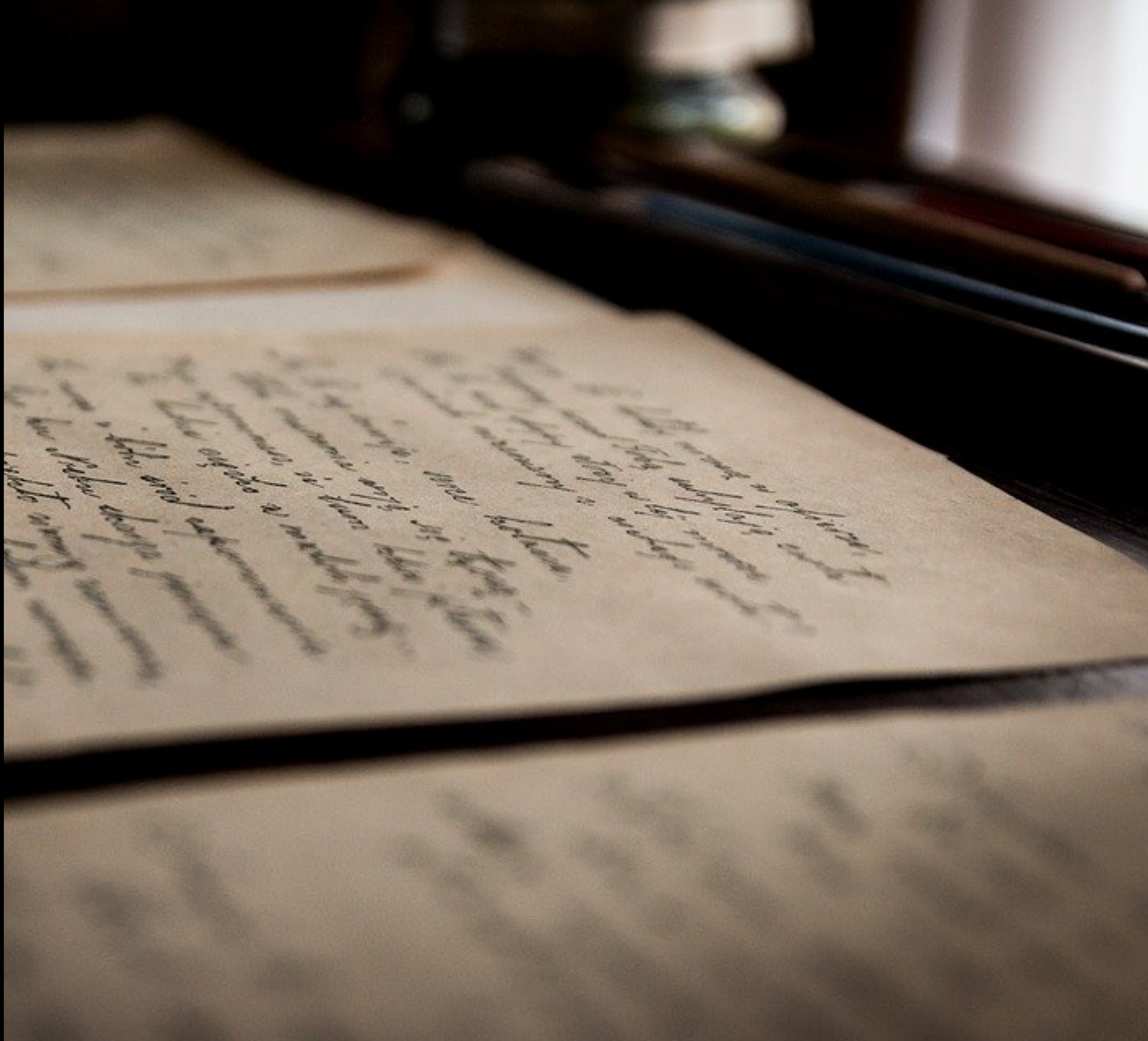
Cursive & Recursive:

Generating Transcriptions of Archival Documents Using Machine Learning

Week 3

Buchanan Fellowship, Spring 2020

Vanderbilt University Library

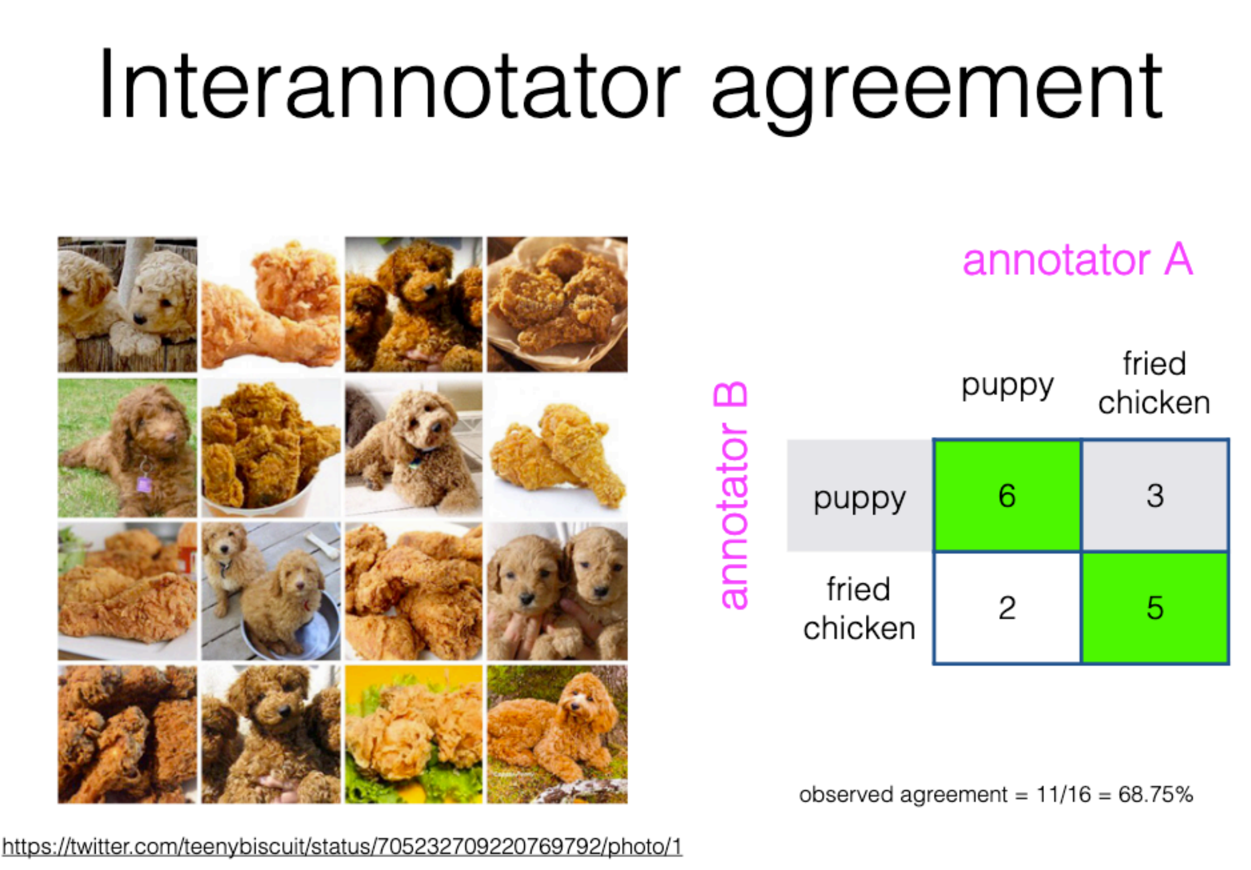


Making training sets

- Supervised machine learning needs a training set
- Training set is human-annotated data
- Can be too time-consuming to do by yourself, but

- Multiple people not always in agreement
- Quality control adjudication (can be just as time-consuming)
- Inter-annotator agreement

Inter-annotator agreement



Interannotator agreement

		annotator A	
		puppy	fried chicken
annotator B	puppy	6	3
	fried chicken	2	5

observed agreement = 11/16 = 68.75%

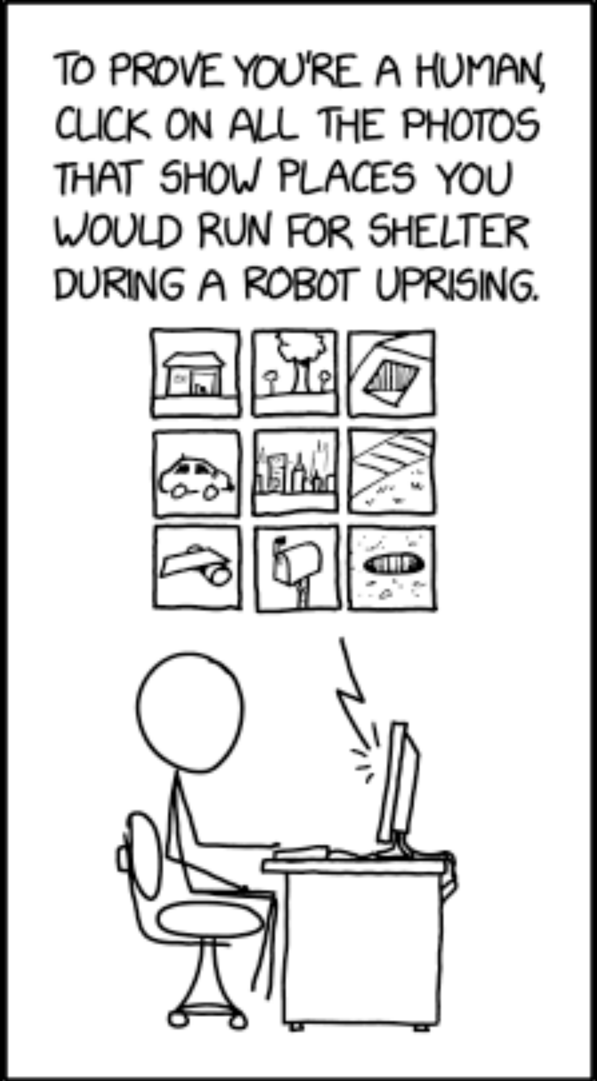
https://twitter.com/teenybiscuit/status/705232709220769792/photo/1

http://people.ischool.berkeley.edu/~dbamman/nlpF17/slides/5_truth_ethics.pdf

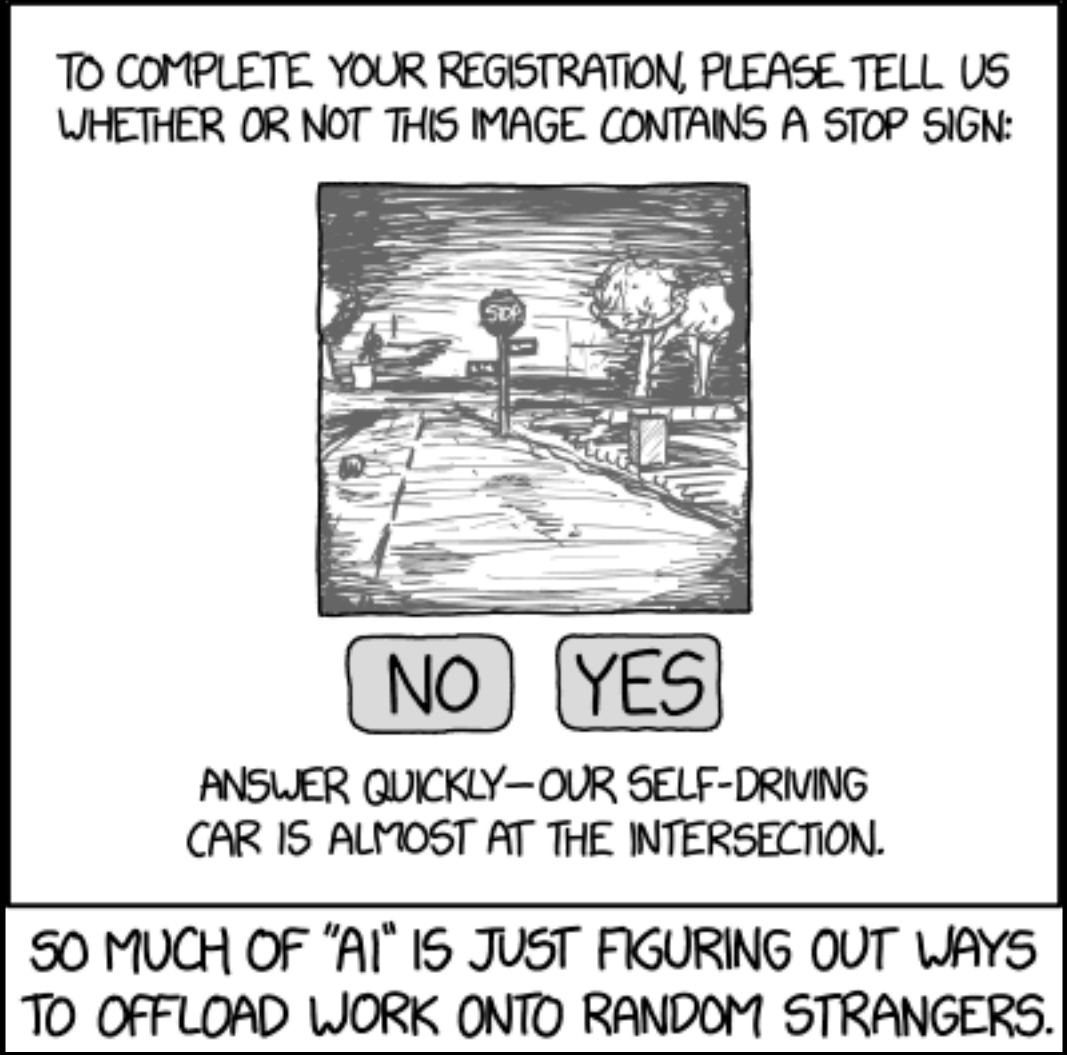
Making training sets through crowdsourcing

- Human Intelligence Tasks (HITs):
  - Amazon’s Mechanical Turk
  - Figure Eight (formerly Crowd Flower)

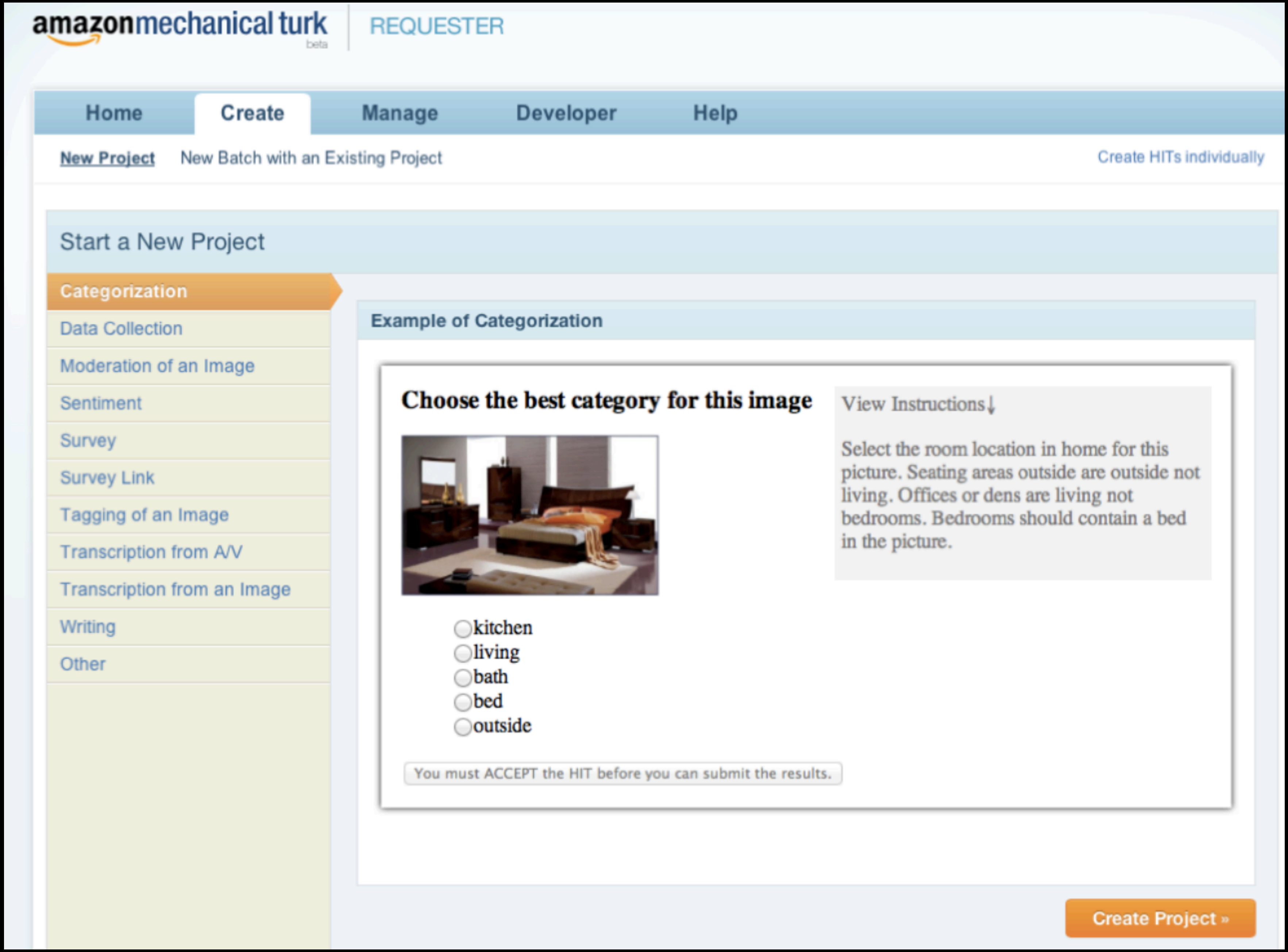
- CAPTCHA: Computer Automated Public Turing test to tell Computers and Humans Apart
- reCAPTCHA: “Stop spam. Read books.”


TO PROVE YOU'RE A HUMAN, CLICK ON ALL THE PHOTOS THAT SHOW PLACES YOU WOULD RUN FOR SHELTER DURING A ROBOT UPRISING.

https://xkcd.com/2228/


TO COMPLETE YOUR REGISTRATION, PLEASE TELL US WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:
STOP
NO
YES
ANSWER QUICKLY—OUR SELF-DRIVING CAR IS ALMOST AT THE INTERSECTION.
SO MUCH OF "AI" IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

https://xkcd.com/1897/



Start a New Project

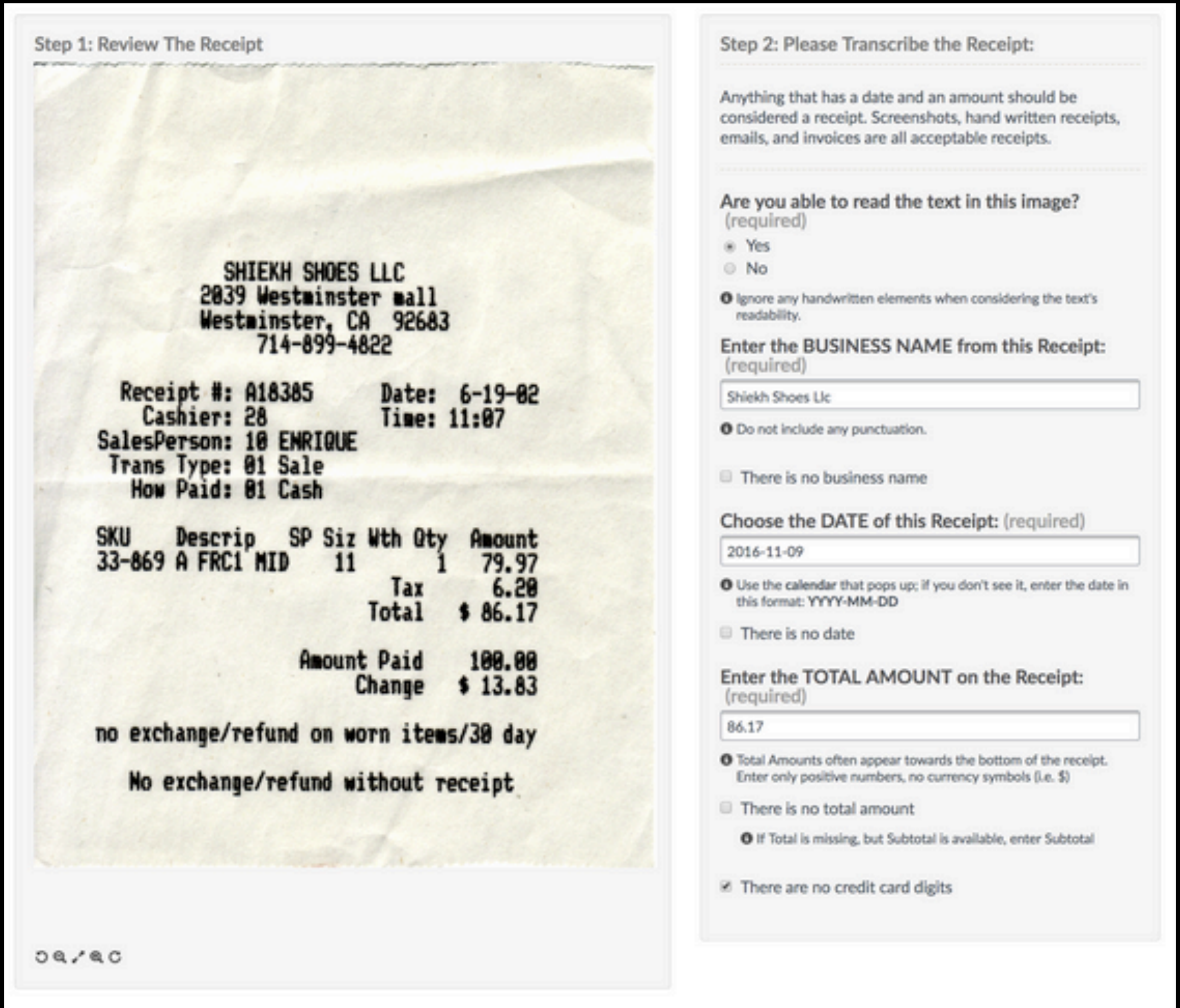
Example of Categorization

Choose the best category for this image

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

○kitchen
○living
○bath
○bed
○outside

You must ACCEPT the HIT before you can submit the results.



Step 1: Review The Receipt

```
        SHIEKH SHOES LLC
      2039 Westminster mall
      Westminster, CA  92683
          714-899-4822

  Receipt #: A18385      Date:  6-19-02
    Cashier: 28          Time: 11:07
SalesPerson: 10 ENRIQUE
 Trans Type: 01 Sale
   How Paid: 01 Cash

SKU    Descrip   SP Siz Wth Qty  Amount
33-869 A FRC1 MID   11        1   79.97
                         Tax       6.20
                       Total   $ 86.17

                 Amount Paid     100.00
                      Change   $ 13.83

no exchange/refund on worn items/30 day

   No exchange/refund without receipt
```

Step 2: Please Transcribe the Receipt:

Anything that has a date and an amount should be considered a receipt. Screenshots, hand written receipts, emails, and invoices are all acceptable receipts.

Are you able to read the text in this image? (required)

● Yes
○ No

Ignore any handwritten elements when considering the text's readability.

Enter the BUSINESS NAME from this Receipt: (required)

Shiekh Shoes Llc

Do not include any punctuation.

□ There is no business name

Choose the DATE of this Receipt: (required)

2016-11-09

Use the **calendar** that pops up; if you don't see it, enter the date in this format: **YYYY-MM-DD**

□ There is no date

Enter the TOTAL AMOUNT on the Receipt: (required)

86.17

Total Amounts often appear towards the bottom of the receipt. Enter only positive numbers, no currency symbols (i.e. $)

□ There is no total amount

If Total is missing, but Subtotal is available, enter Subtotal

☑ There are no credit card digits

Criticisms

CAPTCHA

- Unpaid labor
- Barrier to internet use
- Privacy concerns

Mechanical Turk, et al.

- Low-wage labor
- Language differences
- Cultural differences
- Content moderators at risk for PTSD

Common OCR packages

- ABBYY FineReader
    - Proprietary, not free
    - Best on market
    - Available on Computer 1 in DHC (Windows partition)
- Adobe Acrobat Pro
    - Proprietary, not free
    - Part of Adobe Creative Cloud
    - Available in DHC, Peabody Learning Commons
- Tesseract
    - Open source (free)
    - Originally HP, then Google supported
    - Use on command line
    - Can add to Python script

How to choose an OCR package

- Support for document language and script
- Support for handwriting
- Support for tables and layout
- Support for training or machine learning

- Clean (corrected) vs. dirty (uncorrected) OCR
    - OCR program with 99% accuracy will have ~10 errors / page

Other OCR training engines

- OCRopus
- Kraken
- Transkribus

[more on these later]

Add to cloud service pipeline

Tesseract core underneath

Can be trained on new data

- AWS Textract
- Google Cloud Vision
- Azure Computer Vision / OCR API

Demo Tesseract on Mac - Install

1. Open Terminal (command line)
2. Install Homebrew (Mac package manager)
   ```
   brew.sh
   ```
3. Install xpdf package (OS pdf viewer)
   ```
   brew install xpdf
   ```
4. Install ImageMagick (OS software to read images)
   ```
   brew install imagemagick
   ```
5. Install Tesseract (for all languages)
   ```
   brew install tesseract-lang
   ```

Demo Tesseract on Mac - Convert

Convert pdf with embedded text:

```
pdftotext inputfile.pdf outputfile.txt
```

Convert pdf without embedded text:

1. Convert pdf to tiff

   ```
   convert inputfile.pdf outputfile.tiff
   ```

2. Convert tiff to txt with Tesseract

   ```
   tesseract inputfile.tiff outputfile.txt
   ```

Download some images

- Download from website using right-click
- Or use wget command

```
wget [image file url]
```

```
wget –O example1.jpg
http://www.earlyprintedbooks.com/wp-
content/uploads/10397998-maximum-768x1198.jpg
```

- Be respectful of copyright and usage rights

Demo ABBYY FineReader