

# Labeling

# The bad news first

- Labeling is hard
  - Facebook
  - Points-Of-Interests (Foursquare, etc.)
- Labeling is going to get more difficult
  - Enterprise
  - Personalization
  - Healthcare
  - New data sets
- Some context
  - We assume supervised or semi-supervised learning
  - Large scale
  - Continuous

# What is a label?



• **M. Choi Young Deuk .** <info@undptours.com>  
To: oralonso@yahoo.com



Wed, Mar 31 at 8:16 AM

--

Hello Omar Alonso,

I am a banker working with CIMB bank Cambodia. I contacted you for a reason, one of my late customer have the same family name as yours. He died 6 years ago and left 10.7 million United States dollars in his account. Since then no relative have come to claim his money .. I think we can work things out.

Best regards,

M. Choi Young Deuk .



• **Chase**  
From: contact@melendezlillian.com  
To: oralonso@yahoo.com

**CHASE** 

**Crucial Message**

Message ID: [TR3D71452024](#)



Spam email?  
Label: yes, no

Dear Valued Client :  
oralonso@yahoo.com

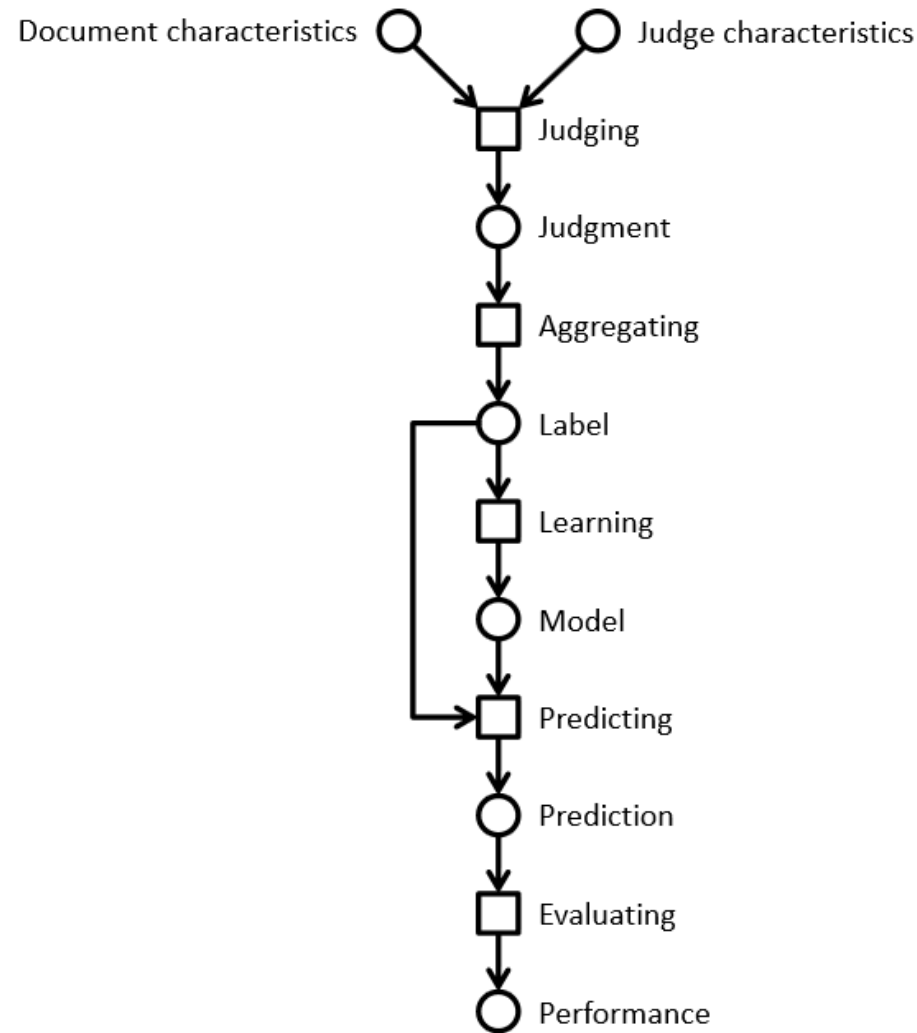
An imperative communication from **Chase** necessitates your immediate consideration. Neglecting to reply promptly could lead to restrictions on your account.

# Why we care?

- Provenance
- Reproducibility & debugging
- Explainability & interpretability
  - How a training set was created
- Bias and fairness
- Data management
  - ML/AI models live & die by the quality of input data
  - Metadata about labels
  - Maintenance

# Lifecycle of a label

- Information retrieval example



Using a crowd to label a data set

Using ML to process the complete data set

# Relevance labels

- Indicate whether a search result is valuable to a searcher
- Key in evaluation and optimization IR systems
- Editors or experts
  - TREC-style
- Crowdsourcing
- LLMs

# Careful with that ~~axe~~ data, Eugene

- In the era of big data and machine learning
  - labels -> features -> predictive model -> optimization
- Labeling perceived as boring
- Tendency to rush labeling
- Quality is key
  - Garbage in, garbage out
- Own the entire stack
  - Labeling, modeling, infrastructure, deployment

# The state of the field

- Human-labeled data is more important than ever
- Requirements
  - Throughput -> ASAP; I need the labels for yesterday
  - Cost -> cheap; if possible free
  - Quality -> top
- Performed as a one-off by 3<sup>rd</sup> party (crowd or editors)
  - Human Intelligence Task (HIT)
  - Micro-tasks
- Needs development work to get good results
- Very limited functionality in current platforms
  - Mechanical Turk, SageMaker (Amazon)
  - Figure Eight (Appen)
  - Toloka (Yandex)
  - Start-ups
- LLMs



# The need for humans

- Many examples where humans are involved
- Adult content and moderation
- Baby sitting algorithms

## The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed

BY ADRIAN CHEN 10.23.14 | 6:30 AM | PERMALINK

Share 60.5k Tweet 7,274 8+1 718 674 Pin it



<https://www.wired.com/2014/10/content-moderation/>

## 'A Permanent Nightmare': Pinterest Moderators Fight to Keep Horrifying Content Off the Platform

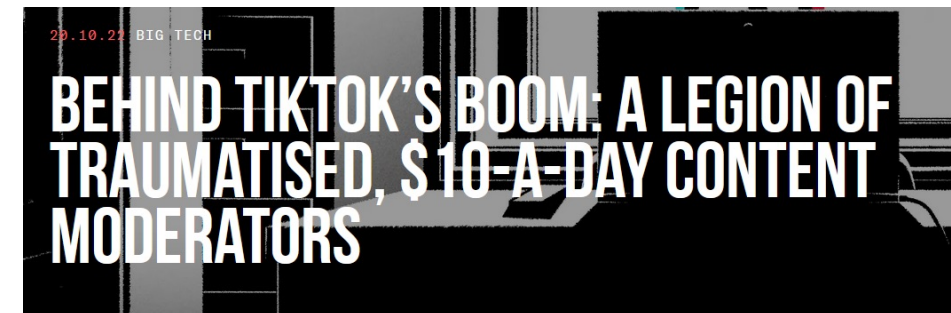
Moderators reported seeing child pornography content 'every couple hours'



Sarah Emerson · Follow

Published in OneZero · 11 min read · Jul 28, 2020

<https://onezero.medium.com/a-permanent-nightmare-pinterest-moderators-fight-to-keep-horrifying-content-off-the-platform-4d8e7ec822fe>



<https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatized-10-a-day-content-moderators>

# Problems

- Monolithic HITs
  - The structure of a HIT mirrors the structure of the task the developer is working on
  - Similar to Conway's law in software engineering
- Task complexity
- Lengthy instructions
  - RTFM doesn't work
- We don't think of HC/crowdsourcing as programming
- How to improve
  - Use established programming practices
  - Careful, we are dealing with humans and not machines

# A spectrum of labeling tasks

Nature of task	Aggregation approach	Evaluation technique
Objective question has a correct answer (objective)	Reliable judge assigns appropriate label for an item	Evaluate workers by comparing individual results with gold set
Judgment question has a best answer (partially objective)	Inter-rater agreement determines label for an item	Evaluate workers by comparing individual results with consensus
Subjective question has consistent answer (subjective)	Repeatable polling determines probability of a label for an item	Evaluate workers by computing the consistency of results between groups

# Prepare the environment

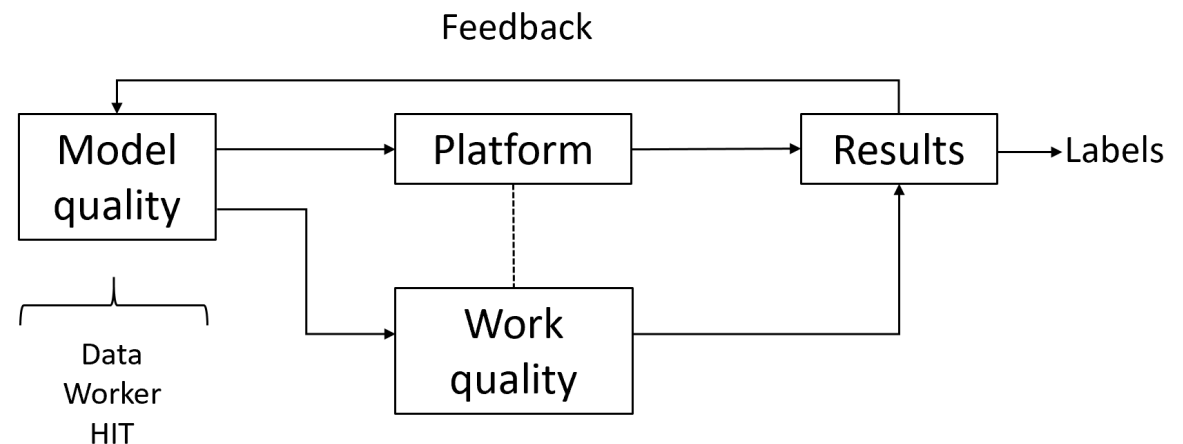
- Homework before you label
  - Assess the lay of the land
  - Identify your use cases
  - Understand your product's data
  - Design your HITs
  - Determine your guidelines
  - Communicate your task
  - Maintain high quality
- Ongoing vs. one-offs HITs
- Labels for the machine != labels for humans

# HIT design principles

- Self-contained, short, and simple
- Document presentation
  - Text alignment & legibility; reading level; multi-cultural and multilingual
- Cognitive biases
  - Implications on the final output: anchor effect, mere exposure, picture superiority
- Task complexity
  - High cognitive load; low usability, specific expertise

# Quality control in general

- Extremely important part of the task
- Approach as “overall” quality; not just for workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester.
- Quality framework
  - Module quality
  - Work quality
- Measuring agreement



# Algorithms used in practice

- Voting
  - Majority vote, Borda, tiers
  - Strong baseline
- Honey pots and programmatic gold
- Expectation-Maximization
- Get another label
- Adaptivity
  - Quality-cost tradeoff
  - How many workers?
  - When to stop?
  - Stopping rules
  - Automatic honey pots creation

# Behavioral features

- Focus on the way workers work instead of what they produce
- Task fingerprinting
- High correlation with work quality
- Wernicke
  - Information Extraction scenario
  - Weighted majority voting
  - Behavioral features outperform performance-based methods

J. Rzeszutarski and A. Kittur. “Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance”. UIST 2011.

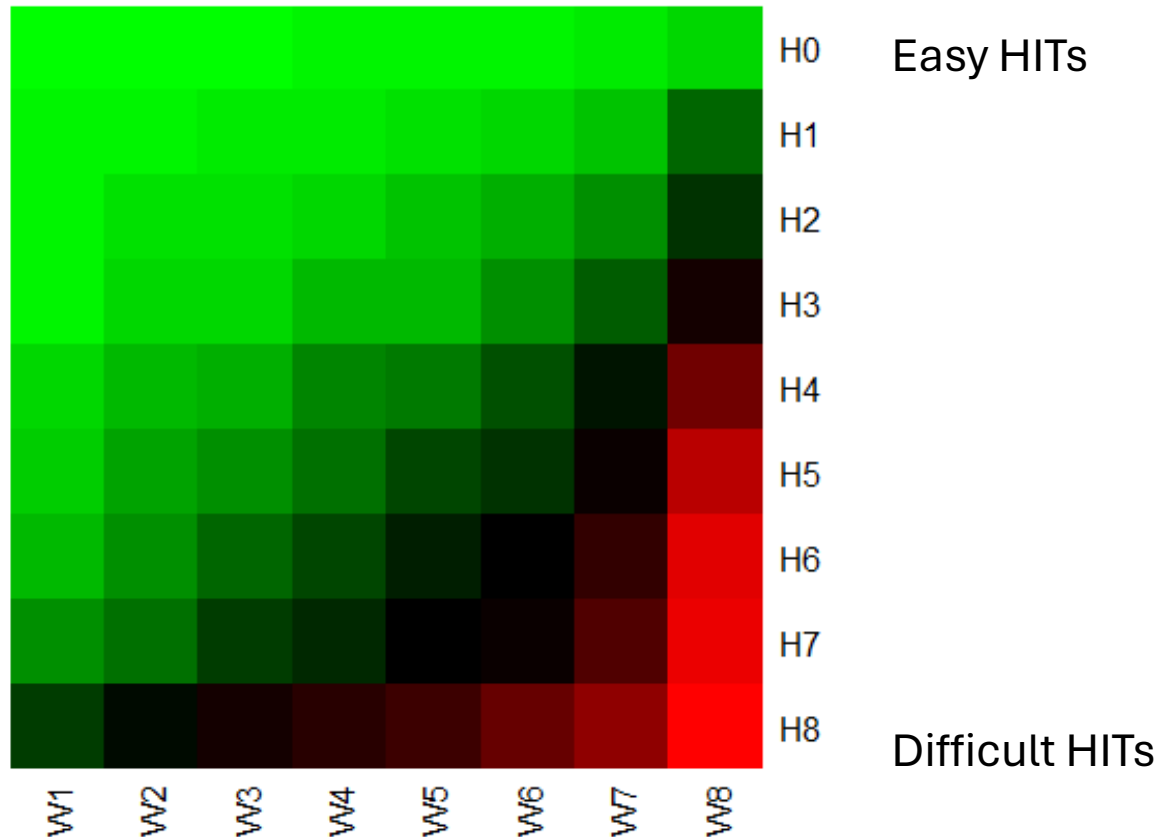
S. Han, P. Dai, P. Paritosh, D. Huynh. “Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control”. ACM TIST 2016



# Active learning

- Accuracy
  - Limited budget for annotating a small % of the unlabeled data
- Speed
  - Model more accurate more quickly
- Diversity
- Uncertainty sampling
  - Least confidence, margin of confidence, ratio of confidence
- Diversity sampling
  - Clustering to partition the data, real-world diversity

# Error rates for different worker/HIT groups



UHRS

2,700 HITs from 20 workloads

For difficult HITs

- Good workers are doing well
- Bad workers are doing poorly

For easy HITs

- Good workers are doing well
- Bad workers are doing well

Difficult HITs

Good  
workers

Bad workers

Let  $\text{error}(W_i, H_j)$  be the average error rate of the workers in the worker group  $W_i$  working on HITs  $H_j$

# Snorkel approach

- Formalizing programmatic labeling
- Models are commodities
  - `pip install <what-you-want>`
- Training data is the interface for software 2.0
- Labeling functions as black boxes that predict a label
- Learn from agreements/disagreements between labeling functions

# LLMs



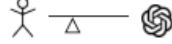
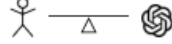
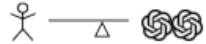
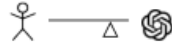
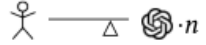

- Human labels are expensive
  - Expert > Crowd-based worker > LLM
  - Automatic label is not a new idea
- How about using LLMs to label documents?
- Potential advantages
  - Cost and performance
  - Allocate humans where are needed the most
- Potential issues
  - Reliability
  - Quality control

P. Thomas et al. "Large language models can accurately predict searcher preferences" [arxiv.org/abs/2309.10621](https://arxiv.org/abs/2309.10621)

G. Faggioli et al. "Perspectives on Large Language Models for Relevance Judgment" ICTIR 2023

# Spectrum of human-machine collaboration

- LLMs judgement quality
- LLMs cost
- Multiple LLMs as judges
- Truthfulness
- Bias
- Explanations/justifications

Collaboration Integration	Task Organization
Human Judgment	
	The human will do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance	
	The human assessor judges an LLM-generated summary of the document.
	Balanced competence partitioning. Humans and LLMs focus on tasks they are good at.
Human Verification	
	Two LLMs each generate a judgment, and a human selects the better one.
	An LLM produces a judgment (and an explanation) that a human can accept/reject.
	LLMs are considered crowdworkers, varied by specific characteristics, and controlled by a human.
Fully Automated	
	Fully automatic assessment.

# Prompting

- In-context learning
- New capabilities can be unlocked in LLMs
- LLM is prompted with a few in-context demonstrations
- Learns to perform a certain task
- Task performance is very sensitive to prompts

# Setup

- Similar to crowdsourcing work
- Take TREC judgement guidelines
  - HIT in Mturk
  - Prompt for GPT or similar LLM
- Compute agreement using Cohen's kappa
- Two main approaches
  - Prompt “as is”
  - Prompt engineering

# Prompt structure

- Relevance evaluation task
- Task instructions
  - You are a search quality rater evaluating relevance of web pages
- Query-document pair to be labelled
  - Query {query}
  - Document {document}
  - Relevant?
- Re-state the task
- Output format



# Prompts compared to HITs

role

You are a search quality rater evaluating the relevance of web pages. Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

2 = highly relevant, very helpful for this query  
1 = relevant, may be partly helpful but might contain other irrelevant content  
0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

## Query

A person has typed [query] into a search engine.

description,  
narrative

They were looking for: *description narrative*

## Result

Consider the following web page.

—BEGIN WEB PAGE CONTENT—

*page text*

—END WEB PAGE CONTENT—

## Instructions

Split this problem into steps:

Consider the underlying intent of the search.

aspects

Measure how well the content matches a likely intent of the query (M).

aspects

Measure how trustworthy the web page is (T).

Consider the aspects above and the relative importance of each, and decide on a final score (O).

multiple

We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.

Produce a JSON array of scores without providing any reasoning. Example: [{"M": 2, "T": 1, "O": 1}, {"M": 1}], [{"M": 1} . . .

## Results

[{

## Document Relevance Evaluation

Please evaluate the **relevance** of a document to the given topic. A document is relevant if it directly discusses the topic. Each document should be judged on its own merits. That is, a document is still relevant even if it is the thirtieth document you have seen with the same information.

### Tips

- Payment based on quality of the work completed. Please follow the instructions and be consistent in your judgments.
- **Bonus** payment if you provide a good justification
- Please justify your answer, otherwise you may not get paid.
- A document should not be judged as relevant or irrelevant based only on the title of the document. You **must** read the document.

### Task

Please evaluate the relevance of the following document about **art, stolen, forged**.

Description: What incidents have there been of stolen or forged art?

More information: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass-produced art reproductions). Pirated software, music, movies, etc. are not relevant.

## CHASE ENDS IN ARREST OF 3 AFTER LATEST JEWEL HEIST;

### CRIME: SANTA ANA ROBBERY FITS A PATTERN OF NEARLY 100 SIMILAR THEFTS IN THE WEST SINCE 1989. THE SUSPECTS MAY BE PART OF A LOS ANGELES COUNTY RING.

By WENDY PAULSON, TIMES STAFF WRITER

#### SANTA ANA

A freeway chase from Huntington Beach to Compton ended with the arrests of three men who allegedly robbed a department store jewelry counter at gunpoint, the latest in a series of Southland jewel heists, police said Thursday.

And although Orange County police were tight-lipped about investigations of two similar robberies in the last month, Los Angeles police said the incidents fit a pattern of nearly 100 similar thefts in the western United States since 1989 that may stem from a criminal network in southwest Los Angeles County.

Please rate the above document according to its relevance to **art, stolen, forged** as follows. Note that the task is about how relevant to the topic the document is.

- ☐ **Relevant.** A relevant document for the topic.  
☐ **Not relevant.** The document is not good because it doesn't contain any relevant information.

Does the topic look difficult? Please rate the difficulty from 1 to 5 (1=easy, 5=very difficult):

- ☐ **1** Easy ☐ **2** Somewhat easy ☐ **3** Neither easy nor difficult ☐ **4** Somewhat difficult ☐ **5** Very difficult

Please justify your answer or comment on your selection. Please use your own words. You may get a bonus payment if your comment is useful.

Submit

# Preliminary results

- With no prompt engineering
- With prompt features
  - R (role), D (description), A (aspects), M (multiple judges)
  - Performance varies per feature
  - Cohen's  $\kappa$  (0.20 to 0.64)

		Model	
		0	1 or 2
TREC assessor	0	866	95
	1 or 2	405	1585

Table 3: Overview of TREC-8 relevance judgment agreement between TREC assessors and each of the LLMs. Based on a sample of 1000 topic-document pairs.

LLM	Prediction	TREC-8 Assessors		Cohen's $\kappa$
		Relevant	Not relevant	
GPT-3.5	Relevant	237	48	0.38
	Not relevant	263	452	
YouChat	Relevant	33	26	0.07
	Not relevant	67	74	

Table 4: Overview of TREC-DL 2021 relevance judgment agreement between TREC assessors and each of the LLMs based on a sample of 400 question-passage pairs. TREC assessments were made on a graded scale from 3 (highly relevant) to 0 (not relevant).

LLM	Prediction	TREC-DL 2021 Assessors				Cohen's $\kappa$
		3	2	1	0	
GPT-3.5	Relevant	89	65	48	16	0.40
	Not relevant	11	35	52	84	
YouChat	Relevant	96	93	79	42	0.49
	Not relevant	4	7	21	58	

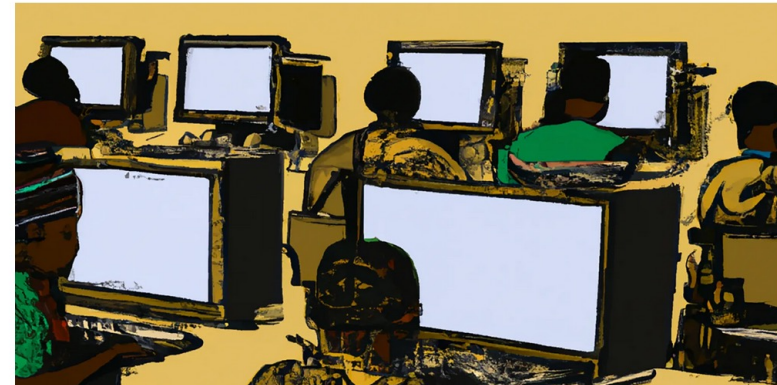
# Discussion

- In favor
  - LLMs are able to produce an explanation
  - This could be used to assist humans in relevance judgements
- Against
  - LLMs are not users
  - IR is about relevance to an information need
  - No proof that evaluation by LLM has any relationship to reality
- Things to consider
  - Reliability over time
  - Cost in prompt engineering
- Caveat
  - LLMs are systems
  - Query intent and answer construction

# LLMs and human computation

- For a LLM like ChatGPT,  $p(w_i|w_1, \dots, w_{i-1})$  is defined by a transformer
- LLM-based system do require editorial work
- Not different from any major property on the Internet

Exclusive: OpenAI Used Kenyan Workers on  
Less Than \$2 Per Hour to Make ChatGPT Less  
Toxic



<https://time.com/6247678/openai-chatgpt-kenya-workers/>

ARTIFICIAL INTELLIGENCE

## ChatGPT is powered by these contractors making \$15 an hour

Two OpenAI contractors spoke to NBC News about their work training the system behind ChatGPT.

<https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892>

# The main process is unchanged

- Regardless if labeling is done by machines or humans
- Three main components
  - Task design
    - HIT or prompt engineering
  - Data
  - Crowd
    - Human-based crowd or LLM-based crowd
- Quality control
- Debugging

	Machine computation	Human computation
Design	Throw away	Reluctant to throw away
Testing	Systematic	Ad-hoc
Debugging	Programmer's fault	Worker's fault

