Дизайн систем машинного обучения

3. Обучающие данные

План курса

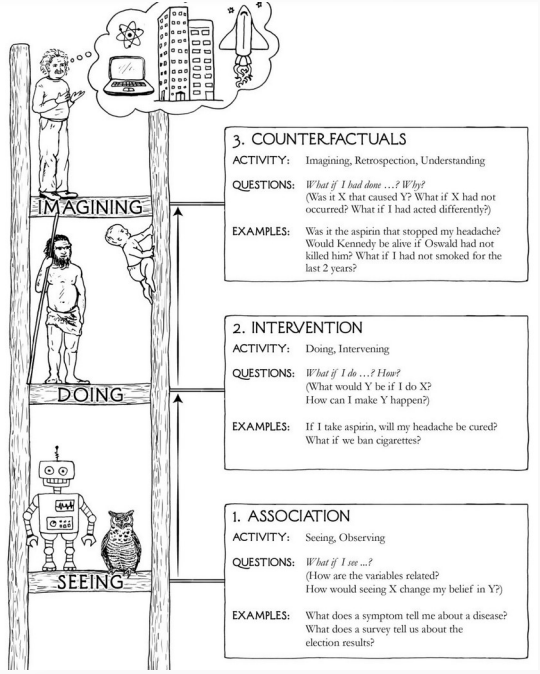
1) ~~Практическое применение машинного обучения~~
2) ~~Основы проектирования ML-систем~~
3) **Обучающие данные — Вы находитесь здесь**
4) Подготовка и отбор признаков
5) Выбор модели, разработка и обучение модели
6) Оценка качества модели
7) Развертывание систем
8) Диагностика ошибок и отказов ML-систем
9) Мониторинг и обучение на потоковых данных
10) Жизненный цикл модели
11) Отслеживание экспериментов и версионирование моделей
12) Сложные модели: временные ряды, модели над графами
13) Непредвзятость, безопасность, управление моделями
14) ML инфраструктура и платформы
15) Интеграция ML-систем в бизнес-процессы

Кто победит?

- Команда с большими мозгами
  - Продвинутые алгоритмы
  - Квалифицированные исследователи
  - Хорошие инженеры
- Команда с большими данными
  - Доступ к огромному датасету
  - Студенты старших курсов
  - Обычные разработчики

Лестница причинности

- Контрфактуалы
  - Альтернативная история
- Эксперимент
  - Достигаем результата
- Ассоциация — встречаются вместе
  - После — не значит вследствие
  - Correlation does not imply causation



The Book of Why: The New Science of Cause and Effect, Judea Pearl

Большие модели = Много данных

- Неудобно хранить webdataset
- Ошибки в разметке LabelErrors
- Дисбаланс классов
- Предвзятость
- Неоднозначная разметка

Разметка

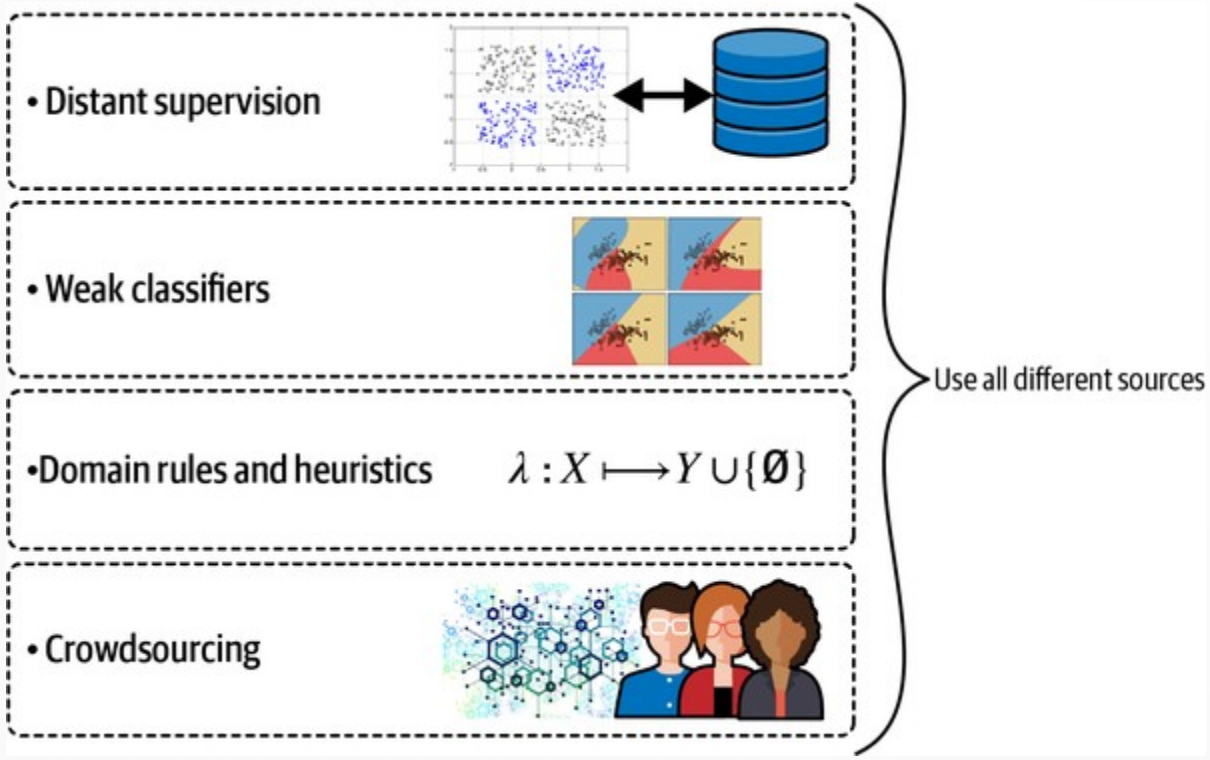
- Ручная разметка данных
- Программируемые датасеты snorkel
- Модели слабого контроля snorkel skweak
- Semi-supervision v7lab
- Transfer learning
- Active Learning Baal или modAL
- Выявление ошибок в данных cleanlab

Ручная разметка данных

- Толока
- LabelMe

- Amazon Mechanical Turk
- Scale
- LabelBox

Программируемые датасеты



Practical Weak Supervision

Snorkel

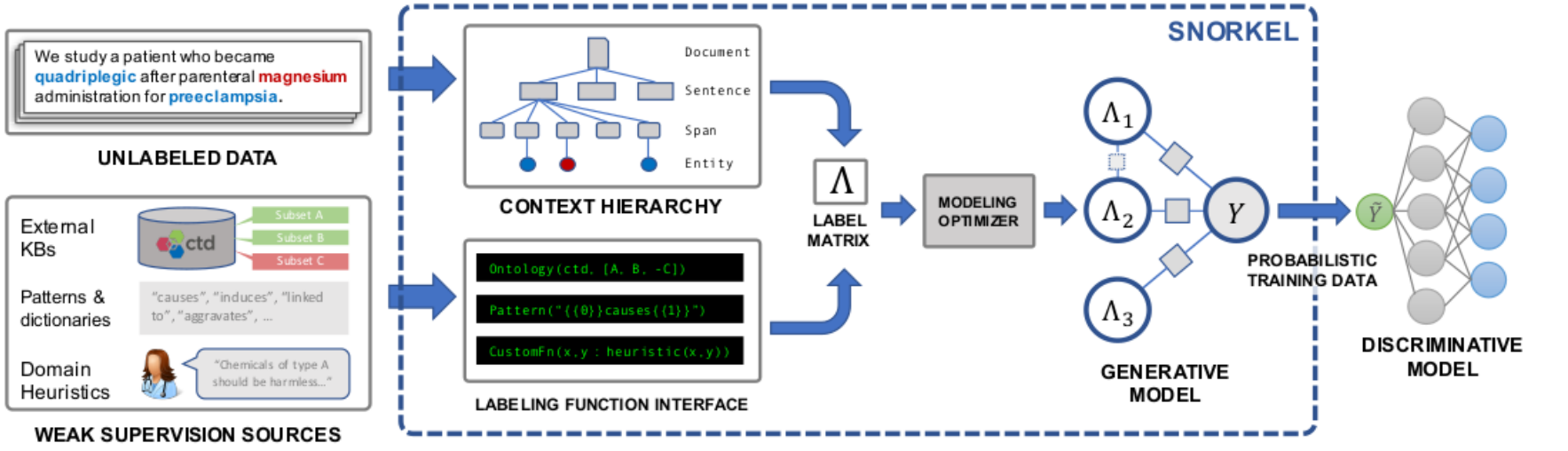


Figure 2: An overview of the Snorkel system. (1) SME users write *labeling functions (LFs)* that express weak supervision sources like distant supervision, patterns, and heuristics. (2) Snorkel applies the LFs over unlabeled data and learns a generative model to combine the LFs' outputs into probabilistic labels. (3) Snorkel uses these labels to train a discriminative classification model, such as a deep neural network.

https://arxiv.org/pdf/1711.10160.pdf

Программный слабый контроль

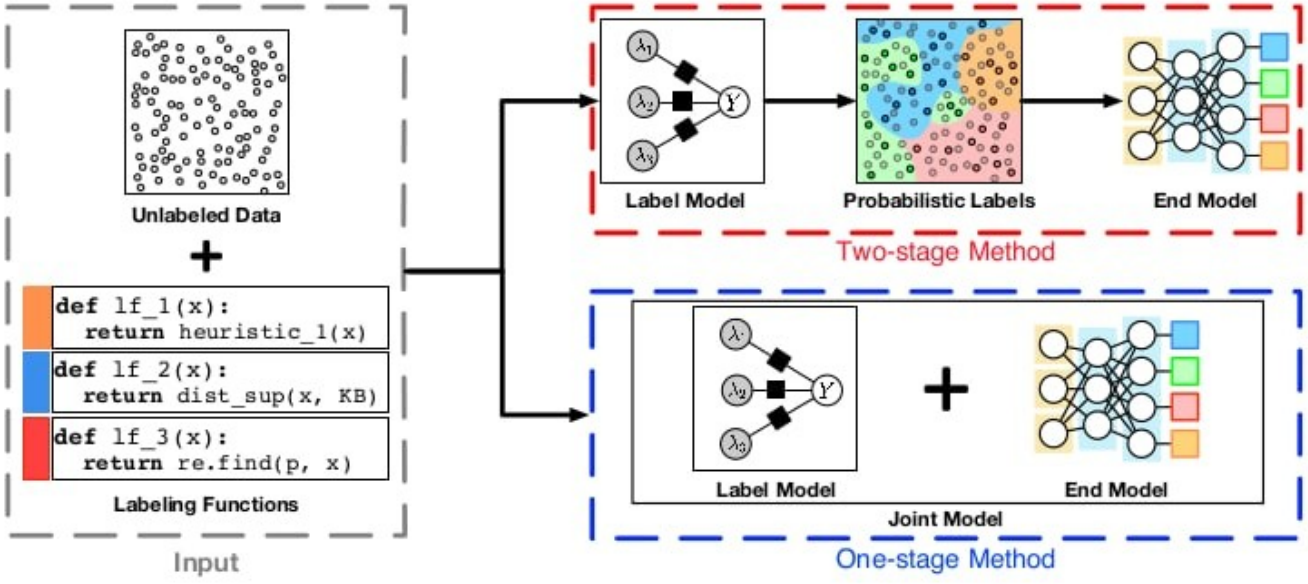


Figure 1: An overview of PWS pipeline [Zhang *et al.*, 2021].

A Survey on Programmatic Weak Supervision

Как отбираем данные для обучения

- Данные стоят дорого, часто нам нужно решить — какие взять
- Купить, собрать, разметить
- Стратегии отбора:
  - Берем то, что доступно
  - Эксперты решили, что важно
  - Снежный ком — собираем связанные данные
  - Квота — решаем, сколько каких примеров нам нужно

Сэмплирование

- Случайное
- Стратифицированная выборка
- Взвешенная выборка — приоритизируем классы
- Выборка по важности — оцениваем «нужность»
- Reservoir sampling
- С возвращением/без возвращения

Баланс классов

- Проблема не в дисбалансе классов
- Проблема в
  - малом количестве примеров для редких классов
  - неподходящем функционале качества.
- Например, у Дьяконова или Мельника
- Дисбаланс меньше чем 10:1 обычно не проблема

Что делать с (дис)балансом классов

- Стратификация выборок
- Дополнительные данные для редких классов
- Взвешивание классов (class_weight="auto" и т. д.)
- Объединение редких классов
- Вероятностные метрики
- Сэмплирование с возвращением во время обучения aka bootstrap
- Focal Loss например, в torchvision
- imbalanced-learn
- Переформулировка задачи — детекция аномалий

Дополнительные материалы

- Practice of Efficient Data Collection via Crowdsourcing at Large-Scale
- Bayesian active learning for production, a systematic study and a reusable library
- A General-Purpose Crowdsourcing Computational Quality Control Toolkit for Python
- Can I use this publicly available dataset to build commercial AI software?
  Most likely not
- A Survey on Programmatic Weak Supervision

Все будет в телеграм-канале