learning is**. They will show you what it takes to build production-ready systems that work.

To succeed in the program, attend the **live sessions**, follow the **code walkthroughs**, complete a few **assignments**, **ask questions**, and **help others**.

The biggest challenge we all face when starting a project is fooling ourselves into solving the **wrong problem** or getting stuck with the **wrong solution.**

best practice

Start every project with a **discovery** phase.
Focus on **understanding** and **framing** the problem.
Finish this phase by building an initial **prototype**.

Customers don’t know what they want.

Help customers identify any **friction points**, explore any **raw edges**, and **ask** the **right questions**.

What's the problem we are trying to solve?

Define the **project scope** and identify the **critical path** to an initial prototype.

What related problems do we want to ignore?

The difference between a mediocre solution and a great one lies in the **rabbit holes** we **avoid**.

Who is our true customer and why do they care?

Cut out the **middleman**. Identify who **benefits the most** and why this is **important** to them.

What do existing solutions look like?

Investigate past **solutions**. Learn **what works**, **what doesn't**, and avoid repeating old **mistakes**.

problem you are solving | problems you are ignoring | customers and why they care | existing solutions | measuring success

How are we going to measure success?

Define a **benchmark**. We can't move without ways to **assess progress** and guide future **improvements**.

best practice

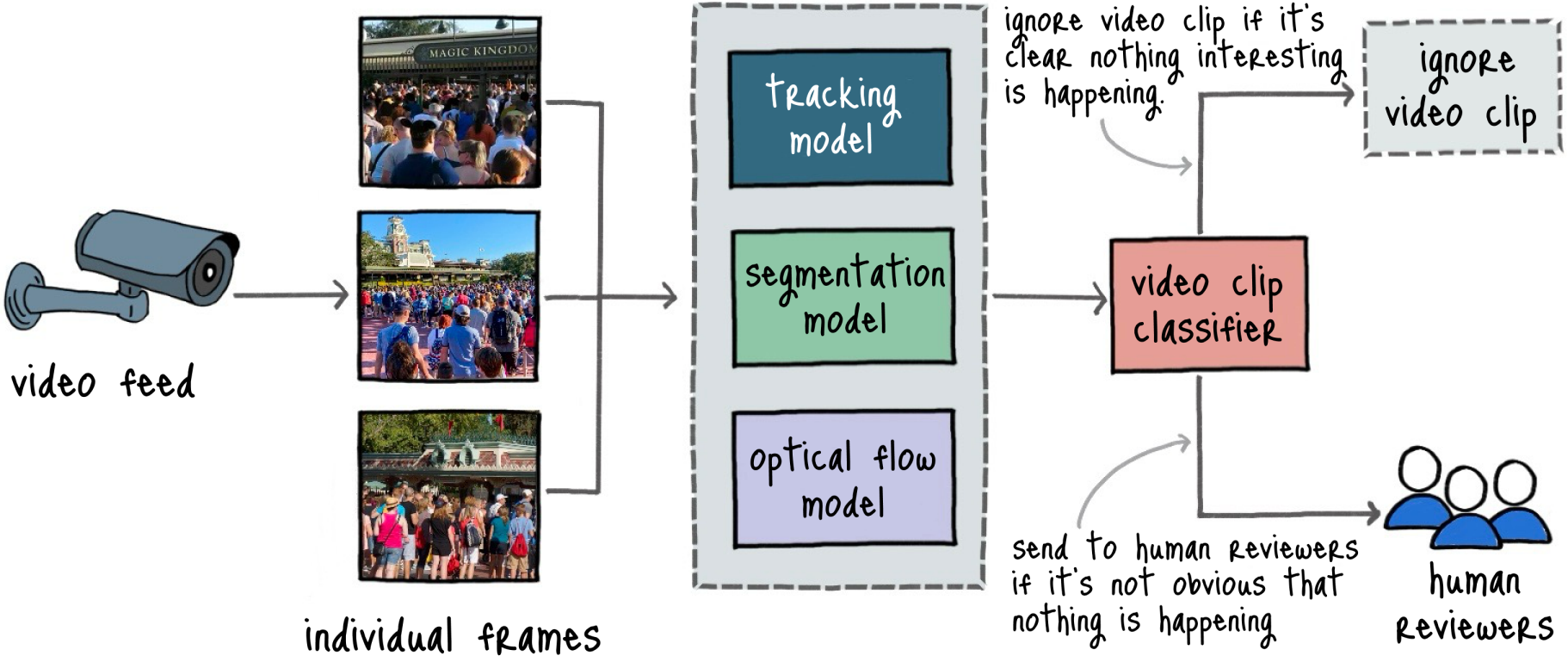
Use questions as your primary tool to uncover **friction points**, challenge **assumptions**, explore the problem's **context**, and define what **success** looks like.

You don’t win with better algorithms or better data.

You win by **framing problems** so that a **useful solution** becomes **inevitable**.

framing full self-driving

people breaking into disney parks



best practice

Use the **haystack principle** to frame problems.

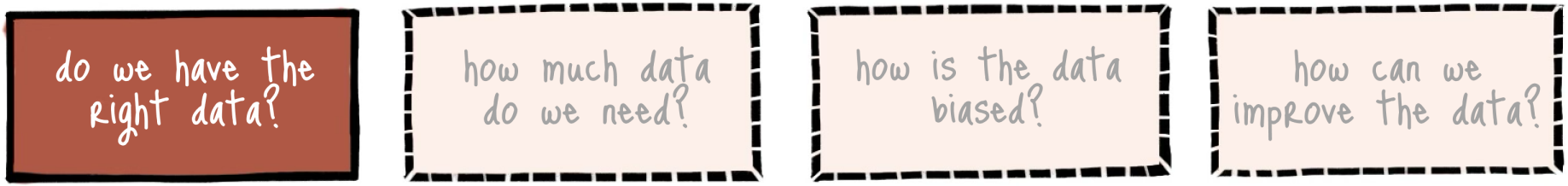
Instead of searching for a **needle** in a haystack,

shrink it until finding the needle becomes inevitable.

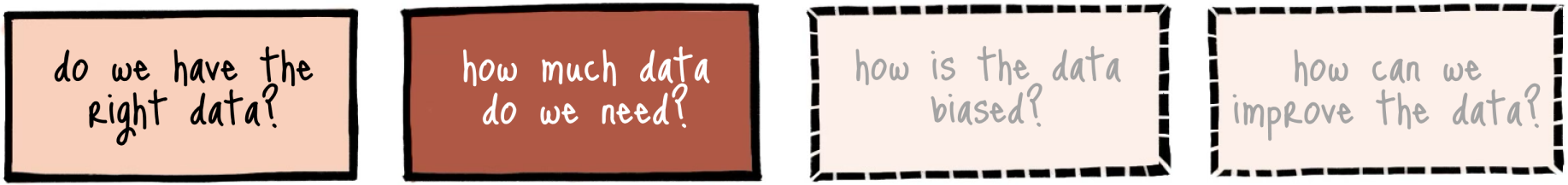
Good datasets contain **predictive** features, have **high-quality**, **diverse** samples, and are **sized** properly so that models can **generalize** effectively.

best practice

Build and **deploy** a **working solution** to capture **production data.** This is the most **effective** way to gather **meaningful, reliable** data to solve a problem.

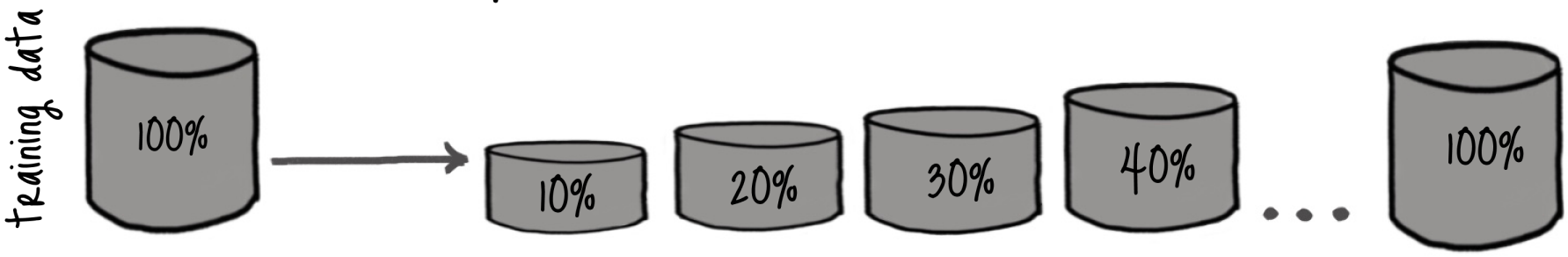


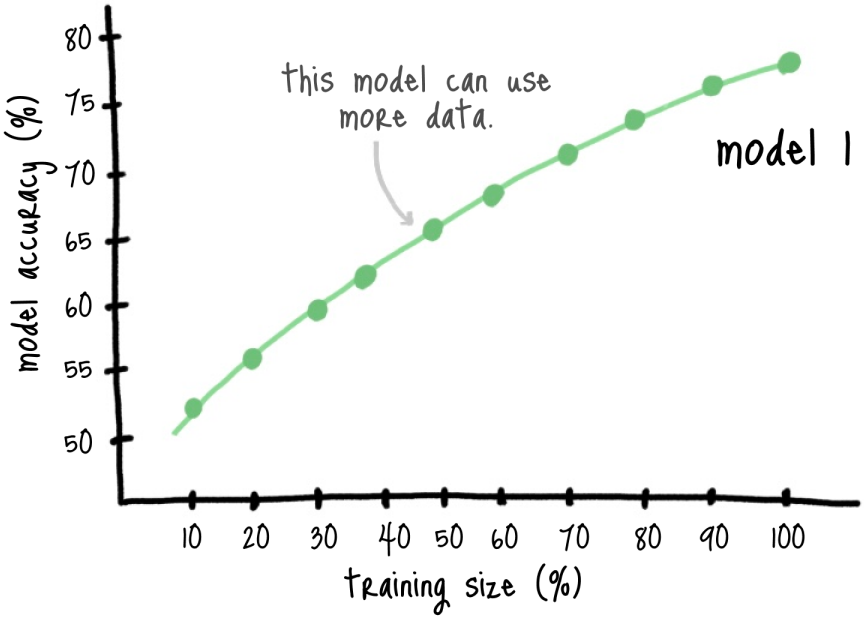
Compare the **data you have** with the **data you need.**
Then plan how to **collect**, **process**, **store**, **secure**,
and **maintain** what's missing.

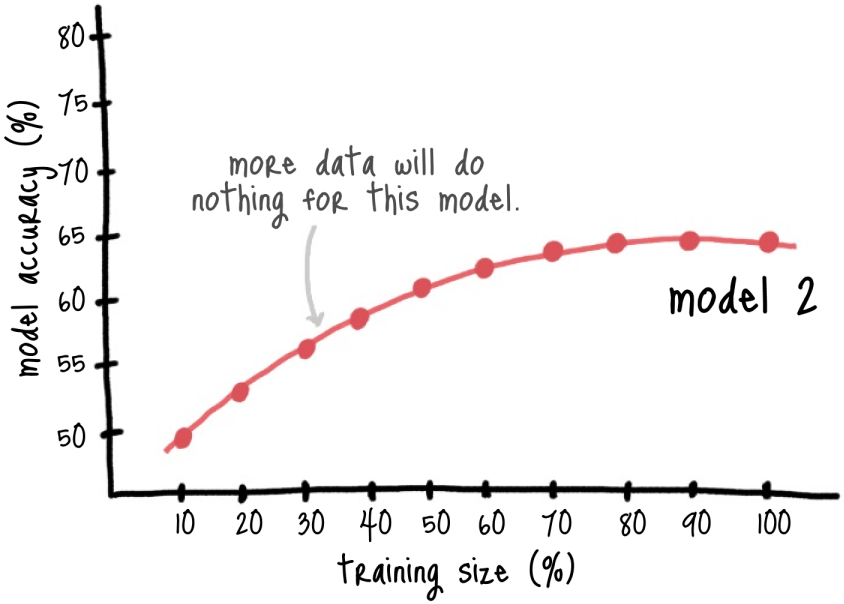


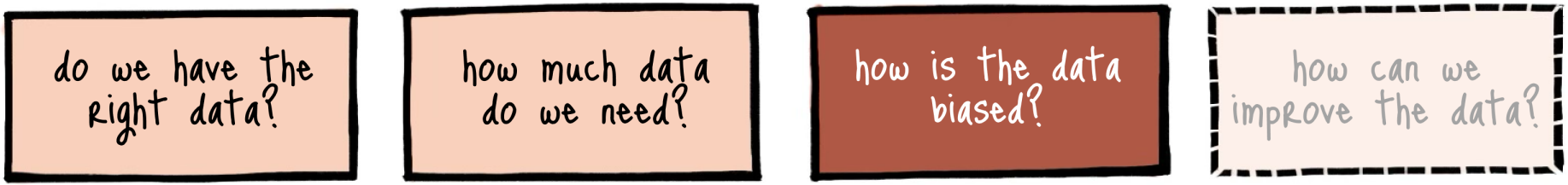
More data isn't always **better**—and it's often **worse.** Prioritize collecting **better data** first, and only scale its volume if you can clearly justify its **value.**

the value of additional data





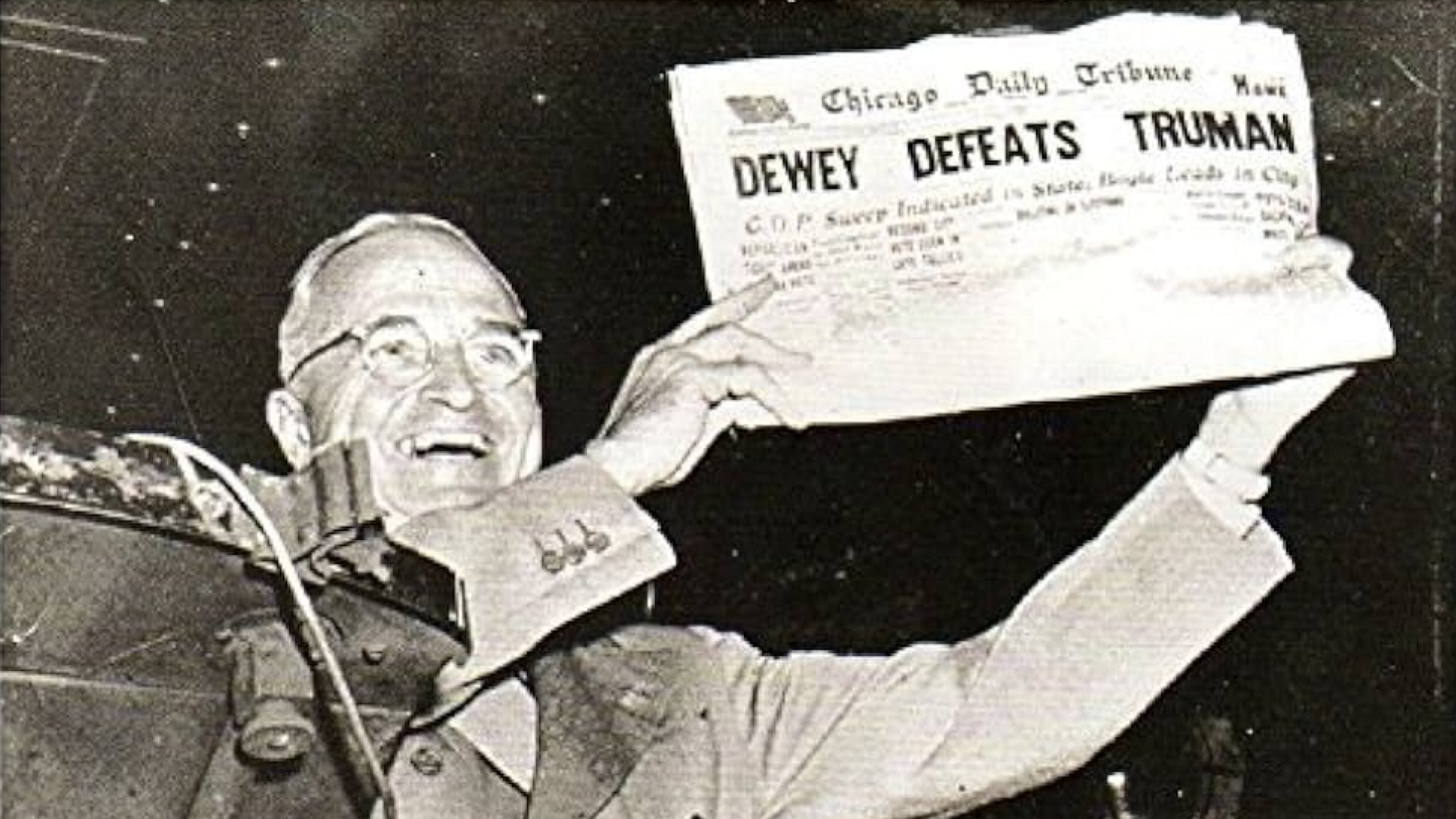


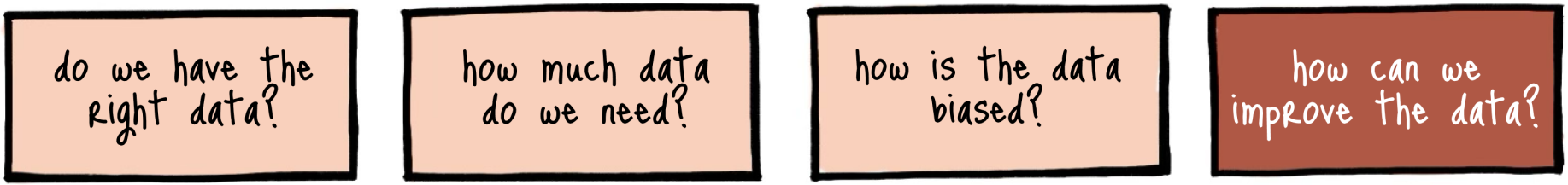


Selection bias is inevitable when collecting data, and it can't be fixed by a model-centric approach.

Awareness is the best antidote against biases.

Common causes of selection bias are **time** (when), **location** (where), **demographics** (who), **response bias** (what), and **availability bias** (convenience).


Chicago Daily Tribune
DEWEY DEFEATS TRUMAN



Better data is **better** than **better models.**

Throughout a project, keep refining your data to improve its general **quality**, **diversity**, and **size**.

best practice

Augment your data with **metadata** that adds context. This metadata is essential for **mitigating biases** and **building** more **accurate** and **reliable** models.

Models need **ground truth labels.**

Some problems require **human annotations,**

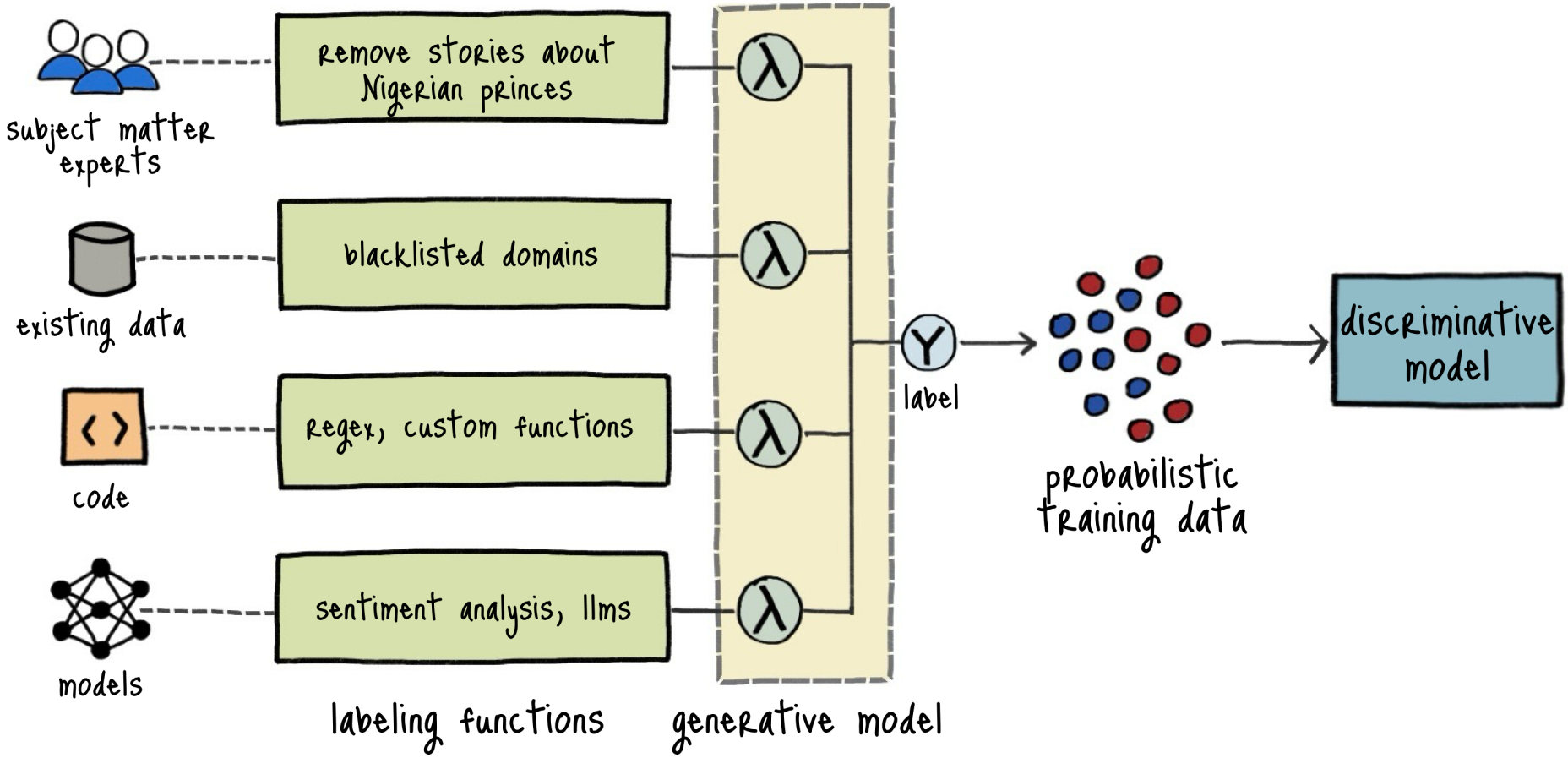
while others have built-in **(natural) labels.**

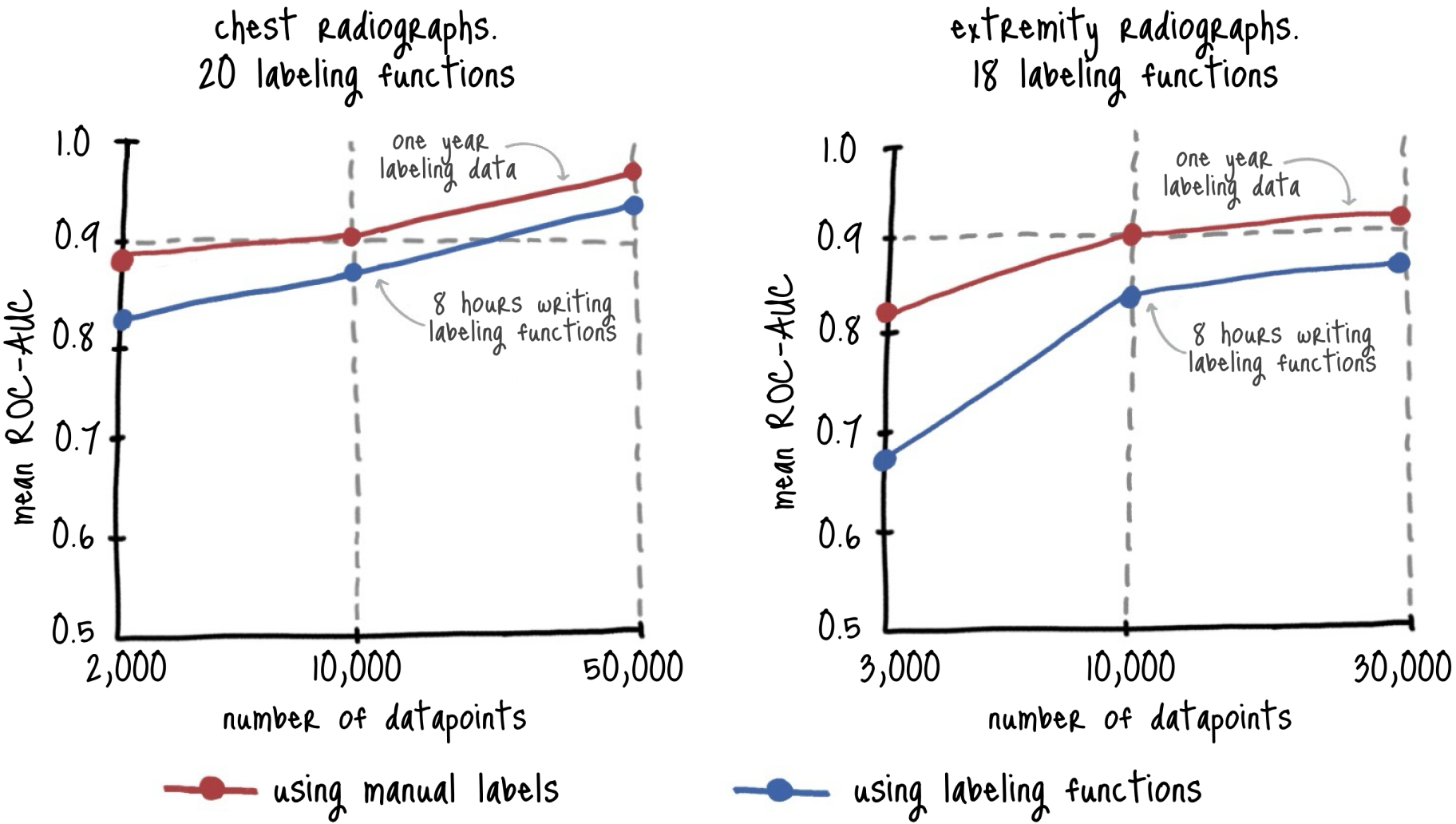
A **lack** of **labeled data** is one of the biggest **bottlenecks** in machine learning. Humans can produce **high-quality** labels, but the process **doesn’t scale**.

1 min of video @ 30 fps
20 objects per frame
60 s x 30 fps x 20 objects = 36,000 boxes
1 box per second labeling speed
10 hours to label 1 minute of video

Weak supervision is a technique that uses high-level (often **noisy**) signals to **quickly** generate **large** training sets, faster than manual labeling.

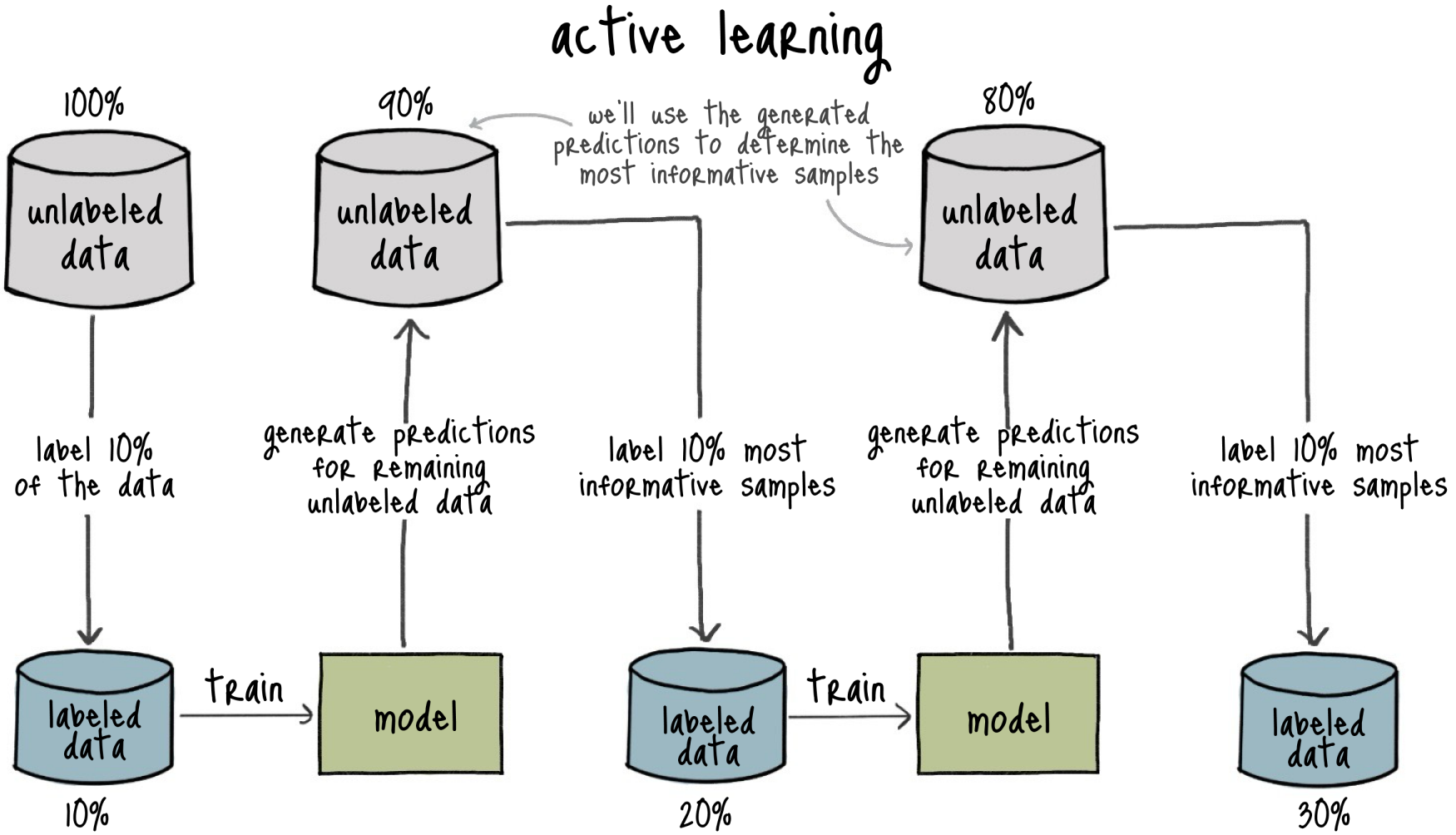
weak supervision




chest radiographs.
20 labeling functions
extremity radiographs.
18 labeling functions
mean ROC-AUC
1.0
0.9
0.8
0.7
0.6
0.5
one year labeling data
8 hours writing labeling functions
2,000
10,000
50,000
3,000
30,000
number of datapoints
using manual labels
using labeling functions

Unlike manual labeling, weak supervision is **faster**, **cheaper**, and leverages **expertise at scale—**but it can produce **noisy** or **inaccurate** labels.

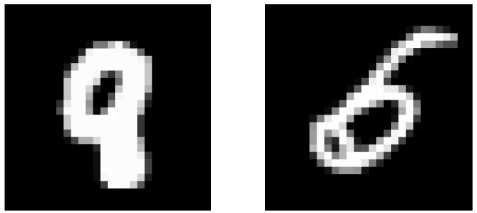
Active learning is a strategy for training a model under a **fixed** annotation **budget**. It's especially helpful when you have **abundant unlabeled data**.


active learning
100%
unlabeled data
label 10% of the data
labeled data
10%
train
model
generate predictions for remaining unlabeled data
90%
unlabeled data
we'll use the generated predictions to determine the most informative samples
label 10% most informative samples
labeled data
20%
train
model
generate predictions for remaining unlabeled data
80%
unlabeled data
label 10% most informative samples
labeled data
30%

Uncertainty sampling and **diversity sampling** are two of the most common **sampling strategies** to find the **most informative** samples to annotate.

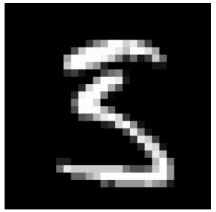
uncertainty sampling

identifies data points near a decision boundary.
these samples have a larger chance of being misclassified by the model.

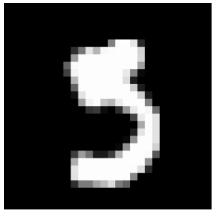


8 or 9

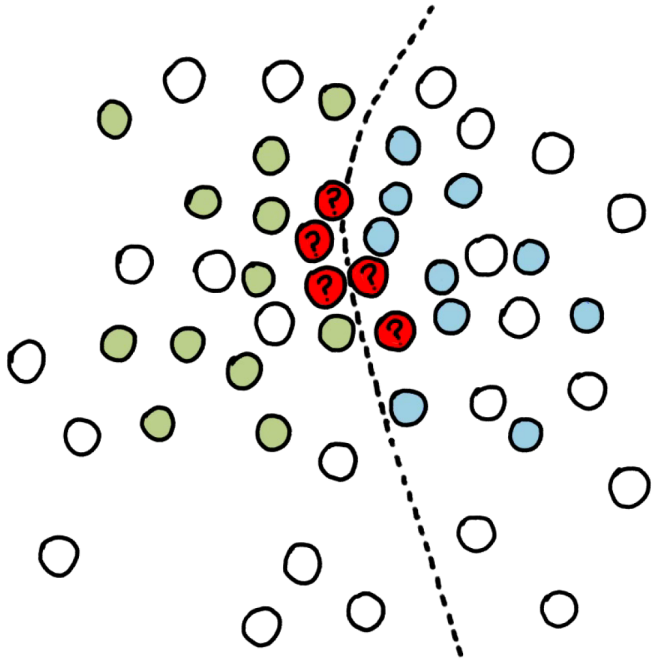
5 or 6



3 or 5

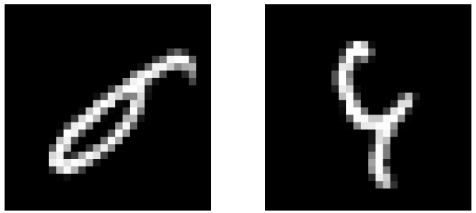


3 or 5



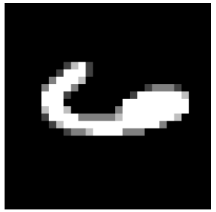
diversity sampling

identifies any unlabeled samples that are unusual, underrepresented, or unknown to the model in its current state.

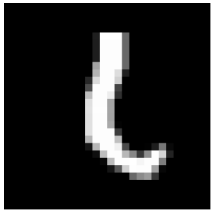


label: 6

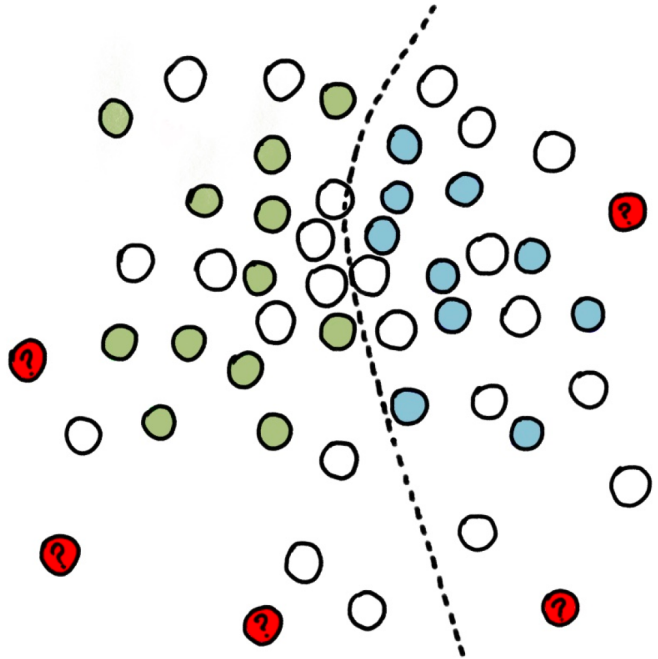
label: 4



label: 6



label: 1



Uncertainty and diversity often **work best together**. **Query-by-committee**, **margin sampling**, and **expected error reduction** are also popular methods.

best practice

Start with a **small**, **diverse** set of **high-quality human** labels to test your **assumptions**. Then use the model's **feedback** to **guide** and **scale** your labeling efforts.

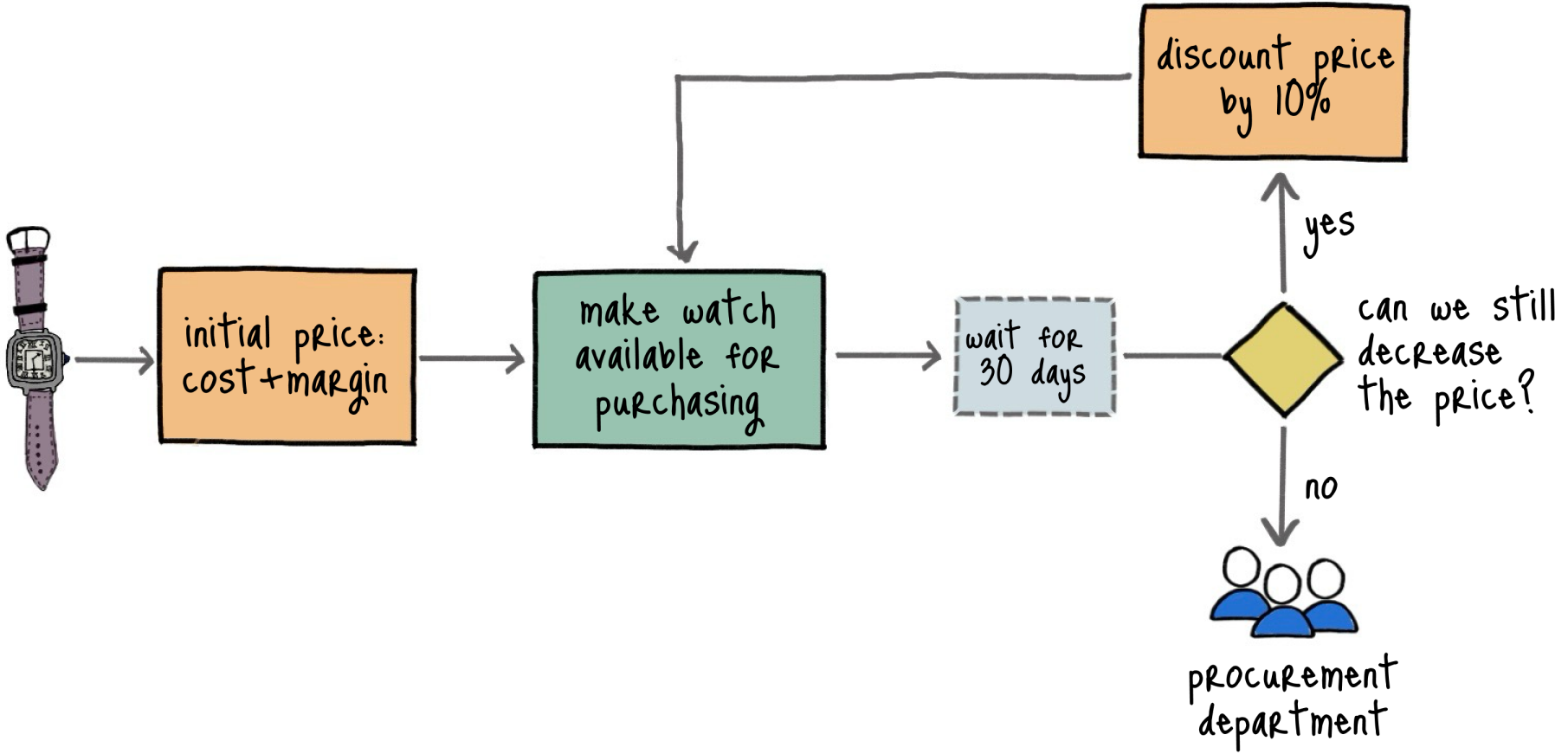
Wrap up the discovery phase of your project by building a **proof of concept**. This is the fastest way to **validate** your **understanding** and collect **feedback**.

The **first rule** of Machine Learning Engineering:

Do not start a project using machine learning.

Instead, start with **simple heuristics**.

a simple rule-based system to price luxury watches



Do things that don’t scale.

Focus on **early**, **frequent feedback** to validate assumptions, reduce risk, and adapt quickly.

best practice

Build your project's **initial prototype** by **doing** the **simplest** thing that could possibly **work.** Make it **work** first; make it **better** later.

the end