From Raw Data to Models

Albert Wendsjö

University of Gothenburg

2026-02-27

Motivation

Motivation

- There are a lot of resources available for how train various machine learning models using training data

Motivation

- There are a lot of resources available for how train various machine learning models using training data
- But where does the training data come from?

Motivation

- There are a lot of resources available for how train various machine learning models using training data
- But where does the training data come from?
- Not so obvious if you want complicated annotations

Agenda, Limitations and Assumptions

- Agenda

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data?

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data?

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model?

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021 (What is ChatGPT?)

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021 (What is ChatGPT?)
  - I will keep it practical

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021 (What is ChatGPT?)
  - I will keep it practical (What is the L1 regularizer?)

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021 (What is ChatGPT?)
  - I will keep it practical (What is the L1 regularizer?)
- Assumptions

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021 (What is ChatGPT?)
  - I will keep it practical (What is the L1 regularizer?)
- Assumptions
  - You want quantitative measures, & you want an automated approach for it

Agenda, Limitations and Assumptions

- Agenda
  - Why do we need annotated data? (Presentation)
  - How can we annotate data? (I share my screen and show)
  - How can we use annotated data to train a model? (We walk through a Colab notebook)
- Limitations
  - A part of me still live in 2021 (What is ChatGPT?)
  - I will keep it practical (What is the L1 regularizer?)
- Assumptions
  - You want quantitative measures, & you want an automated approach for it
  - You know what a regression is

What's the goal?

What's the goal?

- The goal of machine learning is prediction

What's the goal?

- The goal of machine learning is prediction
- Given some data (X), what is the likely value of Y?

What's the goal?

- The goal of machine learning is prediction
- Given some data (X), what is the likely value of Y?
- Two examples:

What's the goal?

- The goal of machine learning is prediction
- Given some data (X), what is the likely value of Y?
- Two examples:
  - Is this text about sports?
  - Is there a football on this image?

What's the goal?

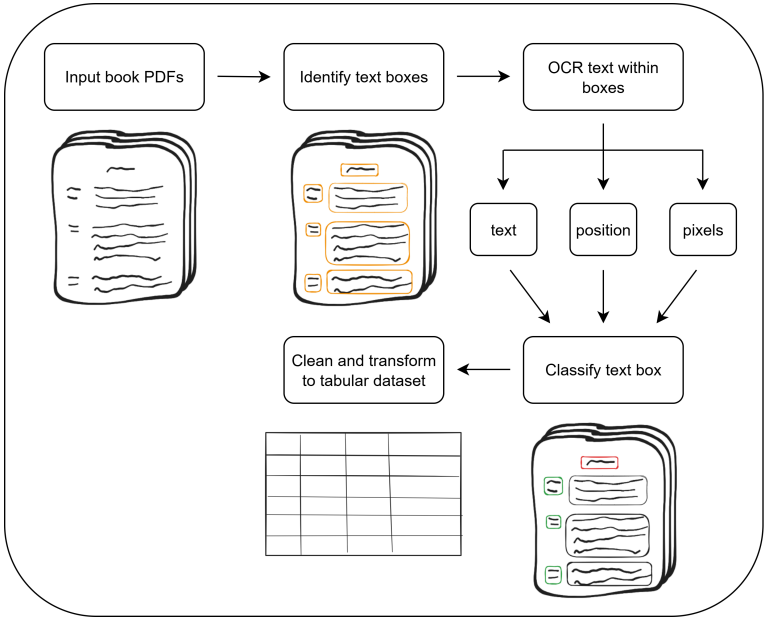
- The goal of machine learning is prediction
- Given some data (X), what is the likely value of Y?
- Two examples:
  - Is this text about sports?
  - Is there a football on this image?
- In both cases the model learns the pattern between words/pixels (X) and sports/football (Y)?

What's the goal?

- The goal of machine learning is prediction
- Given some data (X), what is the likely value of Y?
- Two examples:
  - Is this text about sports?
  - Is there a football on this image?
- In both cases the model learns the pattern between words/pixels (X) and sports/football (Y)?
- The model will learn the best pattern given the data

Why?

- This allows us to transform unstructured data into structured (tabular) data at scale
- Why is this good?
- We get more statistical power, more subgroup analysis, can fit complicated statistical models...



Why do we need annotated data?

- We need annotations for two things

Why do we need annotated data?

- We need annotations for two things
  - Training
  - Validation

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
  - Validation

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
    - We need examples to show the model what we want
  - Validation

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
    - We need examples to show the model what we want
  - Validation
    - How accurate is the model?

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
    - We need examples to show the model what we want
  - Validation
    - How accurate is the model?
    - How can we adjust for measurement error?

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
    - We need examples to show the model what we want
  - Validation
    - How accurate is the model?
    - How can we adjust for measurement error?
- Recent trends suggest we need less data for training ($n \leq 500$), but we still need data for validation

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
    - We need examples to show the model what we want
  - Validation
    - How accurate is the model?
    - How can we adjust for measurement error?
- Recent trends suggest we need less data for training ($n \leq 500$), but we still need data for validation
- Model accuracy is a population parameter; we need sufficient sample size to estimate it!

Why do we need annotated data?

- We need annotations for two things
  - Training
    - Political scientists bring new labels and new contexts, that requires new data (X, Y) for the model to adopt
    - We need examples to show the model what we want
  - Validation
    - How accurate is the model?
    - How can we adjust for measurement error?
- Recent trends suggest we need less data for training ($n \leq 500$), but we still need data for validation
- Model accuracy is a population parameter; we need sufficient sample size to estimate it!
- Measurement error reduces statistical power, requiring annotated data to adjust for it

Running Example

- How can we classify texts? Words?
- How can we classify images? Parts of images?

Running Example

- The Snap Political Ads library contain 72146 political ads (2018-2025) from 54 countries
- We will limit us to the United States and images ($n = 9207$)
- (I provide full script for downloading, structuring and OCR:ing the ads.)

The tasks

- Texts
  - Classify topic (about voting or not)
  - Extract candidates and detect stance toward them
- Images
  - Extract faces
  - Classify faces
    - Gender (male, female)
    - Age (young, middle-age, old)
    - Ethnicity (white, non-white)

What can we do with this?

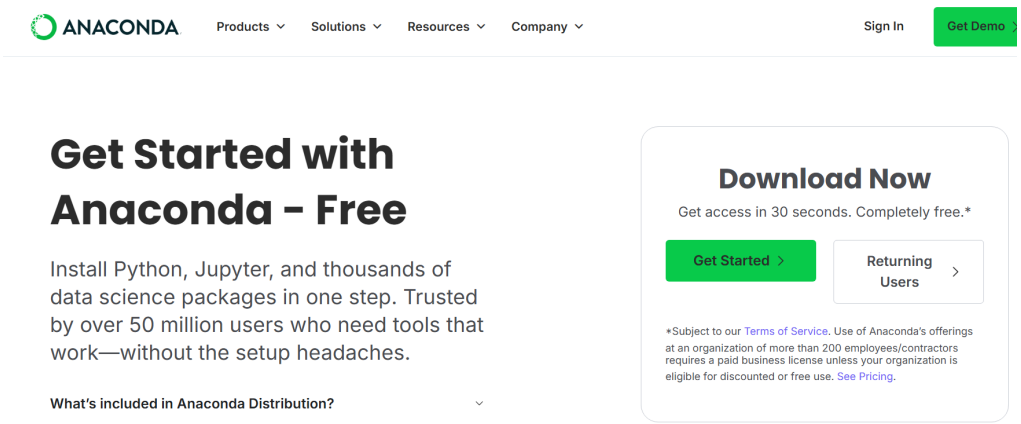
- How much money is allocated to positive and negative campaigns?
- When do political ads target specific demographics? (Visual group appeals)
- What type of ads generates the most interactions?
- For which groups do they encourage voting registrations?

How we will do it

- For annotation we will use Label Studio. It is an open source software that can be applied to many types of annotation tasks
- For training models, we will use Google Colab. I have prepared a notebook with scripts that show how you can import and apply annotated data to train and use various models

How to install Label-Studio

Step 1. Download Anaconda (a Python distributor)



How to install Label-Studio

Step 2. Install Label-Studio the Anaconda prompt (terminal)
Step 3. Start Label-Studio in the Anaconda prompt (terminal)