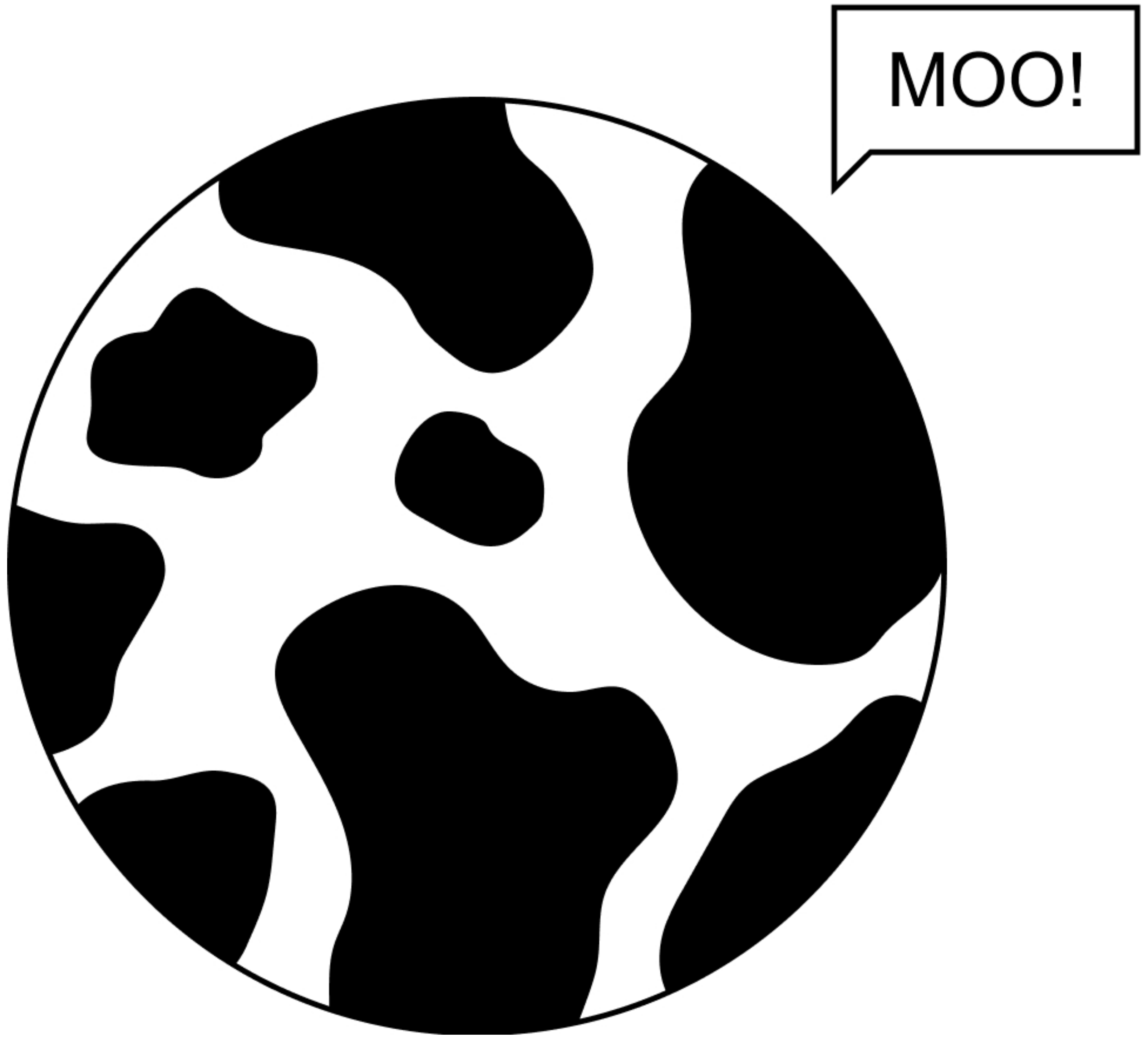
# Making ML Practical with Snorkel

Braden Hancock
Co-Founder & Head of Technology
@SnorkelAI / @bradenjhancock

# "Assume a spherical cow..."




Joke credit: https://en.wikipedia.org/wiki/spherical_cow

OPEN CORE SUMMIT

# Impractical ML assumptions are being made every day

# Outline

- **Impractical** ML assumptions are being made every day
  - 3 Impractical Assumptions
- **Snorkel OSS** was created to make ML **practical** again
  - How it Works
  - Does it Work?
  - 4 Lessons Learned
- **Snorkel Flow** is a **platform** for building AI applications, powered by Snorkel technology
  - 4 Guiding Principles

ASSUMPTION #1:

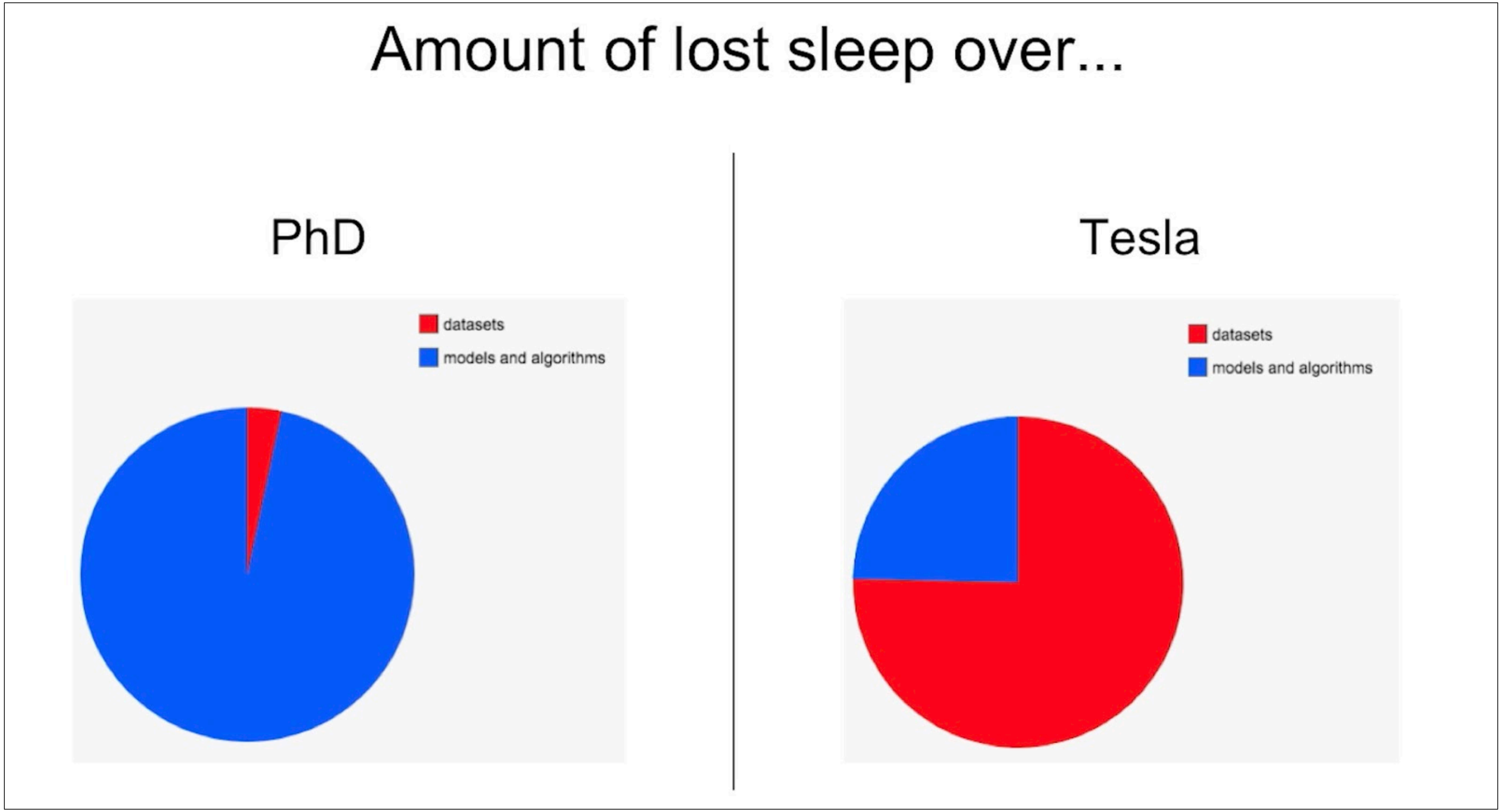
# "Assume a large, high-quality, task-specific training dataset..."

# ML in Academia

- Download the dataset
- Train a model

# ML in Industry

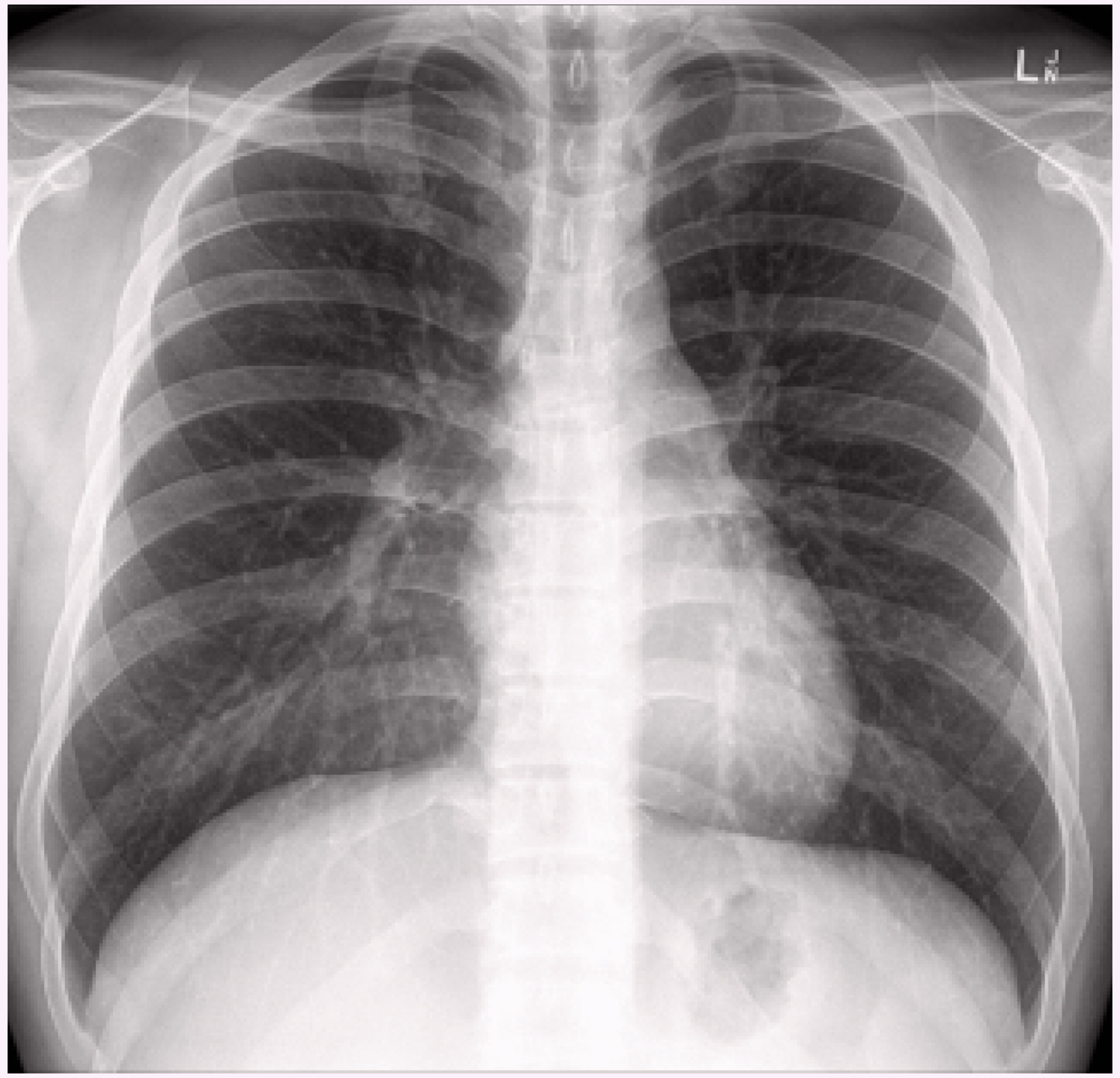
- Create a dataset
  - What training set do I need?
  - How should I collect it?
  - How large does it need to be?
  - Who is allowed to see this data?
  - Do I have enough time/budget for this?
- Train a model
- Refine dataset
  - Are these labels accurate enough?
  - How should I resolve labeler disagreements?
  - What was annotator 42 thinking!?
- Train a model
- Refine dataset
  - …




Karpathy (2018)

ASSUMPTION #2:

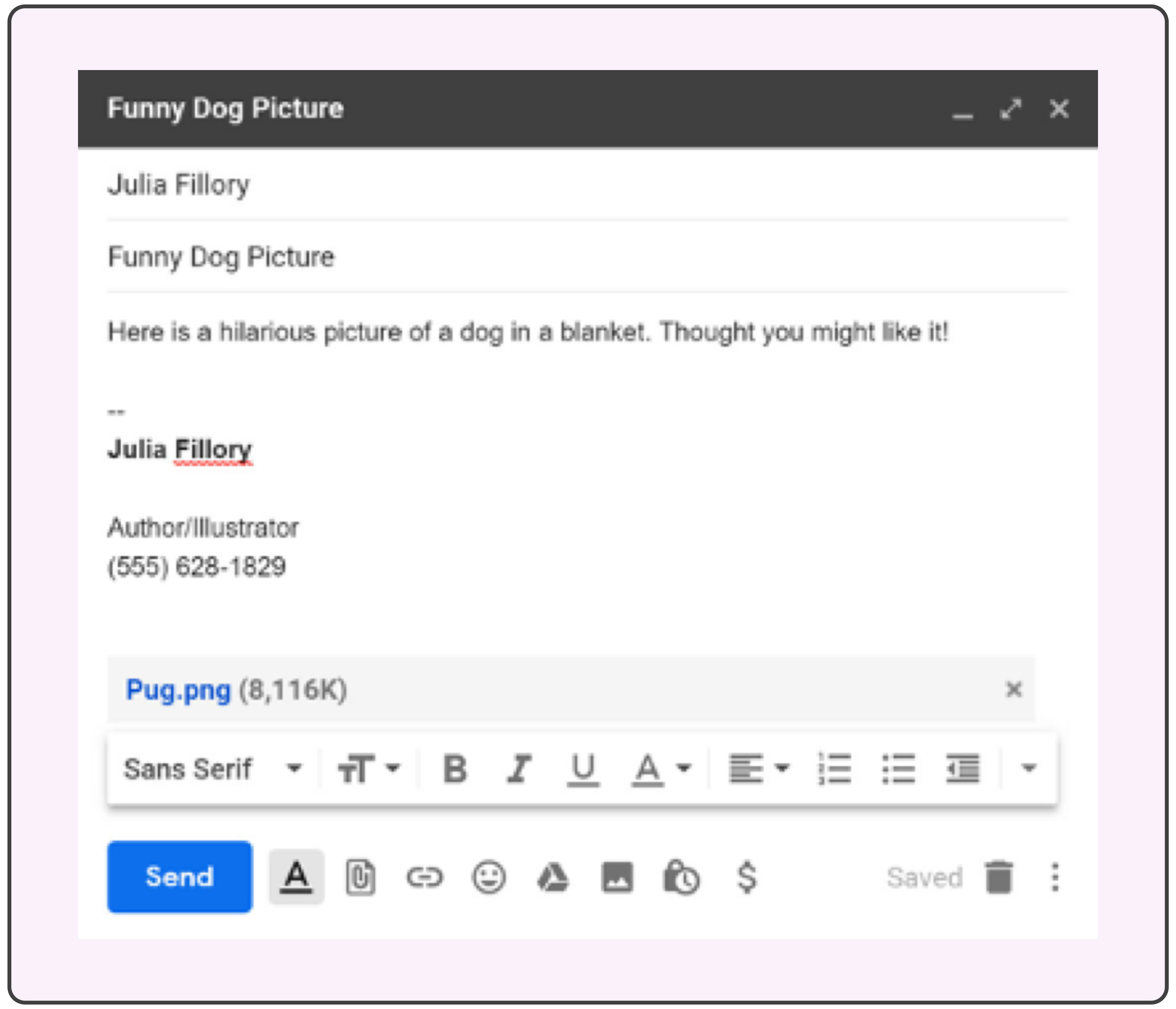
# “Assume an infinite pool of qualified annotators…”



# Expertise Limited

Who has the expertise to label these chest X-rays?*

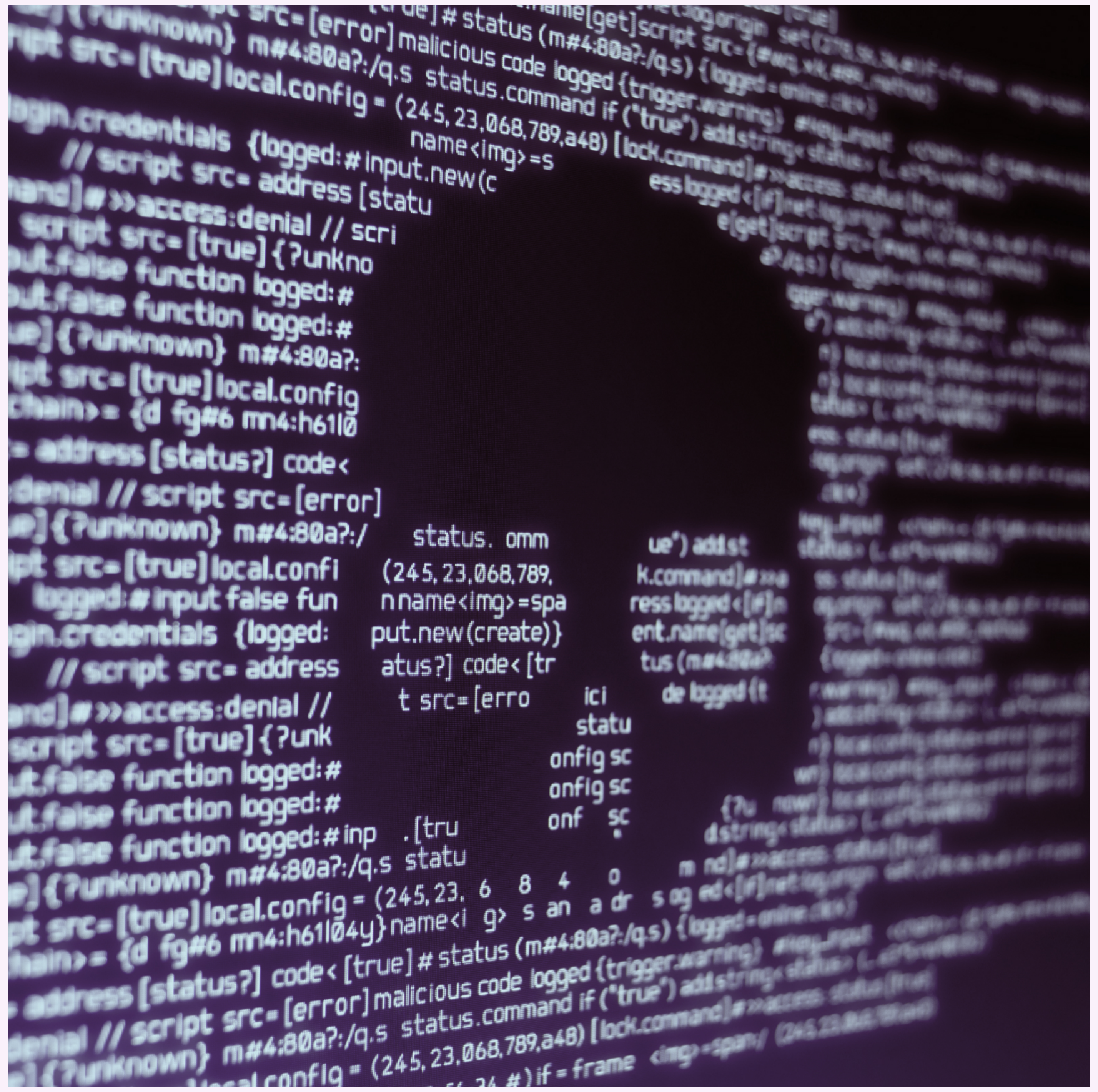
*Based on a real-world deployment of Snorkel




# Privacy Limited

Who should be allowed to read these personal emails?*

*Based on a real-world deployment of Snorkel



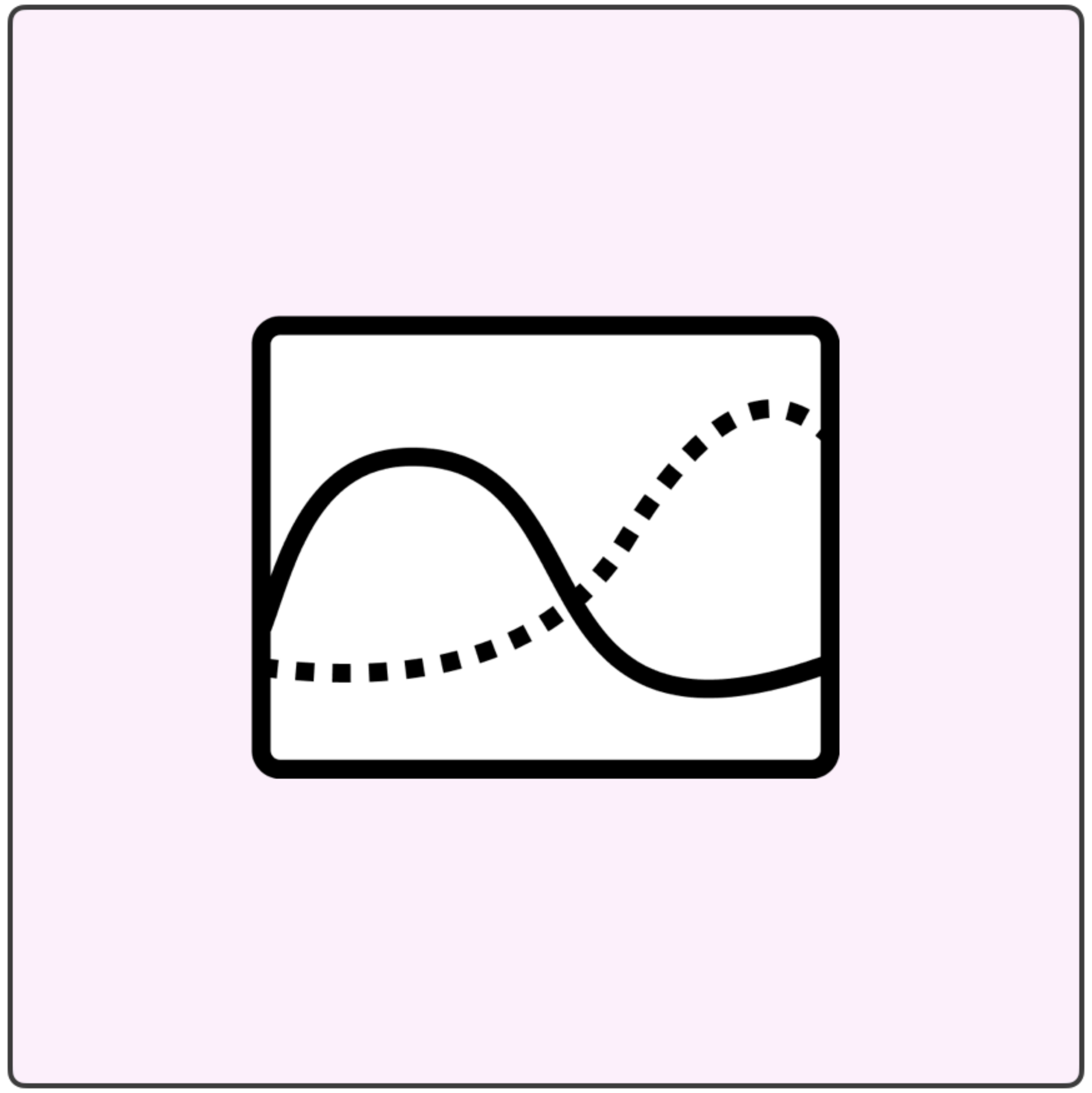
# Latency Limited

Who can turn around new labels fast enough for us to react to new failure modes?*

*Based on a real-world deployment of Snorkel

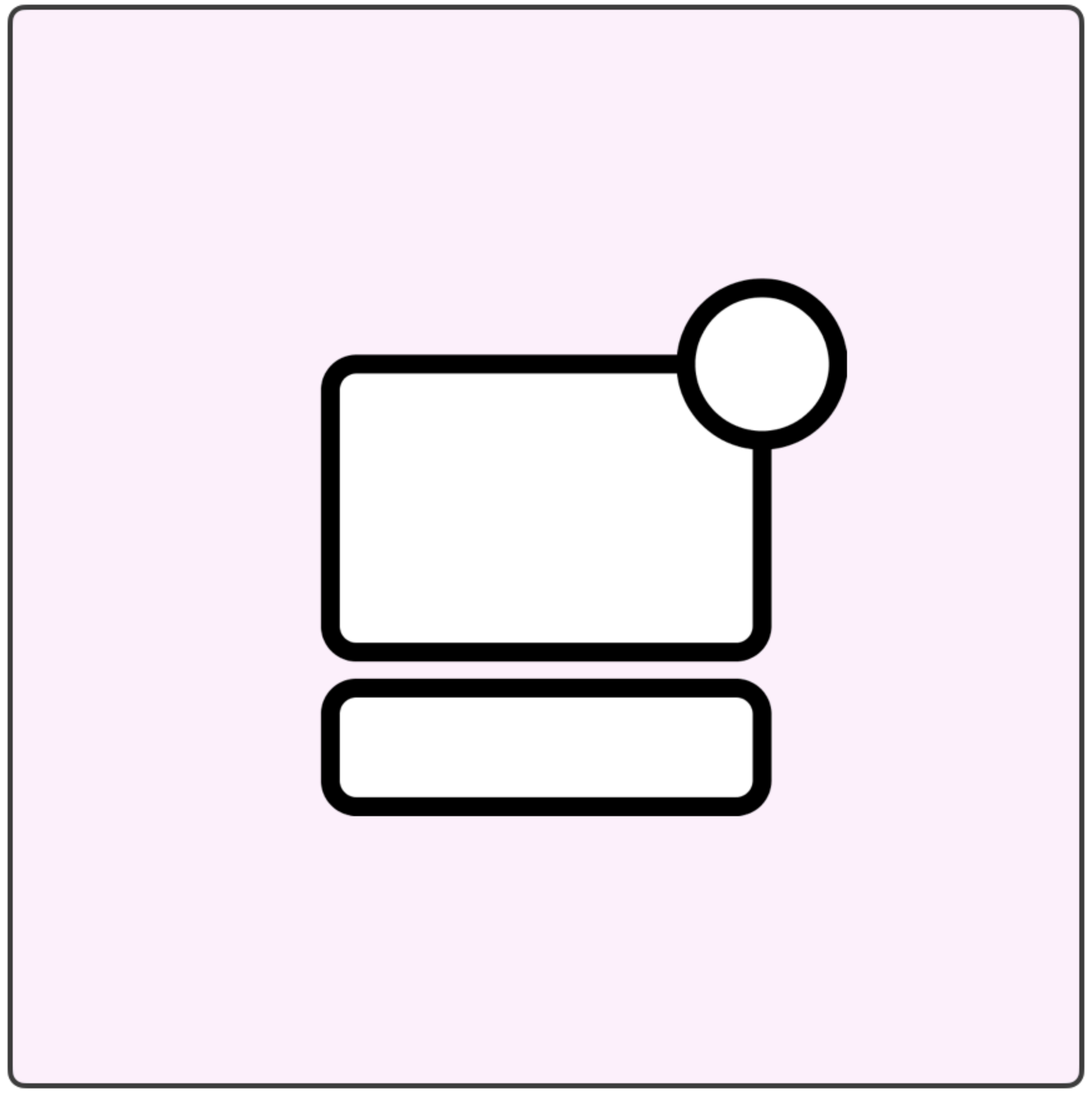
ASSUMPTION #3:

# “Assume a static test distribution…”



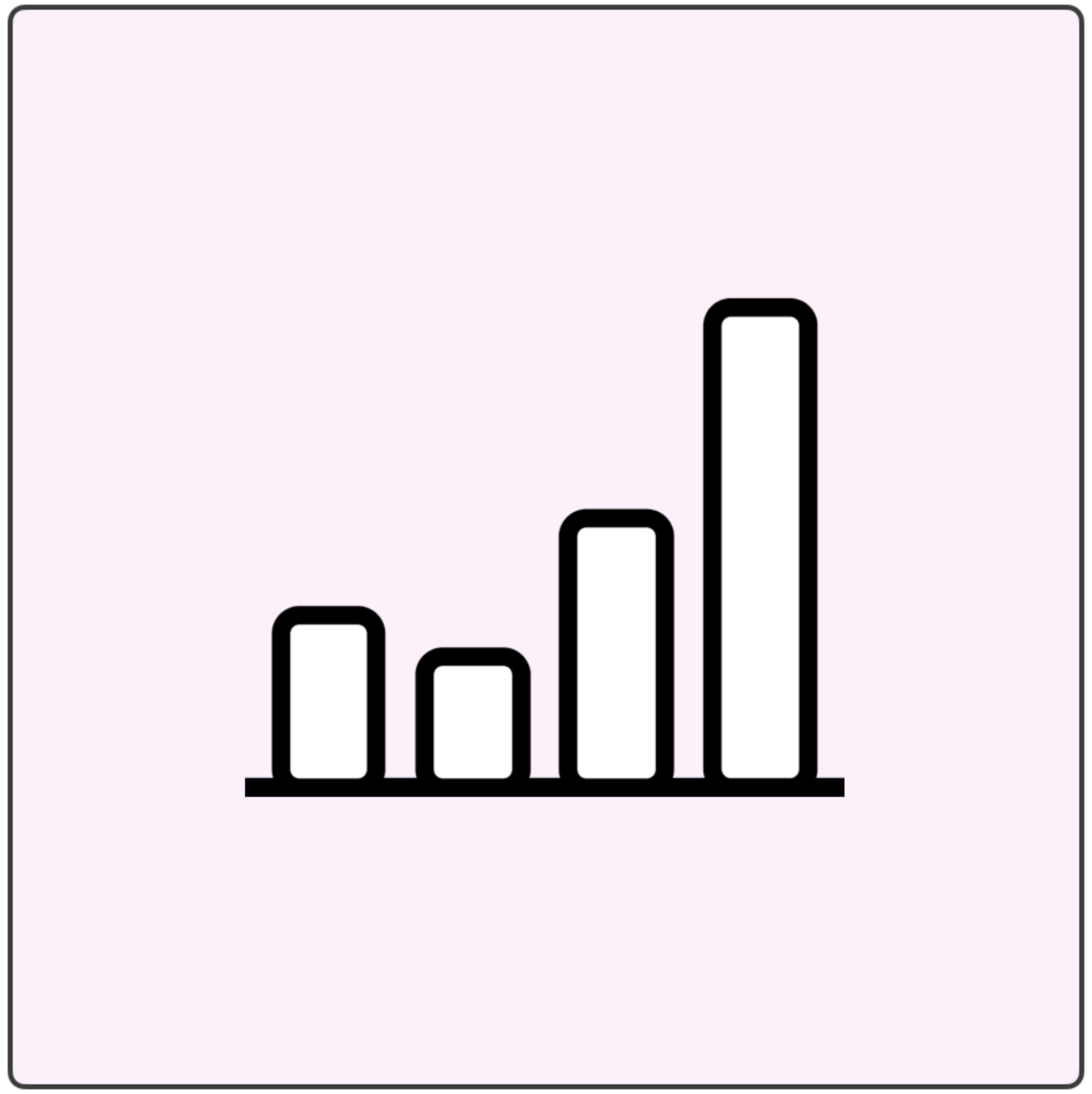
# Semantic drift

Training labels have a shelf-life before they're no longer relevant



# Evolving Needs

“As of Jan. 1, 2021, violations of the terms of service will include...”



# Changing Schemas

Downstream usage now
requires finer label granularity

# Snorkel OSS was created to make ML practical again

QUESTION 1:

# What is Snorkel?

# The Snorkel OSS Project

snorkel

| | |
|---|---|
| ☆ Star | 4.1K |
| Fork | 687 |

Google · FDA · Infosys · facebook · LinkedIn · BASF We create chemistry

YouTube · NIH · NASA · Microsoft · Chegg

teradata. · accenture · Alibaba · intel · TOSHIBA · vmware

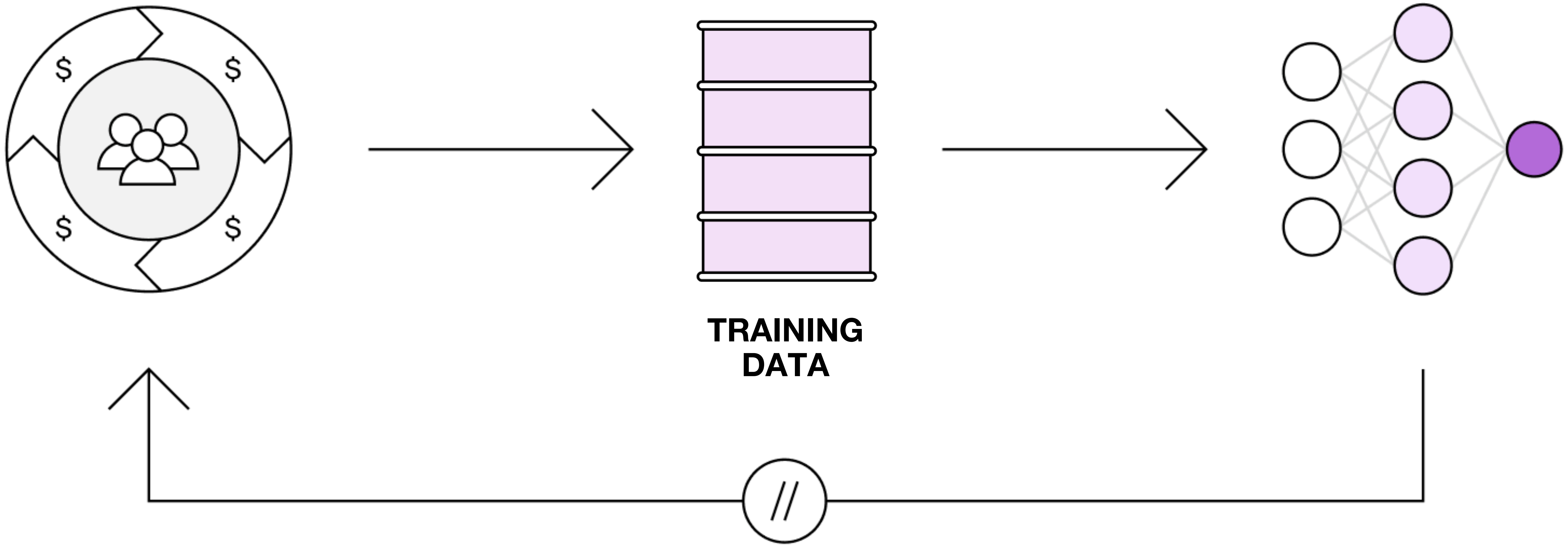
NIH National Institutes of Health · VA U.S. Department of Veterans Affairs · IBM · ANT FINANCIAL · DARPA

NEC · HITACHI · AMERICAN FAMILY INSURANCE · SAP · GORDON AND BETTY MOORE FOUNDATION

4+ year research project at the Stanford AI Lab resulting in
35+ **peer-reviewed publications** and many **production deployments**

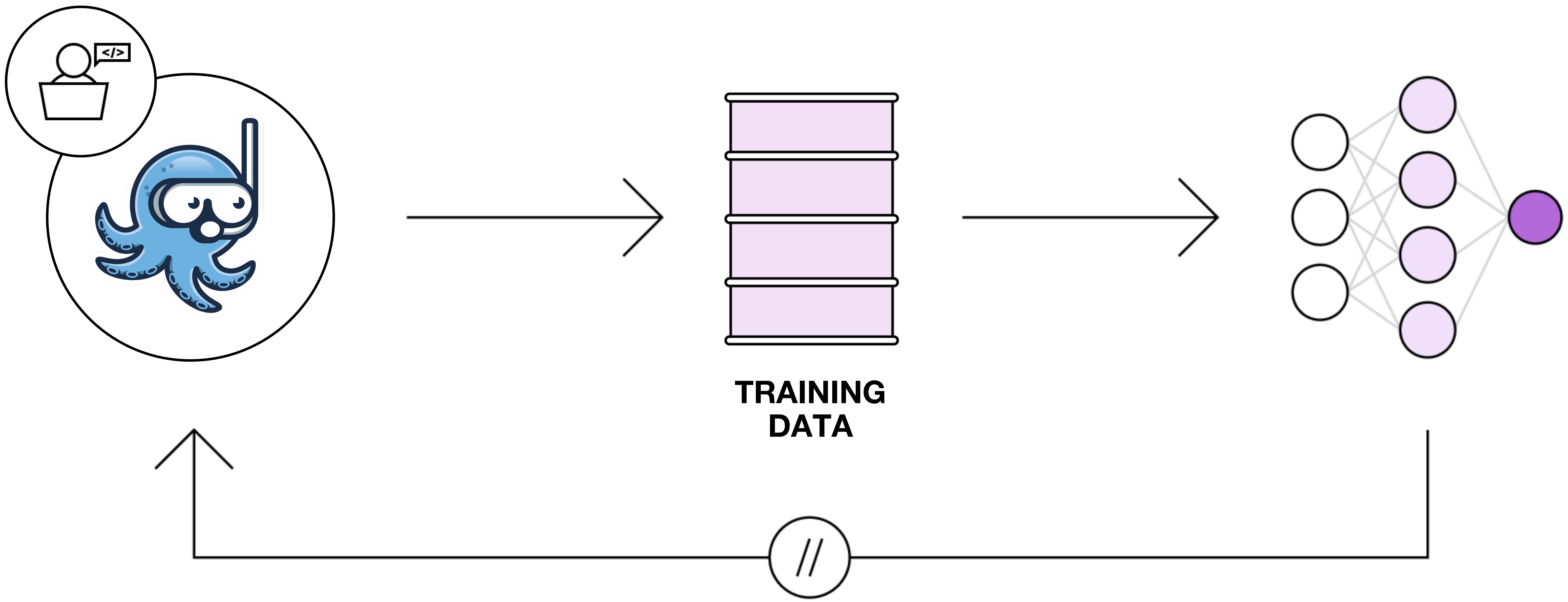
www.snorkel.ai www.snorkel.org

# Without Snorkel



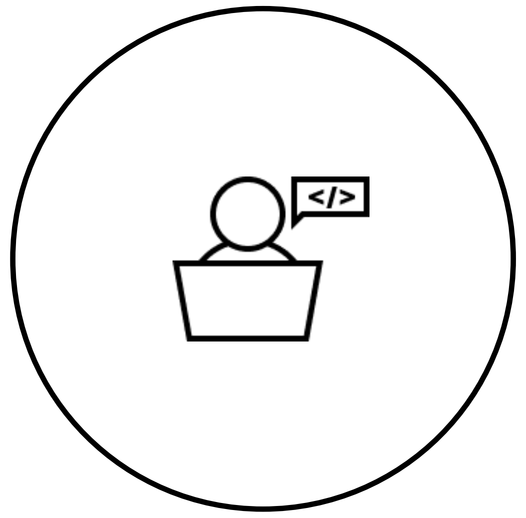

ML is blocked on collecting training data **manually**

# With Snorkel




Key idea: label, build, and manage training data **programmatically**

# What does programmatic supervision look like?



```
def LF_credit_in_title(x):
    if “credit” in get_title(x.text):
        return “Credit Agreement”
    else:
        return None
```

Labeling Functions (LFs) are **black-box functions** that output weak labels

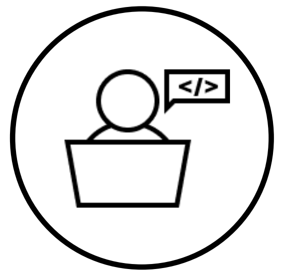


[Ratner et. al. NeurIPS’16; Ratner et. al. VLDB’18; etc.]

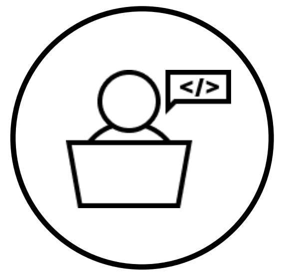
# Where do Labeling Functions come from?

| | |
|---|---|
| Pattern Matching | If a phrase like "send money" is in email |
| Boolean Search | If unknown_sender AND foreign_source |
| DB Lookup | If sender is in our Blacklist.db |
| Heuristics | If SpellChecker finds 3+ spelling errors |
| Legacy System | If LegacySystem votes spam |
| Third Party Model | If TweetSpamDetector votes spam |
| Crowd Labels | If Worker #23 votes spam |

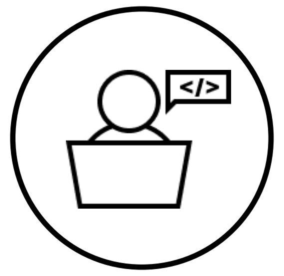
# Making weak labels strong



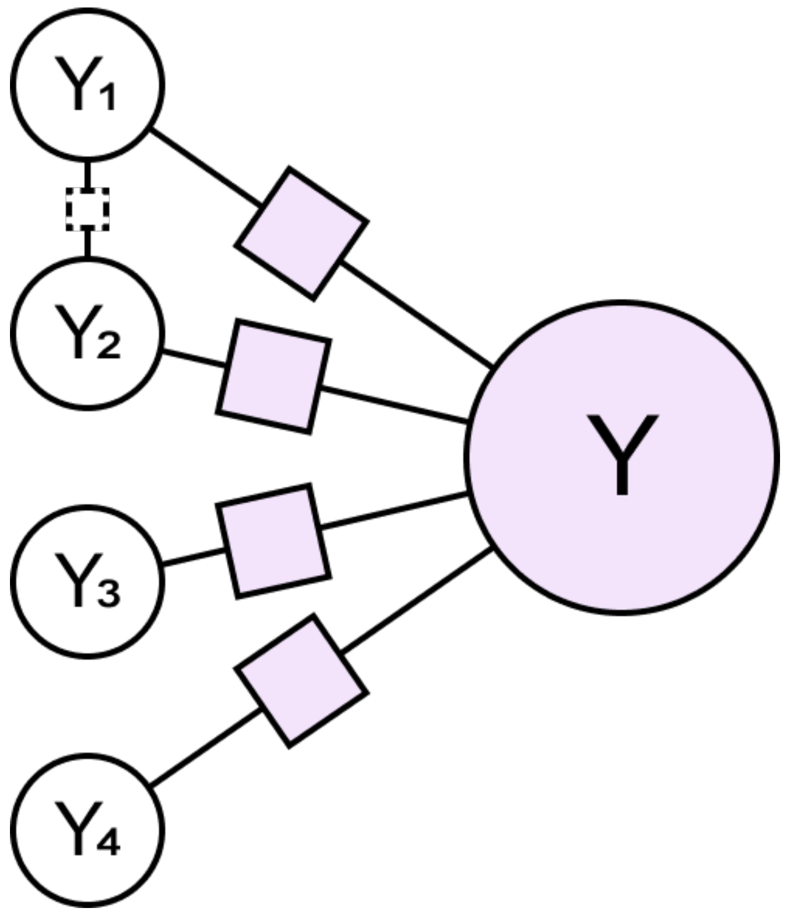
“If it says “credit” in the title…”



“If it matches a list of financial terms…”

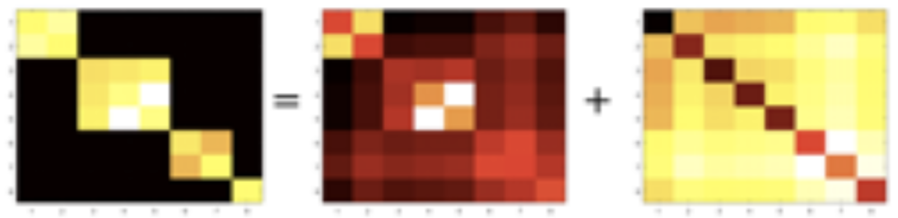


“If our legacy system thinks it’s a credit agreement…”




**[Intuition]**

Look at agreements & disagreements



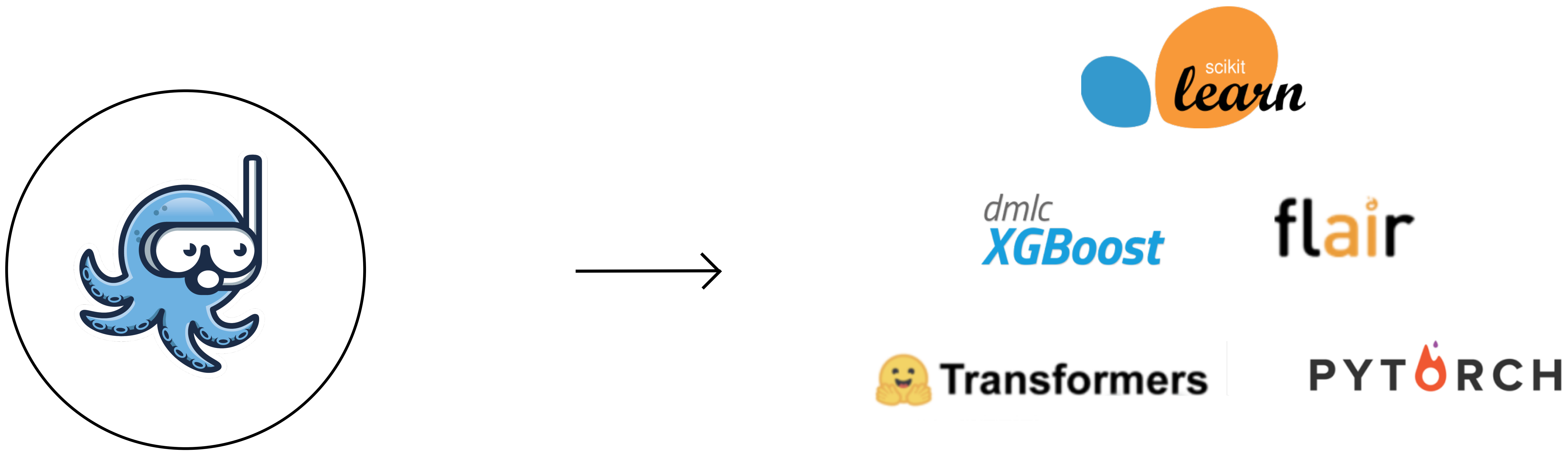
$$(\Sigma^{-1})_0 = \Sigma_0^{-1} + zz^T$$

Provably consistent matrix completion-style algorithm over inverse covariance

Snorkel **learns how to combine your noisy LFs** in an unsupervised way

[Ratner et. al. NeurIPS’16; Bach et. al. ICML’17; Ratner et. al. AAAI’19; Varma et. al. ICML’19l; Sala et. al. NeurIPS’19; Fu et. al. ICML’20]

# Train a discriminative model

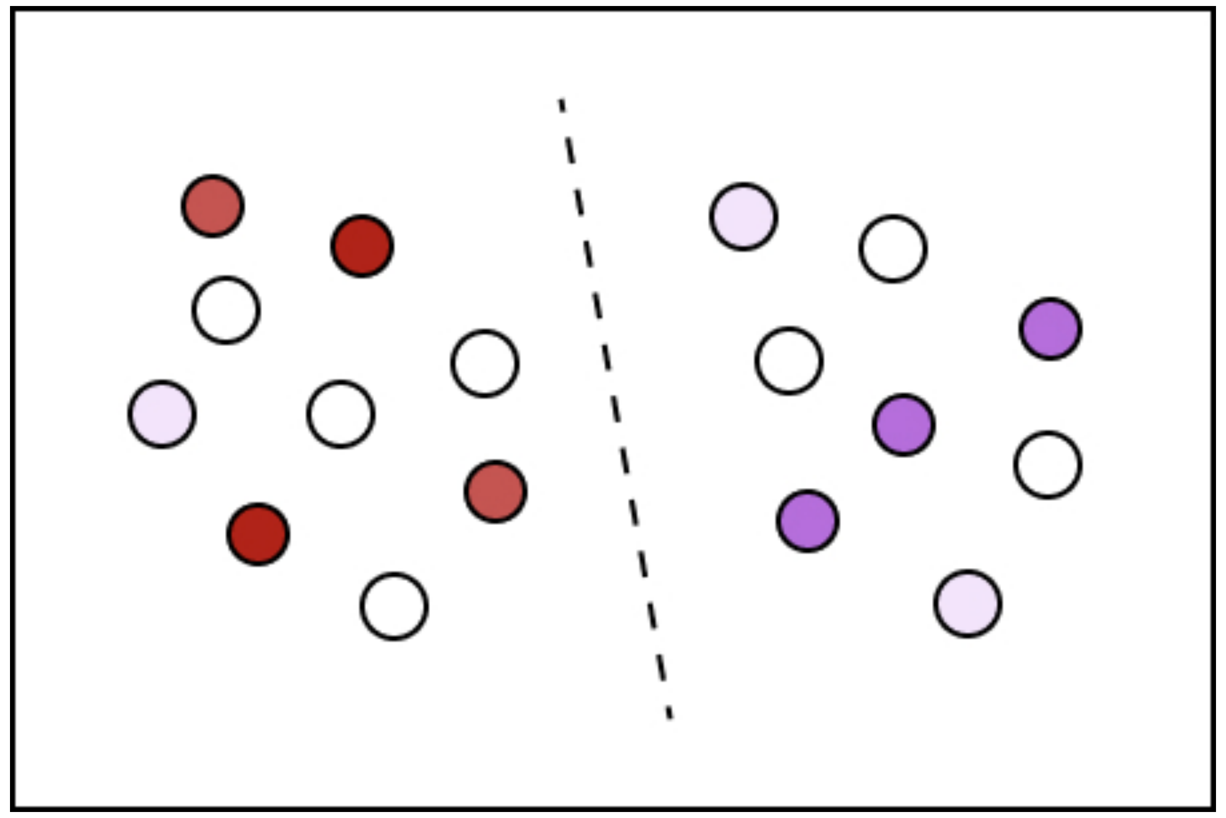



Use your programmatically generated training set to train a model

[Ratner et. al. NeurIPS'16; Bach et. al. ICML'17; Ratner et. al. AAAI'19; Varma et. al. ICML'19l; Sala et. al. NeurIPS'19; Fu et. al. ICML'20]

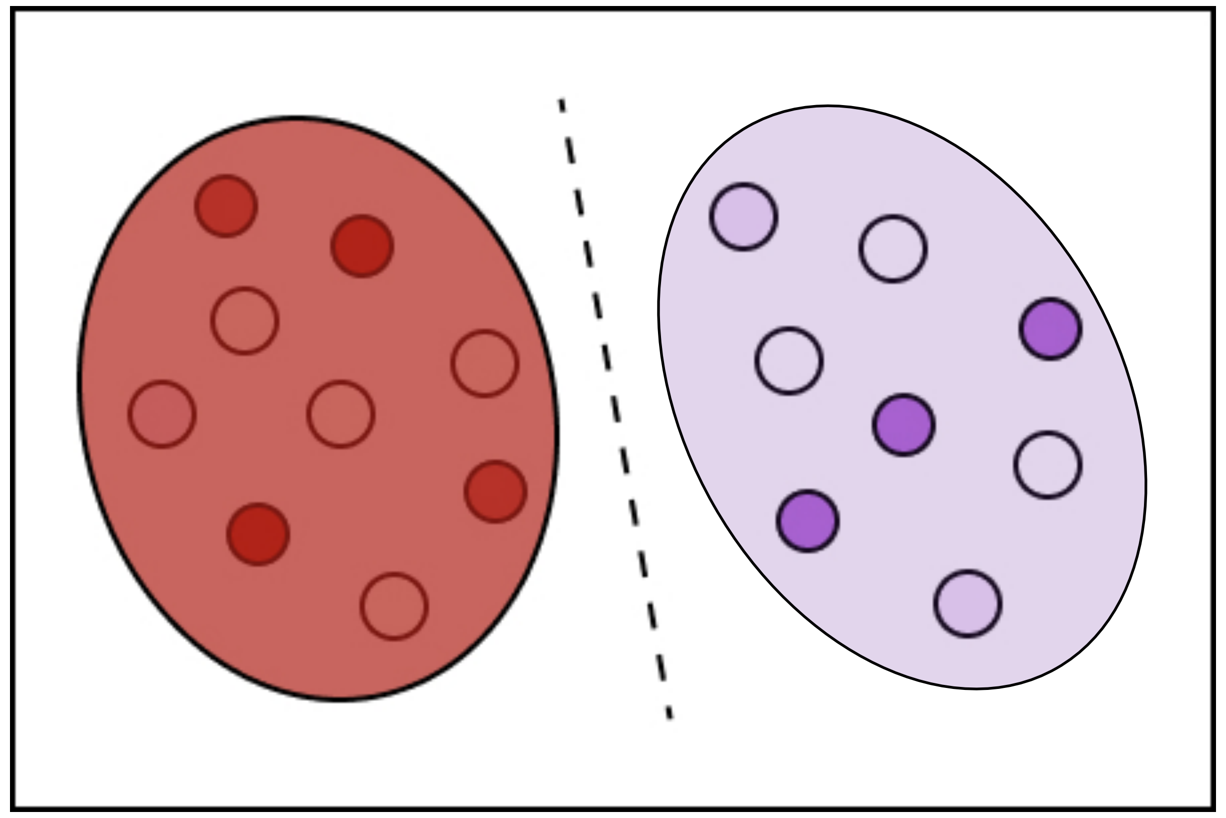
# Input: Rules, Output: Model

Predictions from LFs Alone



High Precision, Limited Coverage

Predictions from Model Trained on those LFs



Similar Precision, 100% Coverage

**Generalize** to new examples not covered by your LFs

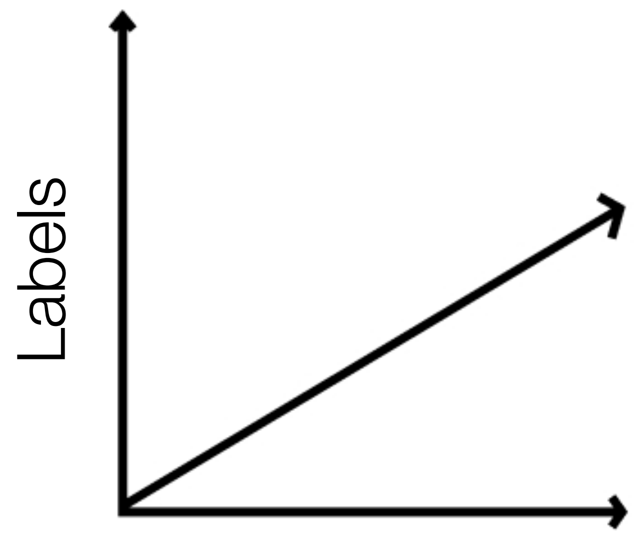
[Ratner et. al. NeurIPS'16; Bach et. al. ICML'17; Ratner et. al. AAAI'19; Varma et. al. ICML'19l; Sala et. al. NeurIPS'19; Fu et. al. ICML'20]

QUESTION 2:

# Does it work?

## Manual

### Slow

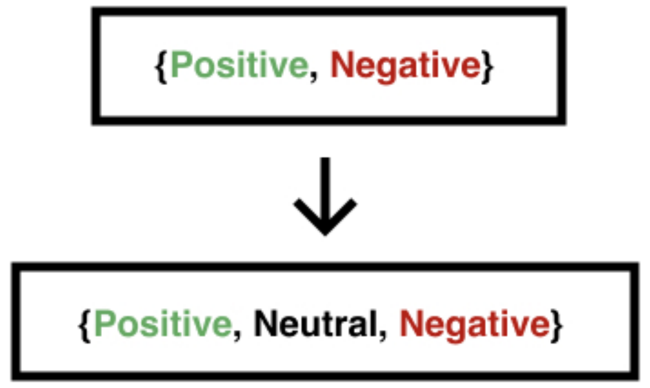



Time

### Expensive

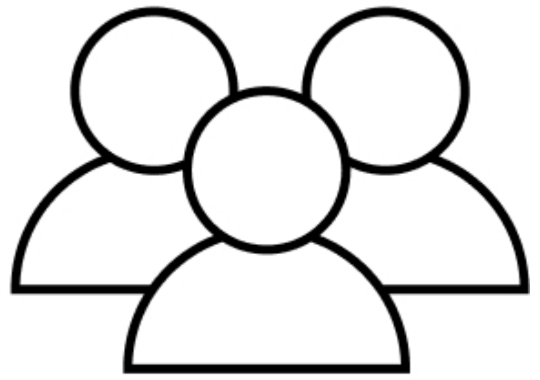
$10 - $100/hr

### Static




Relabeling = Starting Over

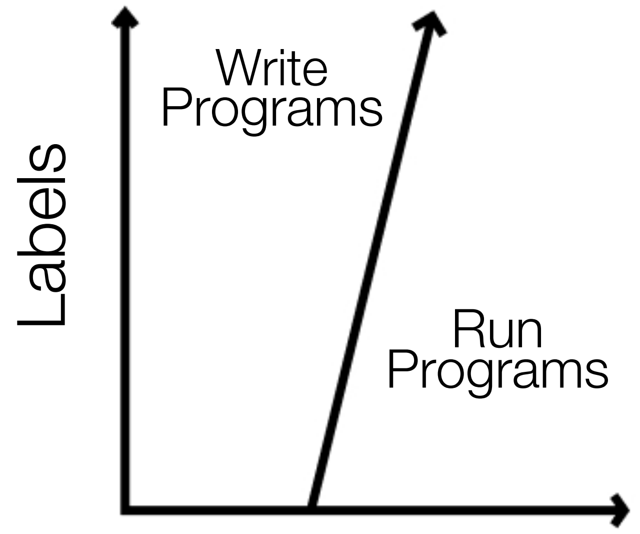
### Exposed



Shared with Crowdworkers

## Programmatic

### Fast

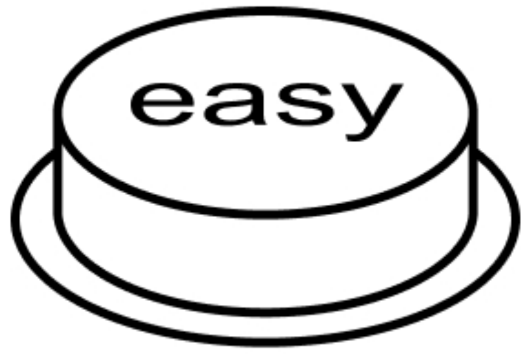



Time

### Cheap


$0.10/hr

### Dynamic




Push-button relabeling

### Private

Shared with Nobody

# Industry Adoption

snorkel

| | |
|---|---|
| Star | 4.1K |
| Fork | 687 |

Google · FDA · Infosys · facebook · LinkedIn · BASF We create chemistry

YouTube · NIH · Microsoft · Chegg

teradata. · accenture · Alibaba · intel · TOSHIBA · vmware

NIH National Institutes of Health · VA U.S. Department of Veterans Affairs · IBM · ANT FINANCIAL · DARPA

NEC · HITACHI · AMERICAN FAMILY INSURANCE · SAP · GORDON AND BETTY MOORE FOUNDATION

www.snorkel.ai www.snorkel.org

# Academic Leaderboards

## GLUE

| Rank | Name | Model | URL | Score |
|---|---|---|---|---|
| 1 | GLUE Human Baselines | GLUE Human Baselines | | 87.1 |
| 2 | Stanford Hazy Research | Snorkel MeTaL | | 83.2 |
| 3 | 王玮 | ALICE large (Alibaba DAMO NLP) | | 83.1 |
| 4 | Microsoft D365 AI & MSR AI | MT-DNNv2 (BigBird) | | 83.1 |
| 5 | Anonymous Anonymous | BERT + BAM | | 82.3 |
| 6 | Jason Phang | BERT on STILTs | | 82.0 |
| 7 | Jacob Devlin | BERT: 24-layers, 16-heads, 1024-hidden | | 80.5 |
| 8 | Neil Houlsby | BERT + Single-task Adapters | | 80.2 |
| 9 | Alec Radford | Singletask Pretrain Transformer | | 72.8 |
| 10 | GLUE Baselines | BiLSTM+ELMo+Attn | | 70.0 |

20 March 2019

https://gluebenchmark.com/

## SuperGLUE

| Rank | Name | Model | URL | Score |
|---|---|---|---|---|
| 1 | SuperGLUE Human Baselines | SuperGLUE Human Baselines | | 89.6 |
| 2 | Stanford Hazy Research | Snorkel Metal | | 74.5 |
| 3 | SuperGLUE Baselines | BERT++ | | 70.5 |
| | | BERT | | 68.0 |
| | | CBOW | | 48.6 |
| | | Most Frequent Class | | 46.9 |
| | | Outside Best | | - |

18 June 2019

https://super.gluebenchmark.com/

# Case Studies

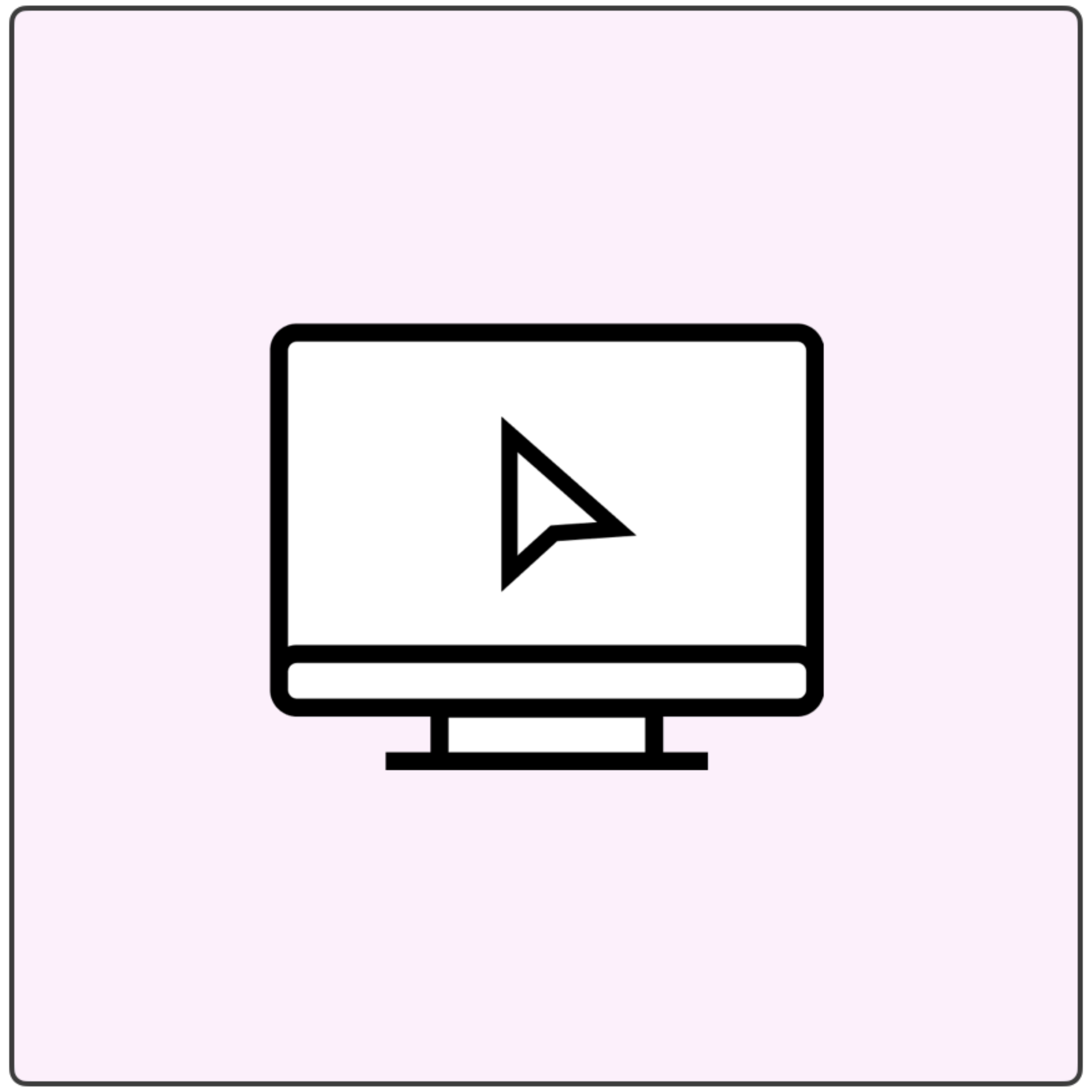
| Google | intel | Stanford MEDICINE | Top-3 US Bank |
|---|---|---|---|
| Content Classification | Social Media Monitoring | Medical Image Labeling | Contact Intelligence |
| 100K+ Hand labels replaced | 6 Crowdworker-months labels replaced | 8 Person-months of labeling replaced | 1 Month of labeling effort in <24 hours |
| 52% Improvement by repurposing resources | +18.5 Precision percentage points | 94% ROC AUC Performance | 99.1% Snorkel Flow Accuracy |
| 6M+ Labels in < 30 min. | +28.5 Coverage percentage points | 50K+ Images labeled in minutes | > 250K # Documents processed |

QUESTION 3:

# What did we learn? (aka The Four “I”s)

# Interfaces

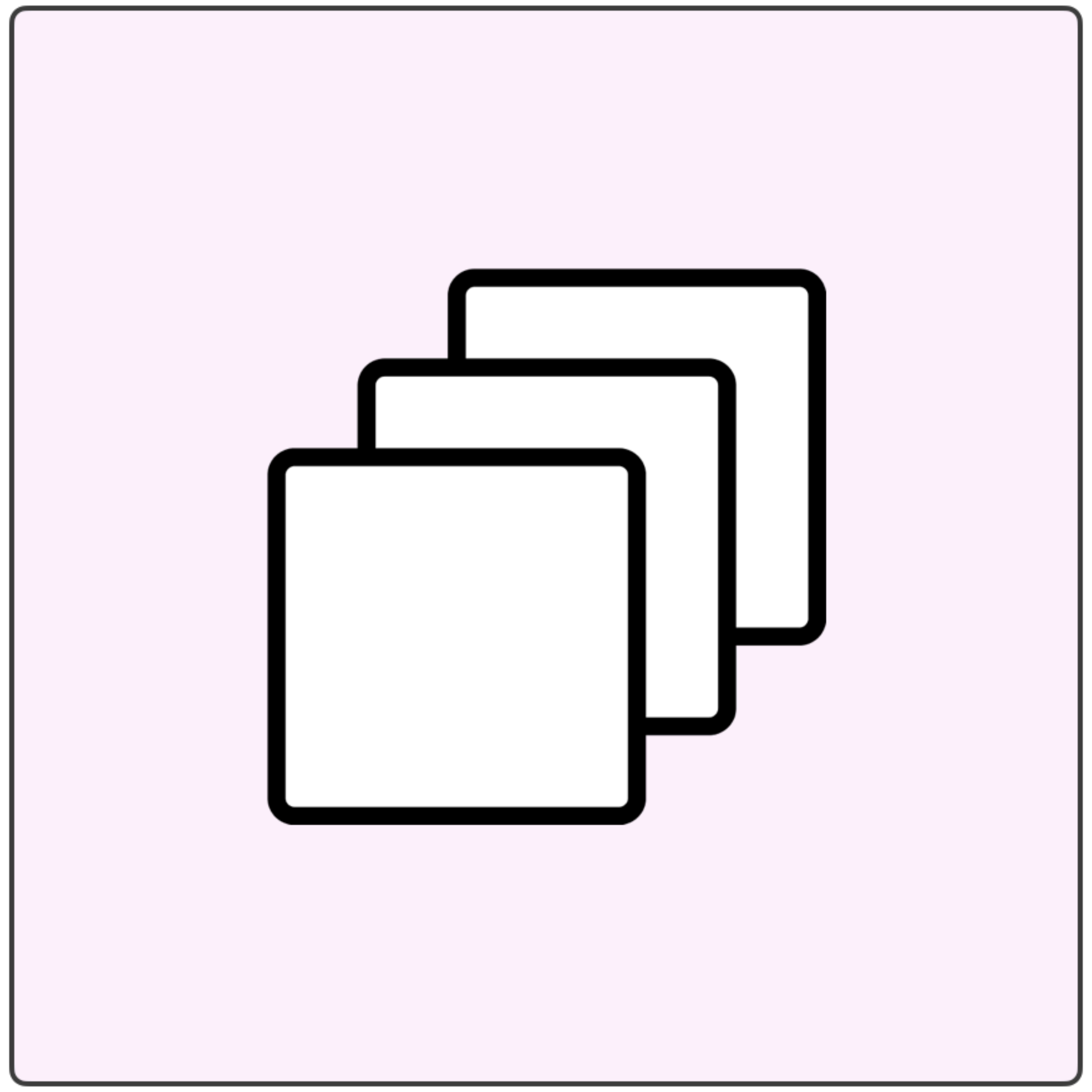
- Jupyter notebooks can't do *everything*
- Common LF types can be templatized
- Heavy 3rd party resources can be cached
- Standard analyses can be made push-button
- User experience can be improved with a GUI



A standalone Python package is ultimately limited in what it can do

# Infrastructure

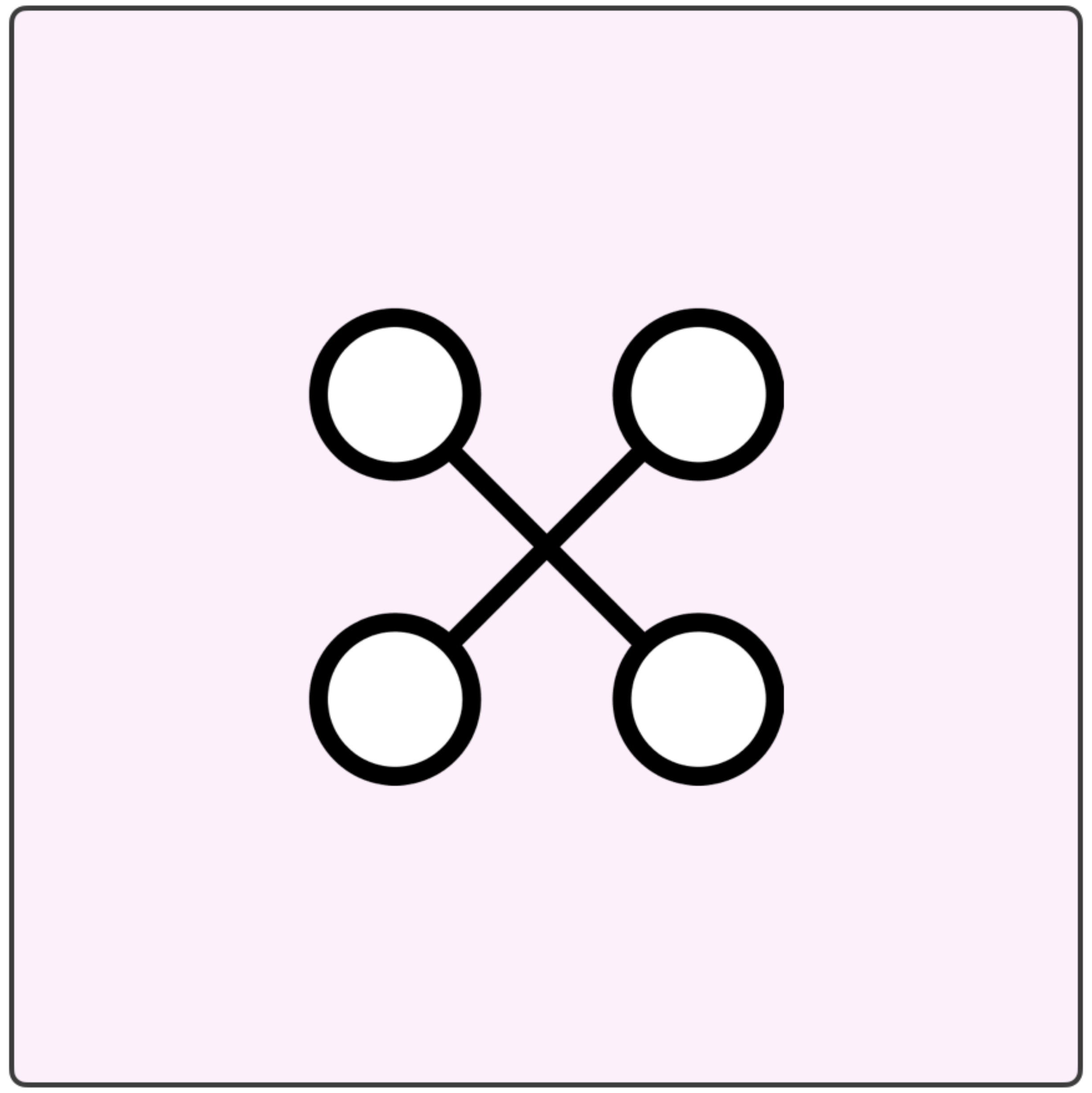
- Parallelization
- Logging
- User Profiles
- Integrations
- Dependencies
- Encryption
- Scalability
- Data Formats
- Versioning
- Etc.



Software that enterprises depend upon needs **enterprise-level support**

# Interactions

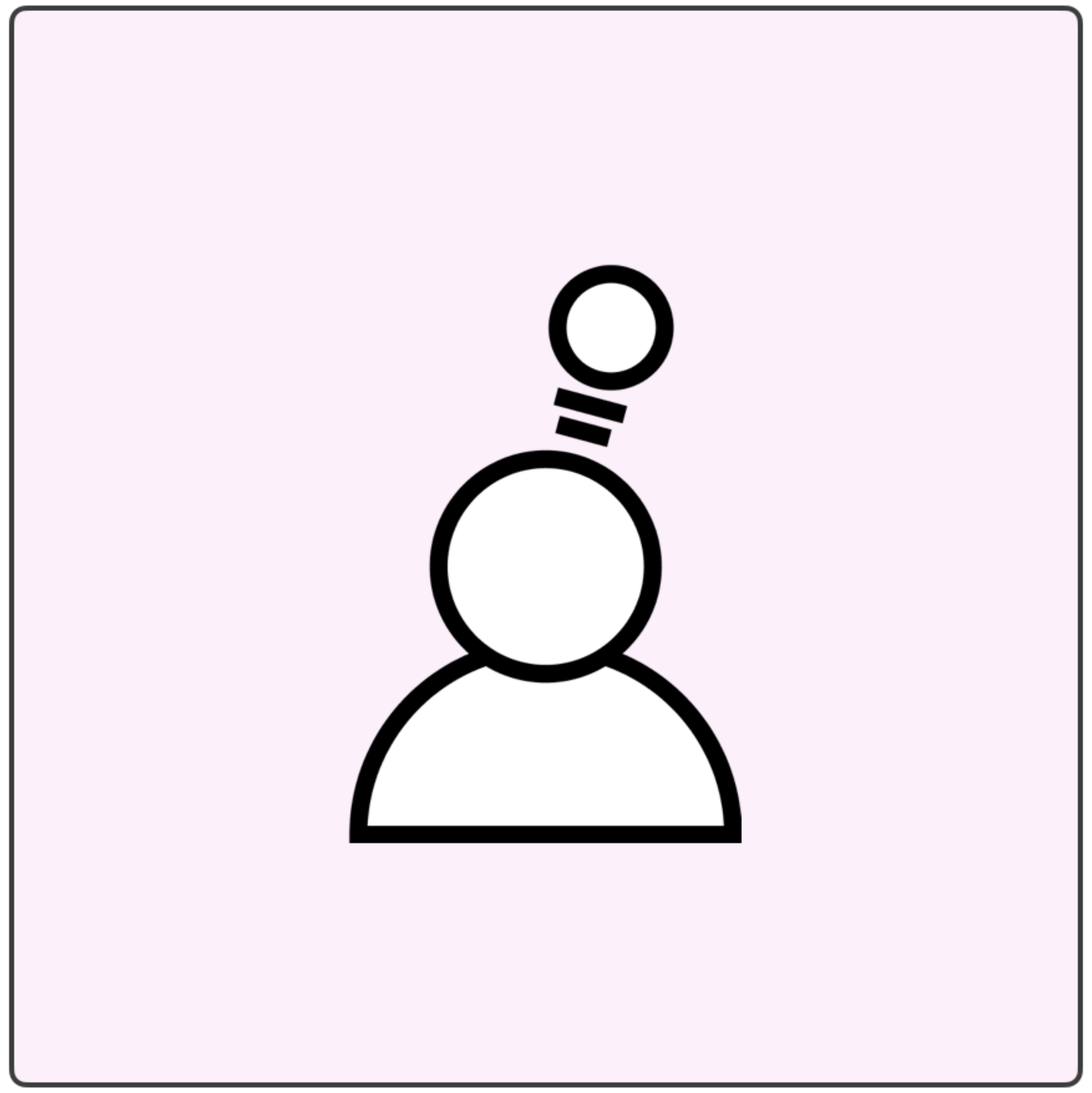
- SME: labels, tags, comments
- DS: program, debug, connect
- Developer: API, SDK, CLI
- Business: dashboards, metrics, samples



Different user profiles require different **views & interaction points**

# Intuition

- How many LFs is enough?
- What LF should I write next?
- Should I focus on precision or recall?
- How do I address bias?
- Will more data help me?
- Is my model or my supervision lacking?

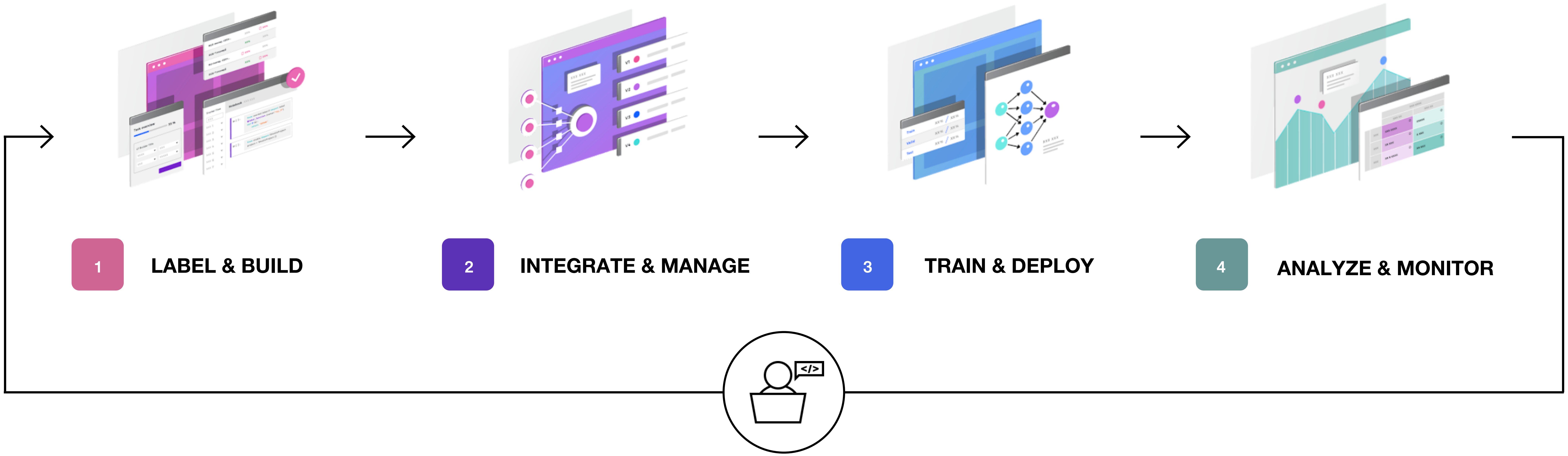


A new interface to ML comes with new **best practices, tips, and tricks**

# Snorkel Flow is a platform for building AI applications

Powered by Snorkel technology

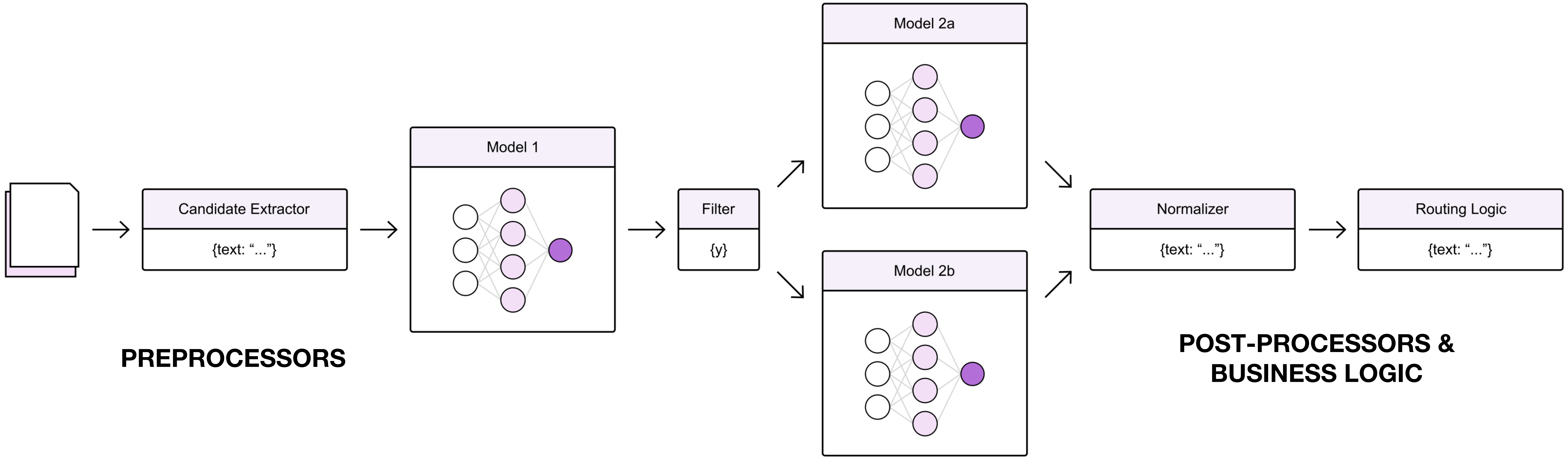
# Snorkel Flow: The Radical New Way to Build AI Applications




Snorkel Flow enables a faster, more practical, **adaptive ML development** process

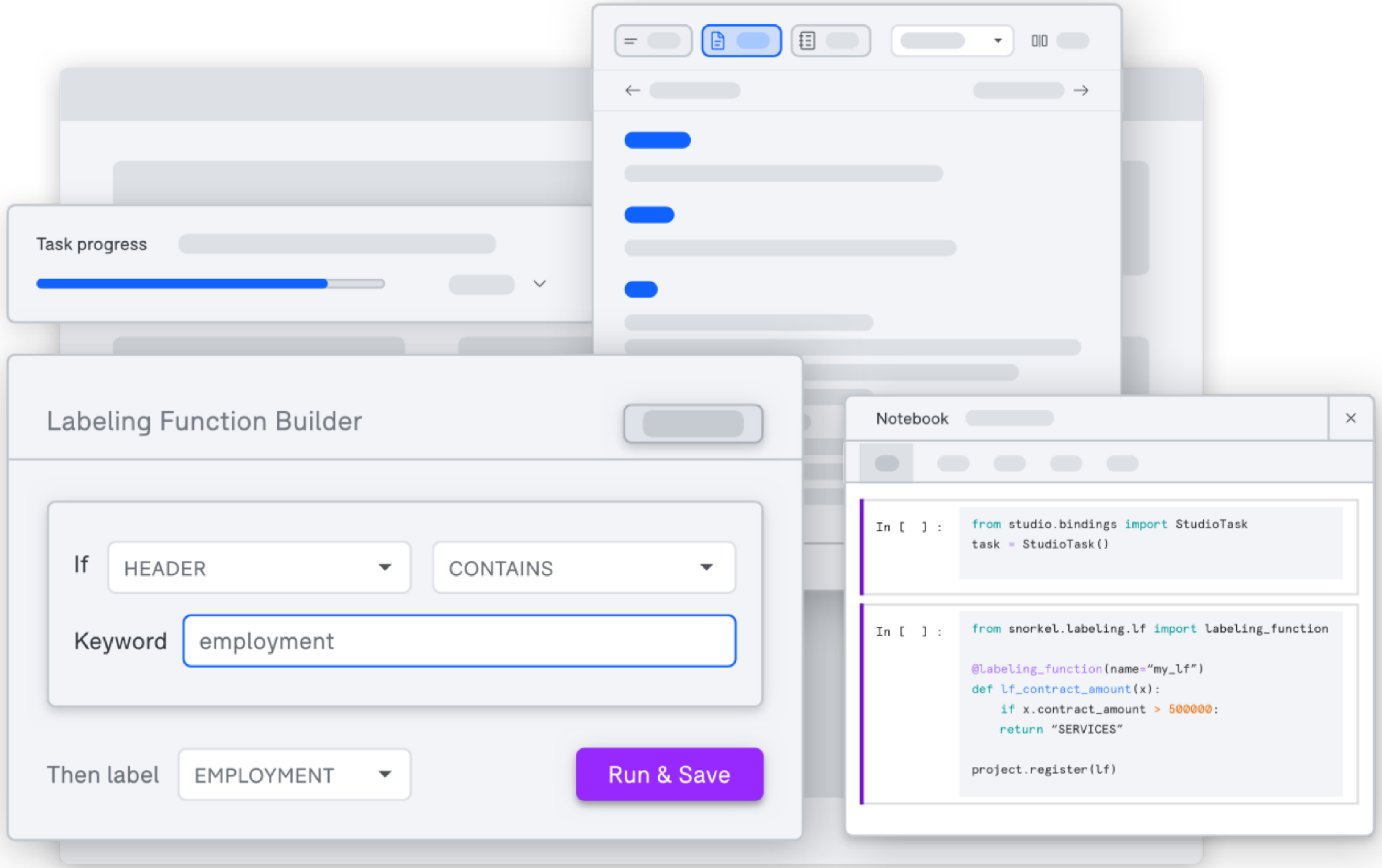
# 4 Guiding Principles behind Snorkel Flow

# Applications > Models



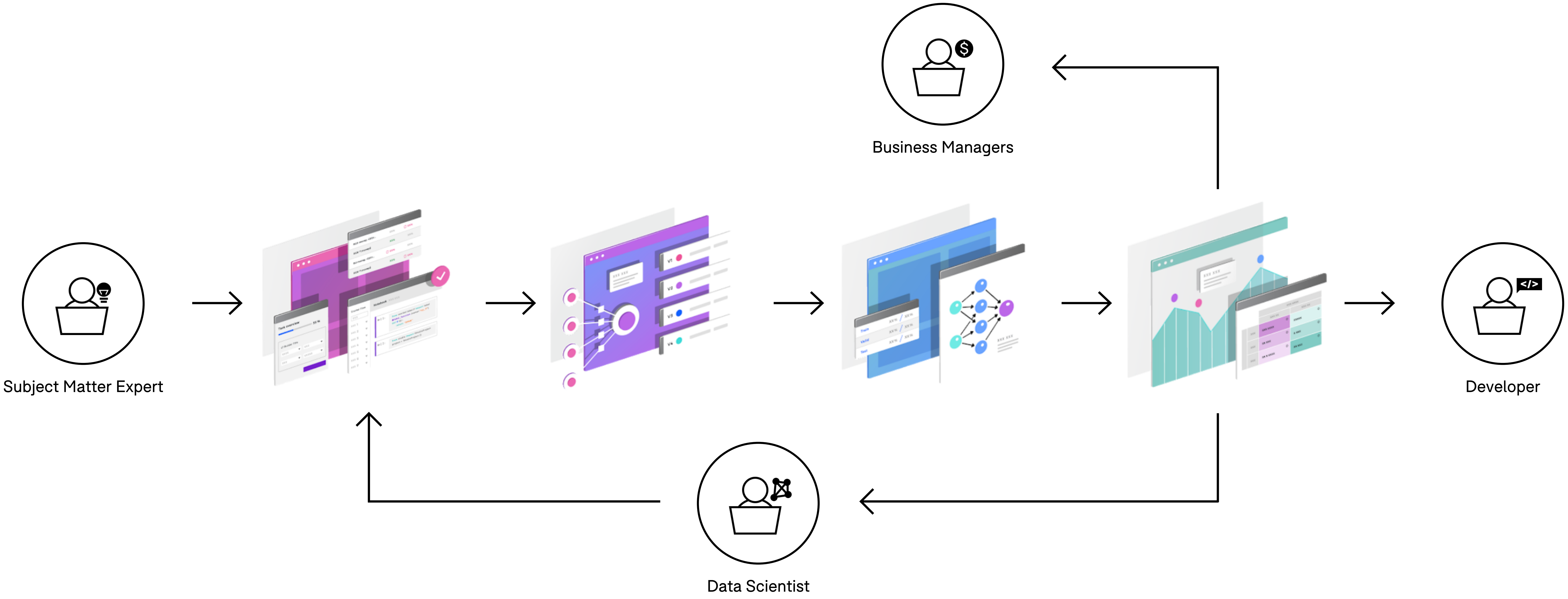

Real-world applications chain **preprocessors**, **models**, and **control logic**

# High & Low



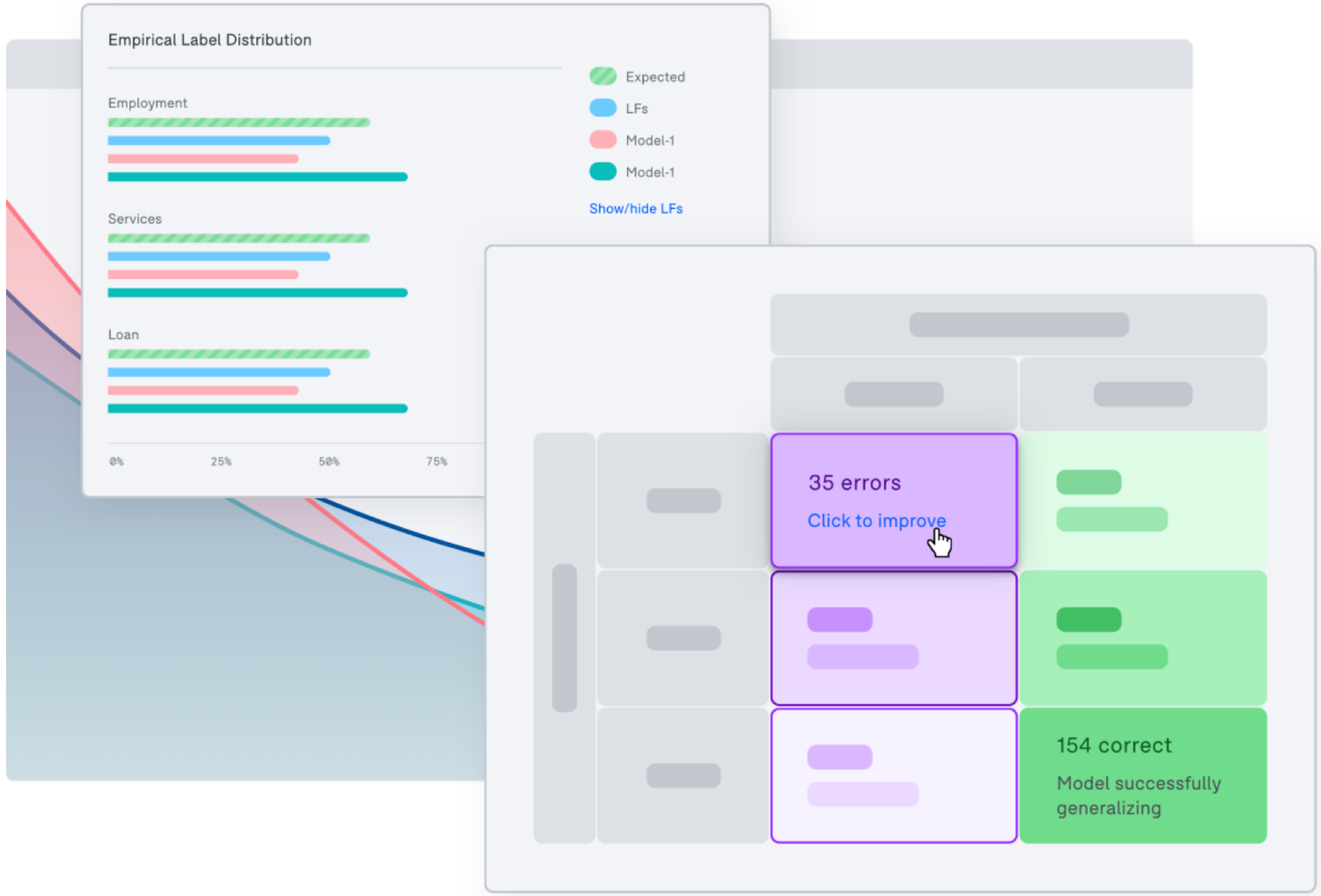

**High-level** where you want it, **low-level** where you need it

# Collaborative Hub




Support the **whole team** responsible for driving ML success in an org

# Iteration Wins




Drive up performance with **systematic, iterative, integrated** guidance

# Request a Demo

https://snorkel.ai/demo

Or email at info@snorkel.ai

# Talk to us on Twitter!

@SnorkelAI

# Thank You