Microsoft

Everything you always wanted to know about labeling (but were afraid to ask)

Omar Alonso

13-May-2019
San Francisco, CA

Disclaimer

The views, opinions, positions, or strategies expressed in this talk are mine and do not necessarily reflect the official policy or position of Microsoft.

Outline

Introduction

Wetware programming

Quality framework and techniques

The human side

Putting all things together

Systems and data pipelines

Conclusion

Introduction

The bad news first

Labeling is hard

Facebook

Snapchat

Points-Of-Interests (Foursquare, etc.)

Labeling is going to get more difficult

Enterprise

Personalization

New data sets

There is hope

Human computation

AI
CS
HCI
Economics
Behavioral sciences

Lots of research and new ideas

Some context

We assume supervised or semi-supervised learning

Large scale

Continuous

Working with people (editors, workers, experts, etc.)

Crowdsourcing != Mechanical Turk

Scope

This is not a comprehensive survey

Emphasis on fundamentals

Description of the many facets

Practical perspective

No silver bullet

“This is the type of AI that I am interested in - what can the human and machine do together, and not in the competition which can arise”

Richard Hamming

What is a label?

Finance Department <notificaciones@pgn.gob.gt>

to

Dear Winner,

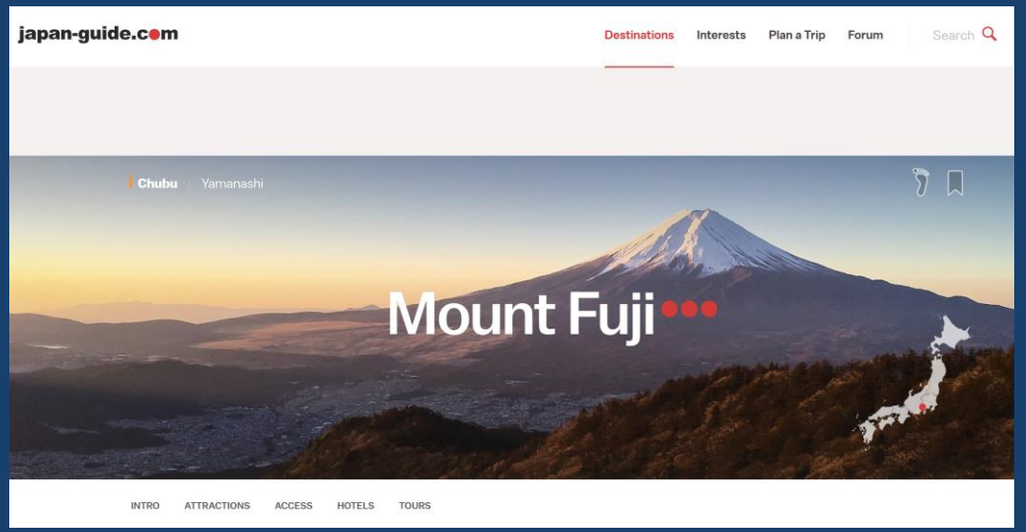
You have been awarded the sum of 8,000,000.00 (Eight Million Pounds sterling) with reference number 77100146. This compensation funds from the United Nation.Send us your personal details to deliver your funds.

Gloria Peter

Spam email?

Label: yes, no

What is a label?



Query = mount fuji

Task: Given the query, is the page relevant?

Answers: very, somewhat, not

Labels: 1, 0.5, 0

Human computation

Use humans as processors in a distributed system

Workers, raters, annotators, judges

Address problems that computers aren't good

Human Intelligence Task (HIT)

Available platforms

Amazon Mechanical Turk
CrowdFlower (Figure Eight)
UHRS (Microsoft)

L. von Ahn and L. Dabbish. "Designing games with a purpose". CACM, 2008

E. Law and L. von Ahn. *Human Computation*. Morgan & Claypool Publishers, 2011

A sample of HITs

KEYBOARD SHORTCUTS: Scroll: Shift + up/down Open Image

Extract purchased items from a shopping receipt (1-2 items)

Hit Reward: $0.01 for first 2 items + Bonus: $0.01 for every 4 items.

Real readable original receipt | Not a receipt or not readable

#	Type	Qty	Item Description	Price	Per Item
			EXAMPLE DESCRIPTION		
#	Item	3	CLOROX BLEACH	26.97	8.99
1.	Item	1			
2.	Item	1			

What is the transaction date & time on the receipt?

05/31/2017 HH : MM

SubTotal:

Sales Tax:

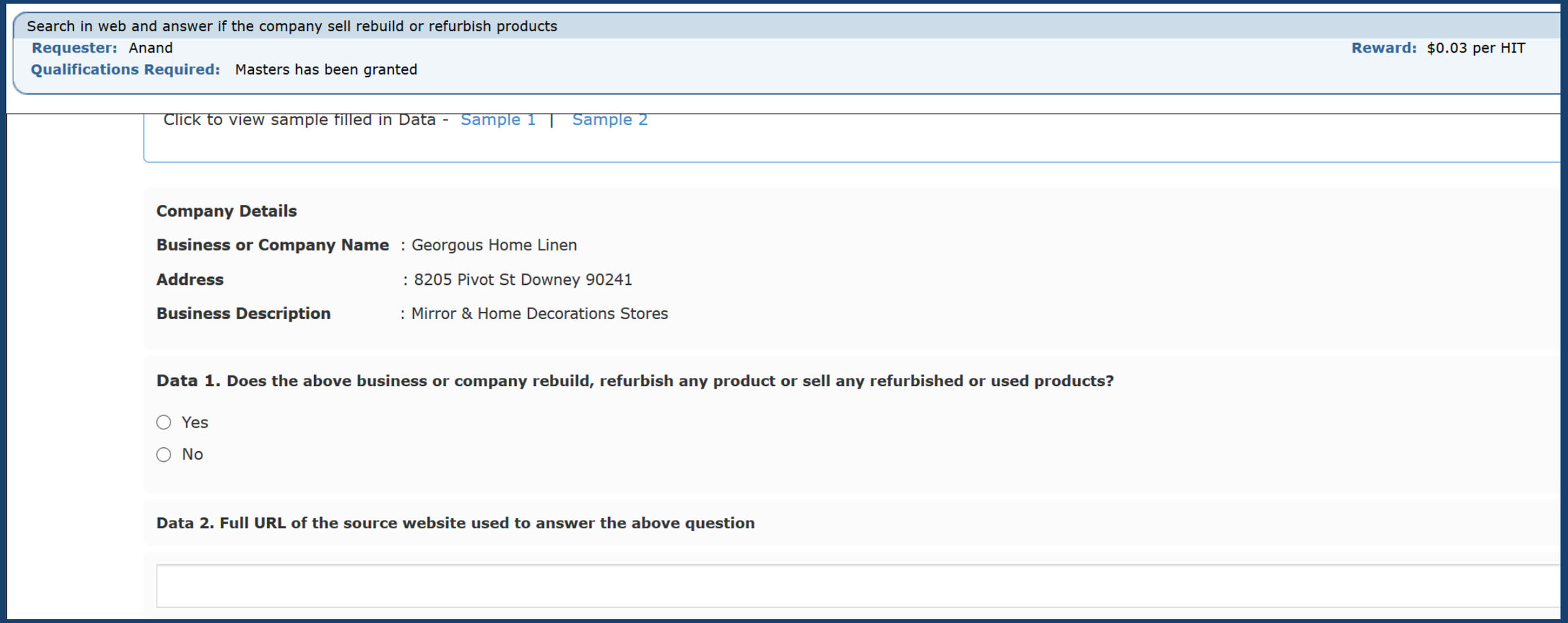
Total:

19.97

If total not captured in image, mark receipt above as "Not Readable or Not Receipt"

Submit

A sample of HITs



A sample of HITs

WHAT ARE THE ATTRIBUTES ON EACH OF THE FOLLOWING FACES?

VALID
- VALID
- INVALID

AGE
- BABY
- CHILD
- YOUNG
- MIDDLE AGE
- SENIOR

HAIR LENGTH
- BALD
- SHORT HAIR
- LONG HAIR
- NOT VISIBLE

HAIR COLOR
- BLACK HAIR
- BLONDE HAIR
- BROWN HAIR
- WHITE HAIR
- RED HAIR
- SALT AND PEPPER HAIR
- NOT VISIBLE

FACIAL HAIR
- GOATEE
- MUSTACHE
- CLEAN_SHAVEN
- BEARD

ETHNICITY
- ASIAN
- BLACK
- SOUTH ASIAN - INDIAN
- WHITE
- MIDDLE EASTERN
- HISPANIC

GENDER
- MALE
- FEMALE

EYEWEAR
- EYEGLASSES
- SUNGLASSES
- NO EYEWEAR

FACE SHAPE
- OVAL FACE
- ROUND FACE
- SQUARE FACE
- LONG FACE

FOREHEAD
- BANGS
- VISIBLE NO LINES
- LINES

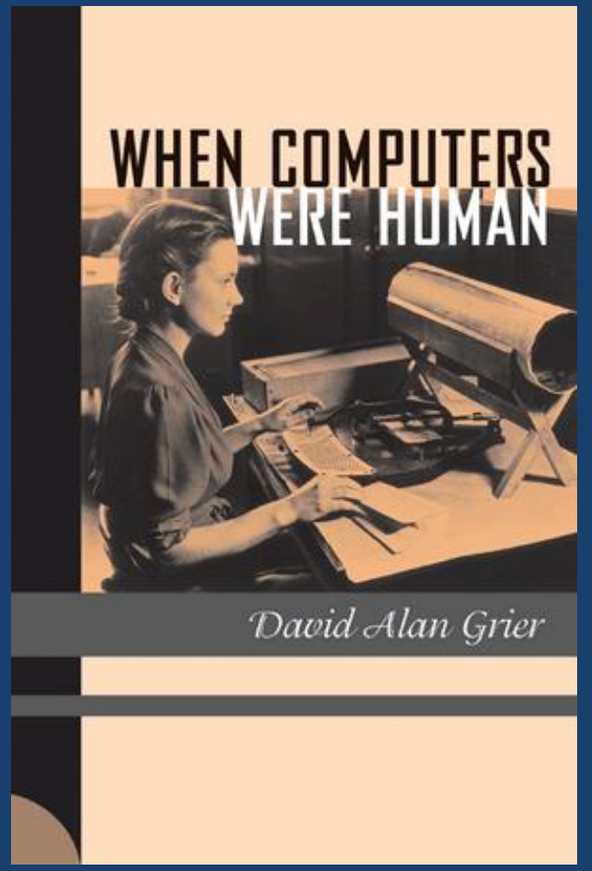
FOREHEAD SIZE
- SMALL FOREHEAD
- LARGE FOREHEAD

EYEBROWS
- THICK BROW
- THIN BROW
- ONE BROW
- NO BROW

EYE SHAPE
- ALMOND EYES
- ROUND EYES
- NOT VISIBLE

In case you didn’t know

You are a computer



Why we need labels?

Information retrieval

Natural language processing

Machine learning

Active learning

Artificial intelligence

A sample of common tasks

Content moderation

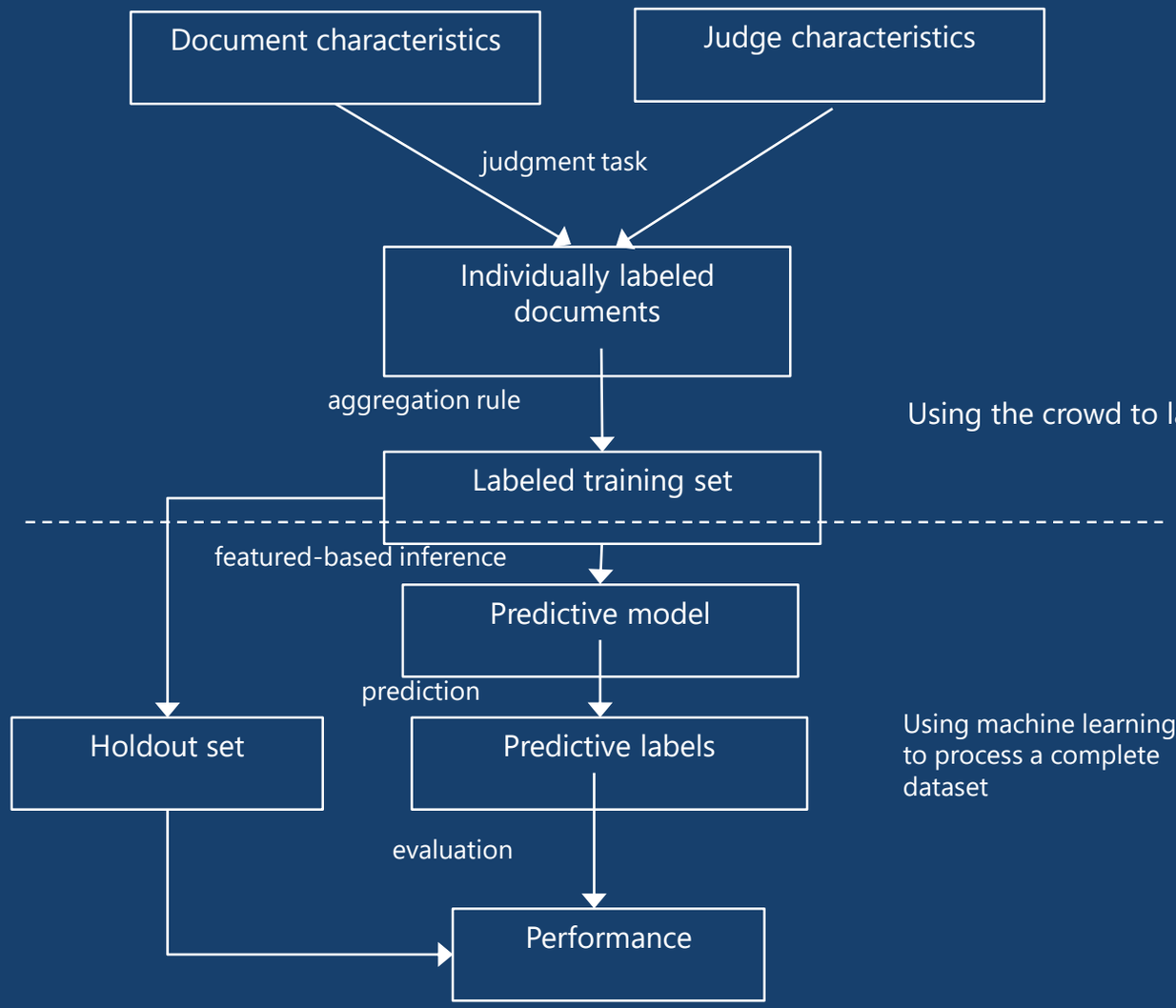
Information extraction

Search relevance

Entity resolution

A. Marcus and A. Parameswaran. *Crowdsourced Data Management.* Foundations and Trends in Databases, 2016

Lifecycle of a label – IR example



Using the crowd to label a dataset

Using machine learning to process a complete dataset

Careful with that ~~axe~~ data, Eugene

In the era of big data and machine learning

labels -> features -> predictive model -> optimization

Labeling perceived as boring

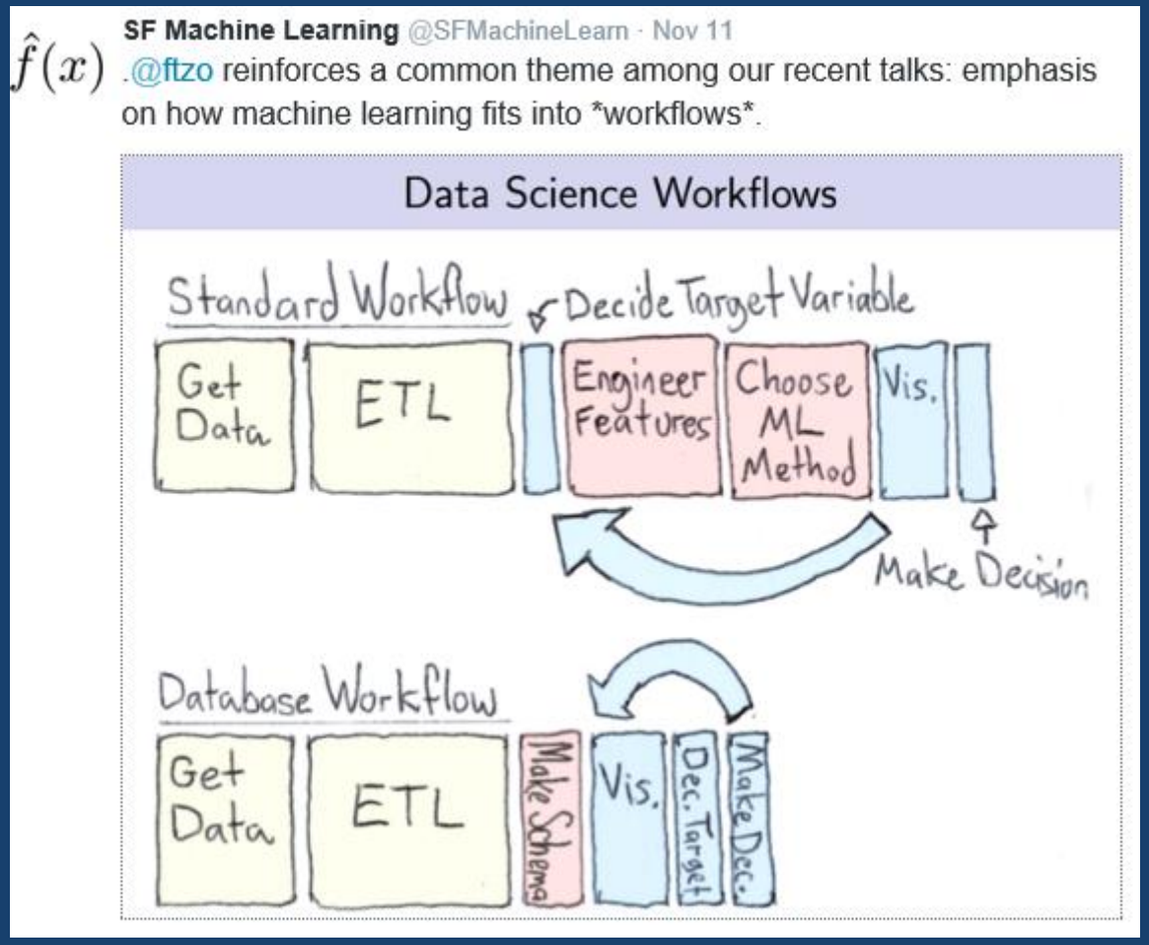
Tendency to rush labeling

Quality is key

Garbage in, garbage out

Big data, ML, and data science

Labels -> features -> predictive model -> optimization



… there is always a human

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 +1 718 Share 674 Pin it 4

Example

Facebook news selection is in hands of editors not algorithms, documents show

The Intersect | Analysis

Facebook has repeatedly trended fake news since firing its humar editors

FROM SLATE, NEW AMERICA, AND ASU

Trending Bad

How Facebook's foray into automated news

Facebook will hire 1,000 and make ads visible to fight election interference

Posted Oct 2, 2017 by Josh Constine (@joshconstine)

Not just Facebook

Kurt Wagner / Recode:

LinkedIn introduces trending topics section curated by human editors, rolling out Wednesday to US users on mobile and desktop

– LinkedIn pulls a Facebook. – LinkedIn is known for helping people find their next job, but now it wants to help people find their news, too.

Moments, the best of Twitter in an instant

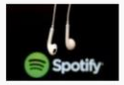
Tuesday, October 6, 2015 | By Madhu Muthukumar (@justmadhu), Product Manager, Moments [12:51 UTC]

Today, most moments are assembled by our curation team, and some are contributed by partners like Bleacher Report, Buzzfeed, Entertainment Weekly, Fox News, Getty Images, Mashable, MLB, NASA, New York Times, Vogue and the Washington Post. While we're working with a small gr

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Silicon Valley Struggles to Add Conservatives to Its Ranks



Tencent Music, Spotify Weigh Stake Swap Ahead of IPOs

China's Tech Giants Have a Second Job: Helping Beijing Spy on Its People

Tencent's Sharp Rally Just Hit the Skids

TECH | KEYWORDS

Without Humans, Artificial Intelligence Is Still Pretty Stupid

There are likely hundreds of thousands of people, world-wide, whose work is sold as AI, says one expert

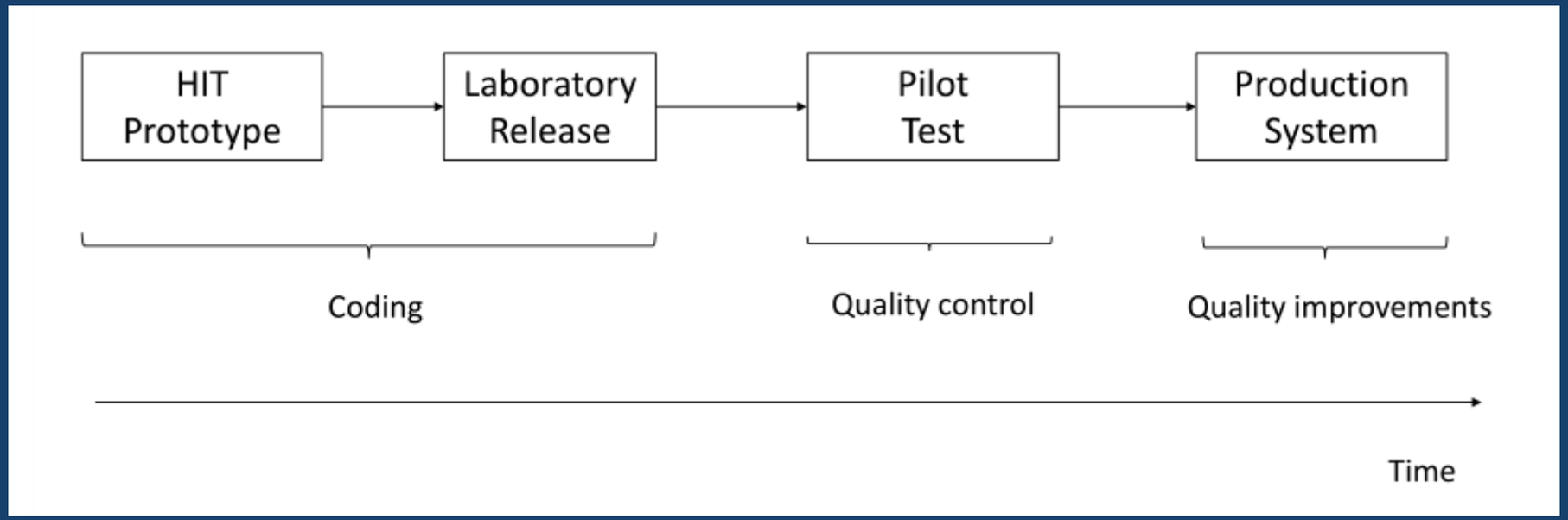
Why Periscope hired an editor in chief

by Brian Stelter @brianstelter

May 2, 2016: 9:15 PM ET

Recommend 877

Incremental development process



Working examples

Couple of HITs

Flip a coin

Query classification

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 3rd party (crowd or editors)

Non trivial amount of work to get good results

Very limited functionality in current platforms

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

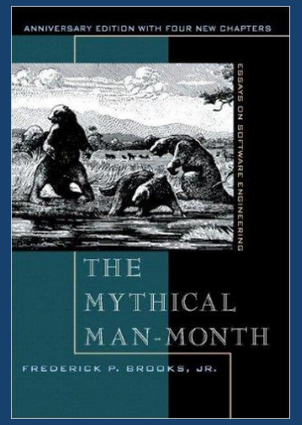
Wetware programming

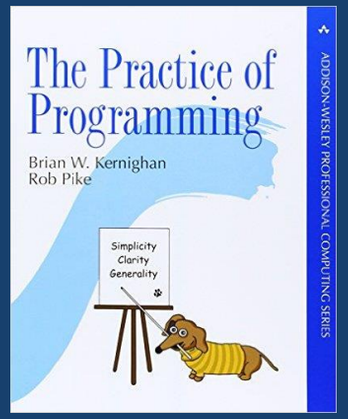
Generic approach

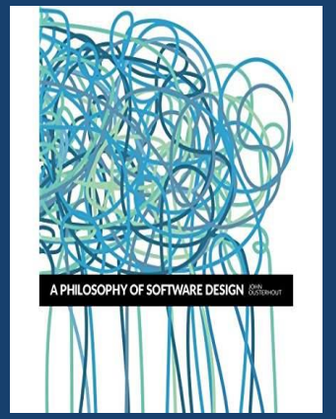
Well-known techniques for writing programs

Humans executing a task on a machine

A programming view for humans and machines







Humans executing code

Instruction set is somewhat unknown

Latency

Cost/incentives

Errors

Task difficulty

Human factors

Asking questions

Part art, part science

Instructions are key

Workers may not be experts so don’t assume the same understanding in terms of terminology

Show examples

Hire a writer

Engineer writes the specification
Writer communicates

HIT design

Self-contained, short, and simple

Document presentation & design

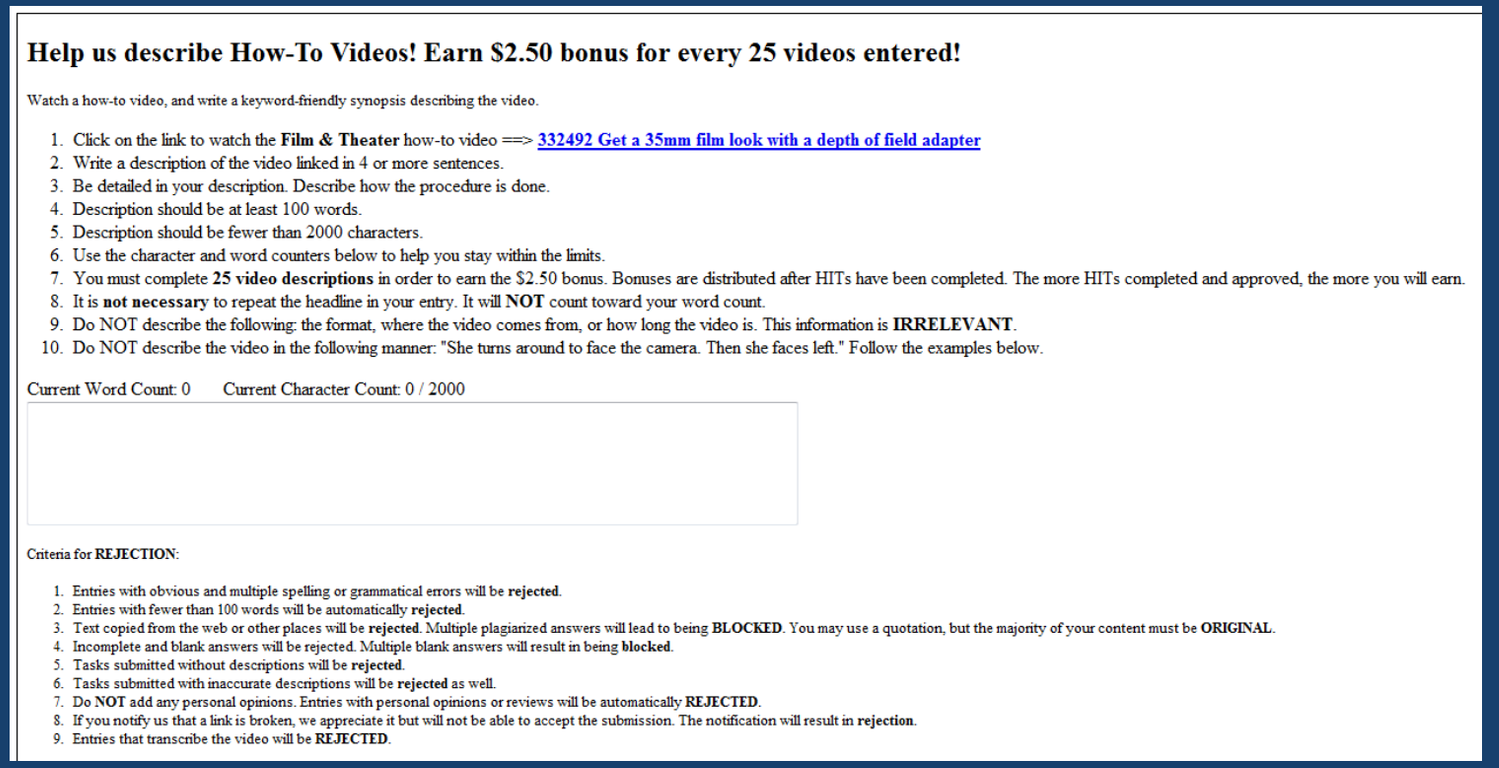
Engage with the worker

Need to grab attention

Localization

Examples - I

Asking too much, task not clear, "do NOT/reject"
Worker has to do a lot of stuff



Help us describe How-To Videos! Earn $2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the **Film & Theater** how-to video ==> **332492 Get a 35mm film look with a depth of field adapter**
2. Write a description of the video linked in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete **25 video descriptions** in order to earn the $2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is **not necessary** to repeat the headline in your entry. It will **NOT** count toward your word count.
9. Do NOT describe the following: the format, where the video comes from, or how long the video is. This information is **IRRELEVANT**.
10. Do NOT describe the video in the following manner: "She turns around to face the camera. Then she faces left." Follow the examples below.

Current Word Count: 0 Current Character Count: 0 / 2000

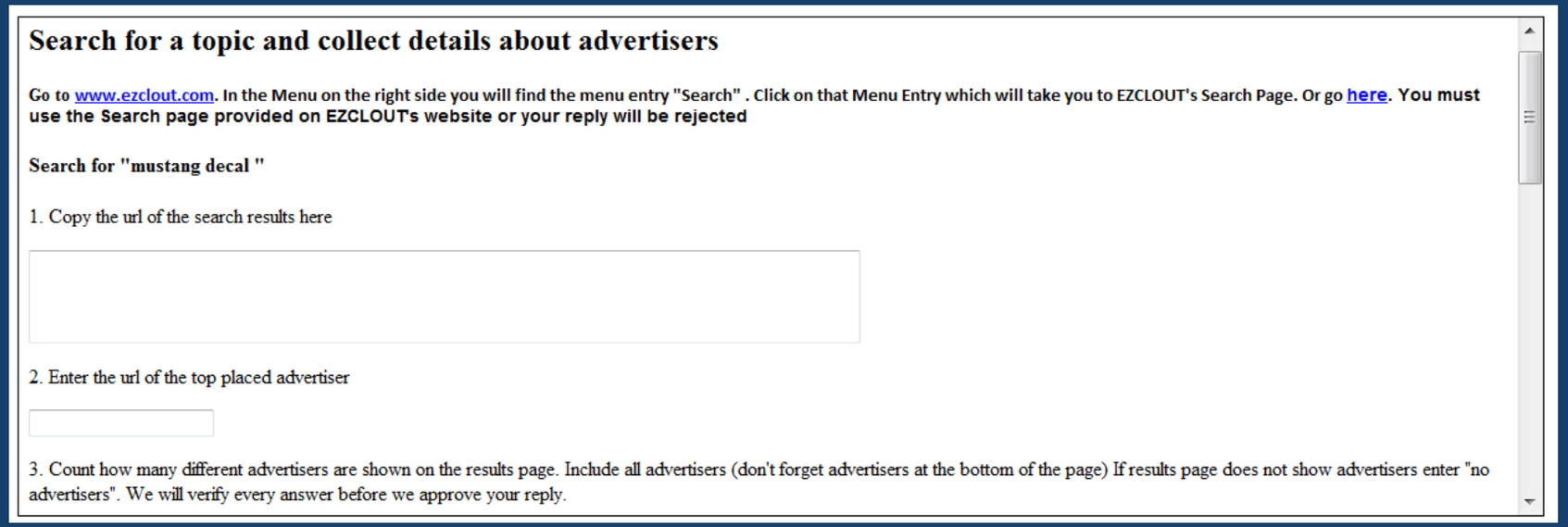
Criteria for **REJECTION**:

1. Entries with obvious and multiple spelling or grammatical errors will be **rejected**.
2. Entries with fewer than 100 words will be automatically **rejected**.
3. Text copied from the web or other places will be **rejected**. Multiple plagiarized answers will lead to being **BLOCKED**. You may use a quotation, but the majority of your content must be **ORIGINAL**.
4. Incomplete and blank answers will be rejected. Multiple blank answers will result in being **blocked**.
5. Tasks submitted without descriptions will be **rejected**.
6. Tasks submitted with inaccurate descriptions will be **rejected** as well.
7. Do **NOT** add any personal opinions. Entries with personal opinions or reviews will be automatically **REJECTED**.
8. If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in **rejection**.
9. Entries that transcribe the video will be **REJECTED**.

Example - II

Lot of work for a few cents

Go here, go there, copy, enter, count ...



Search for a topic and collect details about advertisers

Go to www.ezclout.com. In the Menu on the right side you will find the menu entry "Search" . Click on that Menu Entry which will take you to EZCLOUT's Search Page. Or go here. You must use the Search page provided on EZCLOUT's website or your reply will be rejected

Search for "mustang decal "

1. Copy the url of the search results here

2. Enter the url of the top placed advertiser

3. Count how many different advertisers are shown on the results page. Include all advertisers (don't forget advertisers at the bottom of the page) If results page does not show advertisers enter "no advertisers". We will verify every answer before we approve your reply.

Other design principles

Text alignment & legibility

Reading level: complexity of words and sentences

Multi-cultural and multi-lingual

Special needs communities (e.g., simple color blindness)

Cognitive biases

Implications on final output

Anchor effect

Tendency to rely on the first piece of information

Mere exposure effect

Tendency for people to like things the more they are expose to them

Social desirability

Tendency for people to respond in a socially acceptable direction

Picture superiority effect

Skip things that look ordinary; expect/favor unusual things

Content aspects

Presentation

Data familiarity

Metadata and internationalization

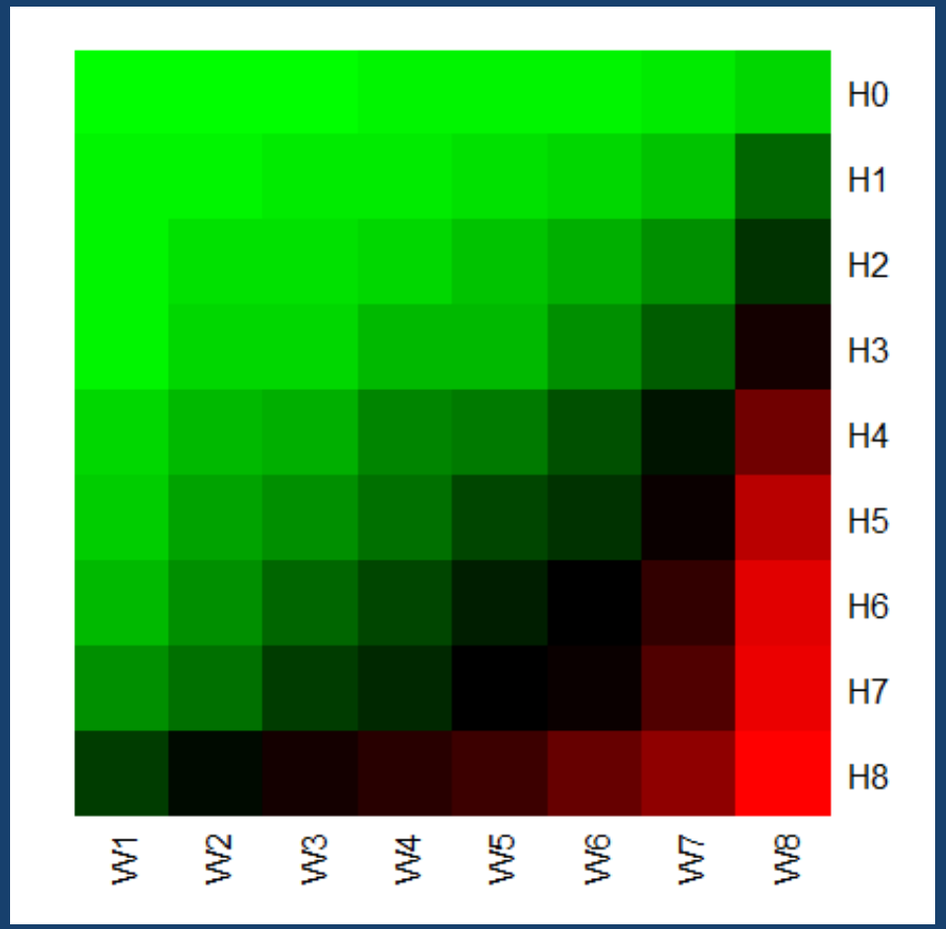
Task complexity

High cognitive load from a worker

Specific expertise to accomplish the work

Low usability

Error rates for different worker/HIT groups



Easy HITs

Difficult HITs

Good workers

Bad workers

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

Sensitive data

Social annotations in a search engine

Likes, RTs

Taxonomy of social relevance aspects

Query, social, content

Social connections

Circle, affinity, expertise, geographical distance, interest valence

Personal data

Very difficult to evaluate

P. Pantel, M. Gamon, O. Alonso, K. Haas. Social Annotations: Utility and Prediction Modeling, SIGIR 2012

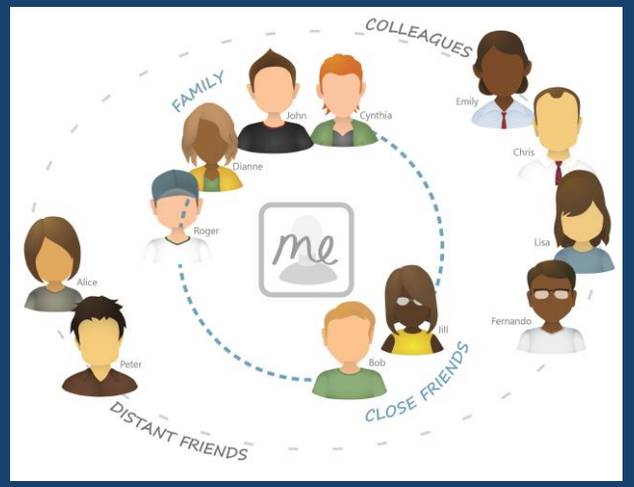
Generating social annotations

Bing search logs

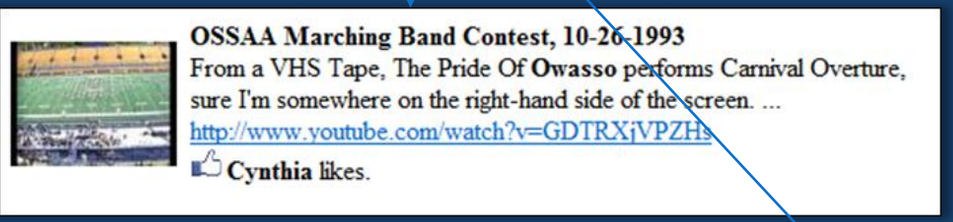
Original annotation = <Thumbnail + Title + Snippet + URL>

Scenario template

Generator



Simulated social network



What is the value of this social annotation?

Simulated social network construction

Scenario template generation

Task

You query Bing for **muniets to midnight** and one of the results is illustrated below. **Jill**, someone in your social network, has liked, disliked or shared this result. Recall that Jill is a **CloseFriend**. Also, assume that **Jill** is a **Expert** who **Dislike** the web page. Local information about **Jill**: **NA**

What is the added value of this social annotation?

Linkin Park - "Leave Out All The Rest
The official music video for "Leave Out All The Rest" from the album Minutes **To Midnight**. Directed by: Joe Hahn.
http://www.youtube.com/watch?v=LBTXNPZPfbE
Jill Dislike.

Please answer the following question:

How relevant is the annotation to the web results?

- **There is significant added value**. The annotation is substantially relevant, useful, or of interest to you.
- **There is some added value**. The annotation is somewhat relevant, useful or of interest to you.
- **No added value**. The annotation is not relevant, useful or of interest to you.
- **Don't know**. I don't have enough information to assess this annotation (please add a comment in the box below).
- **Non English/Service error**. Can't judge because content is non-English or there is a service error (e.g., 404 message, image didn't load, etc.) (please add a comment in the box below).

Quality Assurance

Label quality

Quality

Meets internal customer needs

Free from deficiencies

Process

Don’t rush labeling

Don’t outsource

Own it end to end

Large scale

Continuity

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

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.

When to assess work quality?

Beforehand

How: “qualification tests” or similar mechanism
Purpose: screening, selection, recruiting, training

During

How: assess labels as worker produces them
Like random checks on a manufacturing line
Purpose: calibrate, reward/penalize, weight

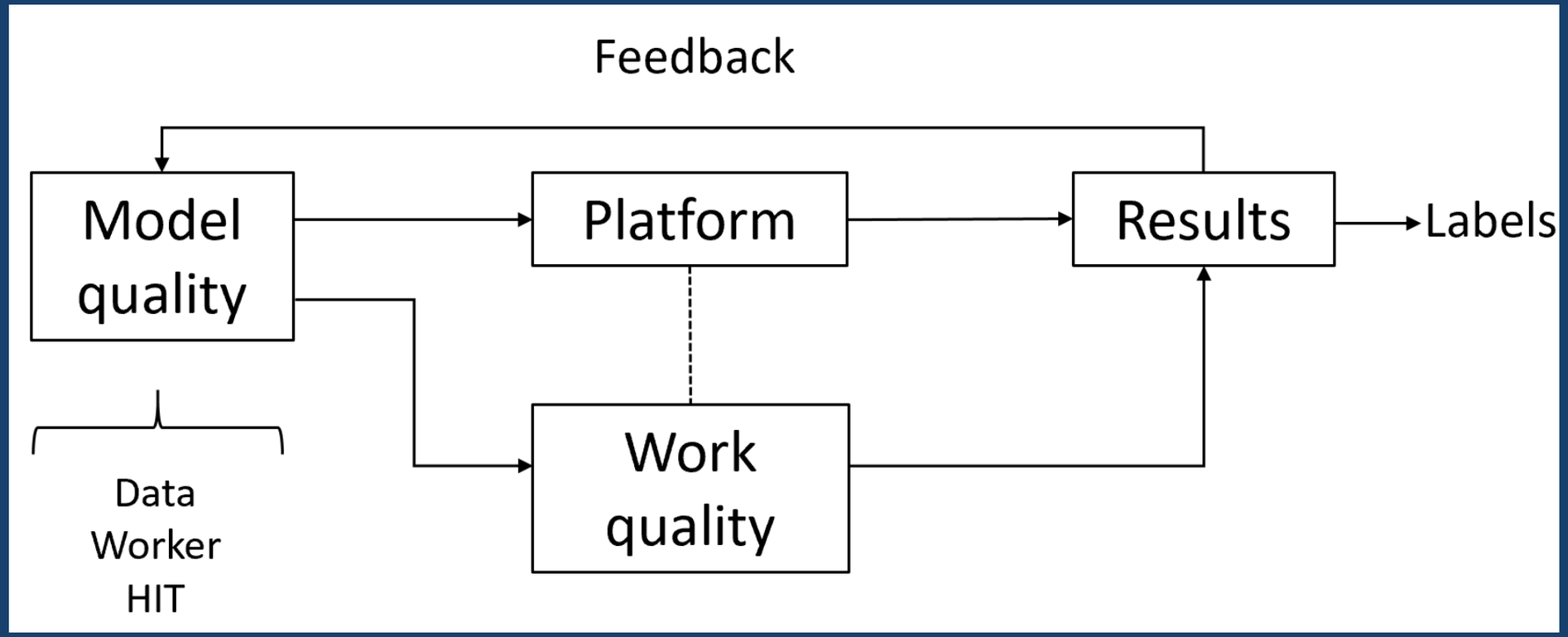
After

How: compute accuracy metrics post-hoc
Purpose: filter, calibrate, weight, retain

Quality framework

Module quality

Work quality



Worker qualification

Qualification tests

Advantages

Great tool for controlling quality

Adjust passing grade

Disadvantages

Extra cost to design and implement the test

May turn off workers, hurt completion time

Refresh the test on a regular basis

Hard to verify subjective tasks like judging relevance

Try creating task-related questions to get worker familiar with task before starting task in earnest

Reliability and validity

Redundancy

What to look for

Agreement, reliability, validity

Inter-agreement level

Agreement between judges

Agreement between judges and the gold set

Statistics

Agreement and consistency

Cohen's kappa (2 raters)

Fleiss' kappa (any number of raters)

Krippendorff's alpha (any number of raters; missing values)

Kuder-Richardson (KR-20)

Sample code

R packages psy and irr

```
>library(psy)
>library(irr)
>my_data <- read.delim(file="test.txt", head=TRUE, sep="\t")
>kappam.fleiss(my_data,exact=FALSE)

>my_data2 <- read.delim(file="test2.txt", head=TRUE, sep="\t")
>ckappa(my_data2)
```

k coefficient

Different interpretations of k

For practical purposes you need to be >= moderate

Results may vary

k	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Stats discussion

A high value for k or alpha is not enough

Contingency tables

Patterns of disagreements

Detection theory

Sensitivity measures

High sensitivity: good ability to discriminate
Low sensitivity: poor ability

Stimulus Class	"Yes"	"No"
S1	Hits	Misses
S2	False alarms	Correct rejections

Hit rate H = P("yes"|S2)
False alarm rate F = P("yes"|S1)

A background story

Twitter classifier

Detect if a tweet is interesting or not?

Standard ML approach

Get labels

Feature engineering

Modeling with a tool (e.g., Weka, etc.)

Production classifier

Moderate kappa values

What's going on?

HIT as baseline

	B1 (older, random)	B2 (recent, random)
% interesting	16.7%	14.3%
Krippendorff's α	0.013	0.052

Worker reliability and expertise

Borrowed idea from reCAPTCHA: use of control term

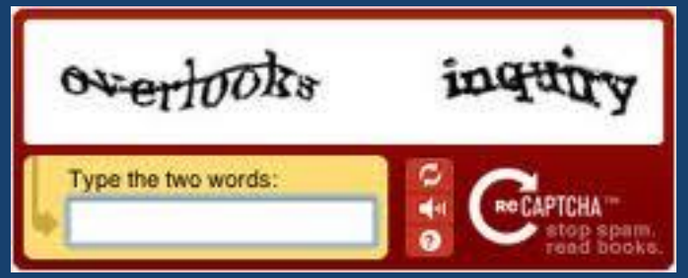
Human Intelligence Data Driven Enquires (HIDDEN)

2 more questions as control

1 algorithmic

1 semantic

Adapt your labeling task

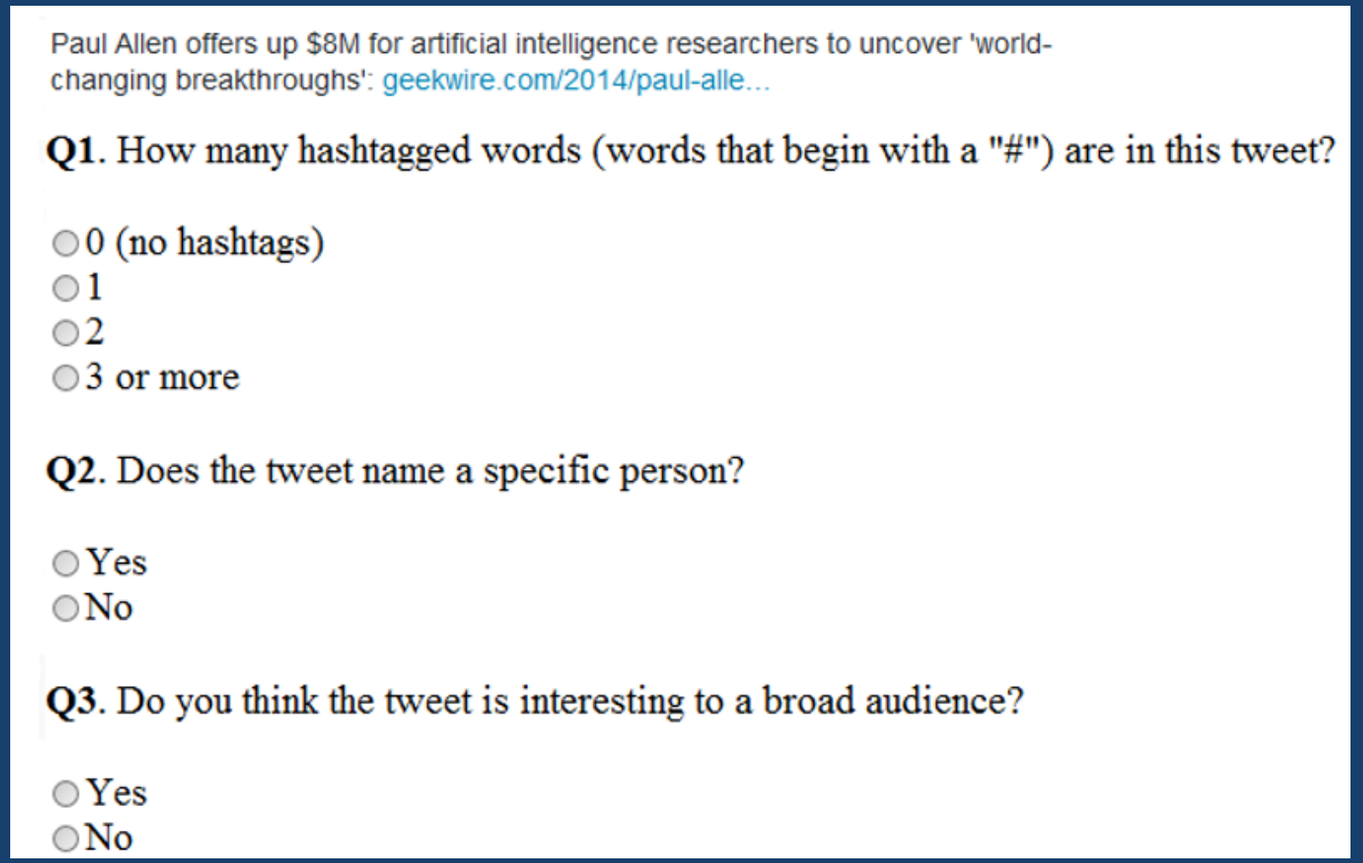


HIT with HIDDENs

Tweet de-branded

HIDDENs

The main question



Q1 (alpha = 0.888)

Q2 (alpha = 0.708)

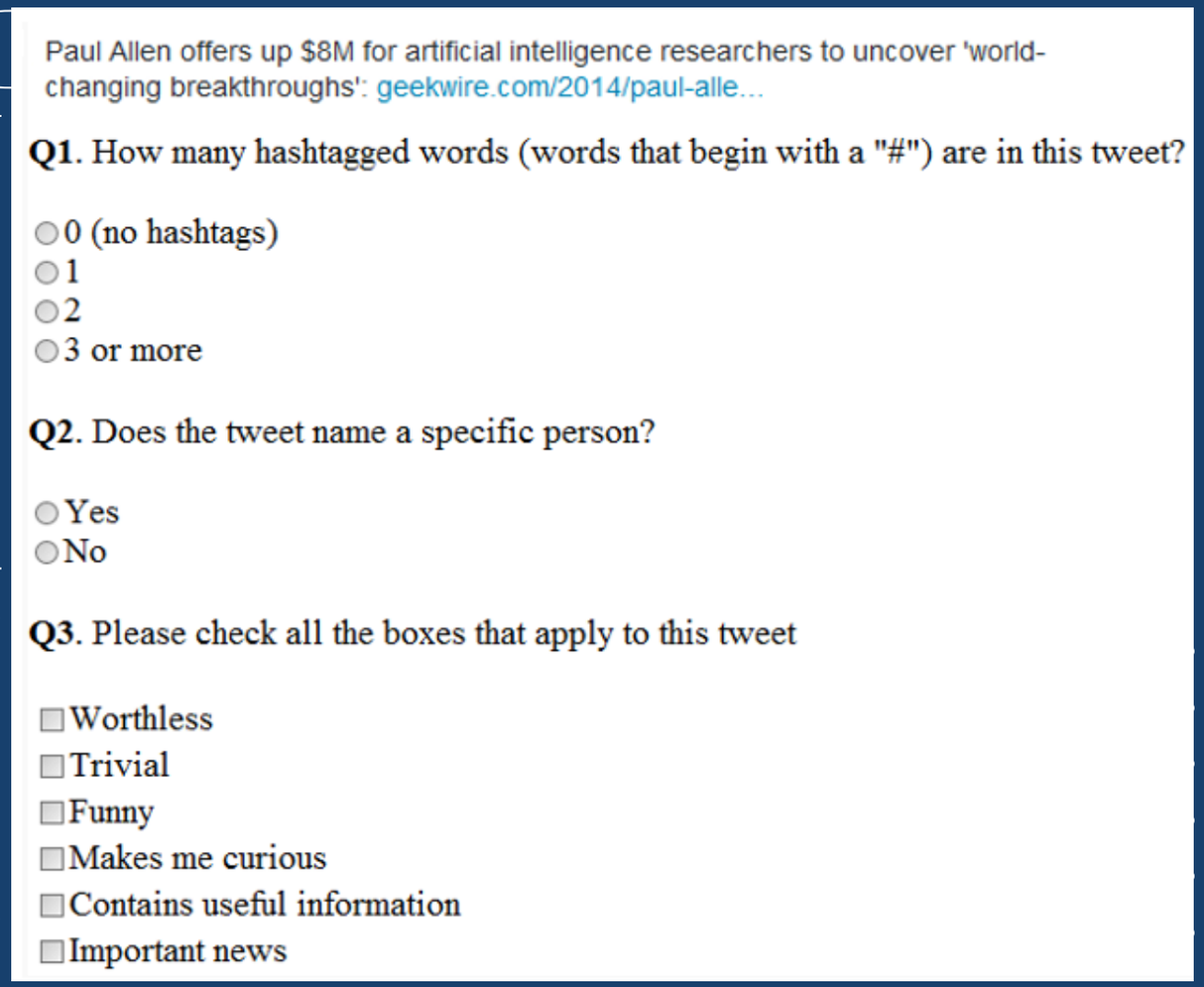
Q3 (alpha = 0.160)

HIT re-design

Tweet de-branded

HIDDENs

Breakdown by categories to get better signal



Q1 (alpha = 0.910)

Q2 (alpha = 0.758)

Q3 Worthless (alpha = 0.384)
Q3 Trivial (alpha = 0.097)
Q3 Funny (alpha = 0.134)
Q3 Makes me curious (alpha = 0.056)
Q3 Contains useful info (alpha = 0.079)
Q3 Important news (alpha = 0.314)

Findings from designs

No quality control issues

Eliminating workers who did a poor job on Q1 didn't affect inter-rater stats on Q2 and Q3

Interestingness is a subjective notion

We can still build a classifier that identifies tweets that are interesting to a majority of users

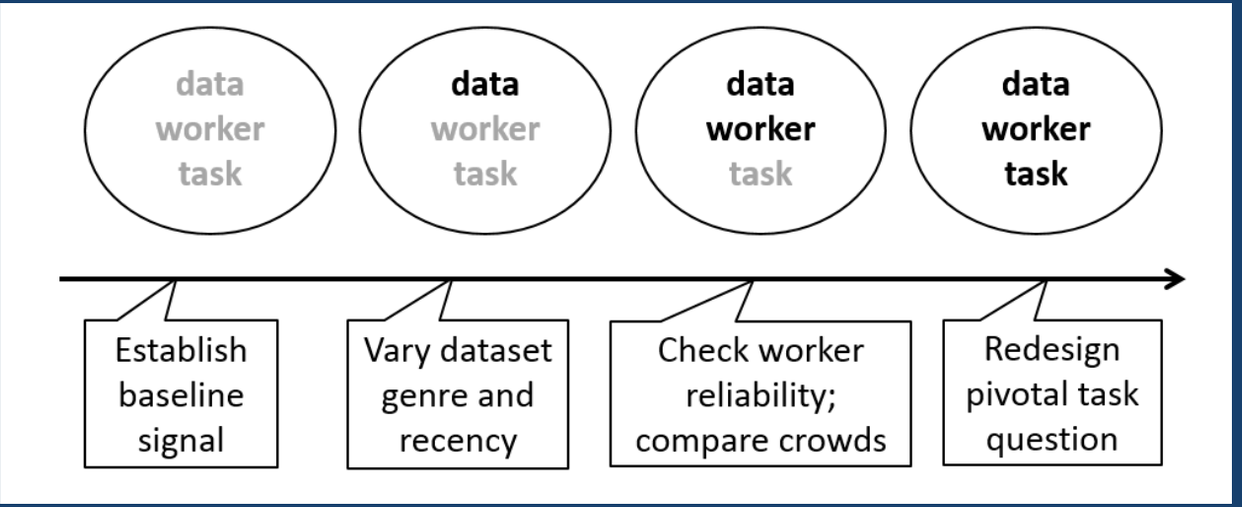
Debugging framework

Human computation tasks are difficult to debug

Multiple contingent factors

Framework

Data-worker-task
Rapid iteration
Small data sets
Emphasis on testing before scaling



O. Alonso, C. Marshall, M. Najork. “Debugging a Crowdsourced Task with Low Inter-rater Agreement”. JCDL 2015

Algorithms and Techniques

How do we measure work quality?

Compare worker's label vs.

Known label

Other workers' labels

Model predictions of workers and labels

Verify worker's label

Yourself

Tiered approach

Comparing to known answers

Gold, honey pots, verifiable answer

Assumes you have known answers

Cost vs. Benefit

Producing known answers (experts?)

% of work spent re-producing them

Finer points

What if workers recognize honey pots?

Maintenance

Comparing to other workers

Consensus, plurality, redundant labeling

Well-known metrics for measuring agreement

Cost vs. Benefit

% of work that is redundant

Finer points

Is consensus “truth” or systematic bias of group?

What if no one really knows what they’re doing?

Low-agreement across workers indicates problem is with the task

Algorithms used in practice

Majority vote

Honey pots and programmatic gold

EM

Get another label

V. Sheng, F. Provost, P. Ipeirotis. “Get Another Label? Improving Data Quality Using Multiple, Noisy Labelers”. KDD 2008.

D. Oleson et al. “Programmatic gold: Targeted and scalable quality assurance in crowdsourcing”. In Human Computation Workshop, 2011.

Voting

Aggregation functions

Majority: $l \geq \left\lceil \frac{N+1}{2} \right\rceil$

l is collectively accepted iff there is a majority of individual accepting it

Unanimity: $l \geq N$

l is collectively accepted iff all individuals accept it

Quota: $l \geq t$

l is collectively accepted iff there are at least t individuals that accept it

Distance: find closest consensus using distance function

Borda

Micro-breaks

Micro-diversions as relief

Improves worker retention and answer speed

Work quality

Contextual effects

P. Dai, J. Rzeszotarski, P. Paritosh, E. Chi. "And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions". CSCW 2015.

Honey pots

Include predefined gold data in the data set

Effective solution in practice

Initial cost for producing and maintaining honey pots

Problems when rejecting work

Programmatic gold

Process of generating gold units by injecting known types of errors in the dataset

Step 1: identifying worker errors through manual inspection.

Step 2: define set of data transformations that alter certain attributes to produce a new unit that

(1) differs from the original violating task requirements
(2) looks like the original.

Each data manipulation produces a gold unit that correspond to a particular error.

Scalable solution: takes the same effort to create 100 or 1000 gold units.

Crowd-workers reviewing work

Soylent

Find-fix-verify

Interactive crowdsourcing

FamilySearch

Arbitration

Peer review

Technique does work

Feedback leads to better work and motivates workers

D. Hansen et al. “Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing”, CSCW 2013

M. Bernstein et al. “Soylent: A Word Processor with a Crowd Inside”, UIST 2010

Workers reviewing work

Human-powered technique

Easy to implement

May slow down the process

Examples

Find-fix-verify

Arbitration

Peer-review

Justification

Ask works to justify an answer

MicroTalk

Assess

Justify

Reconsider

Expectation Maximization

Estimate workers' accuracy and the final HIT result at the same time

Expectation (of missing data) step and a Maximization (maximum likelihood estimation step).

1. Obtain some initial estimates of the missing data
2. Calculate the maximum likelihood estimates for the quantities of interest as if the missing data had been found
3. Now calculate new estimates of the missing data
4. Repeat steps 2 and 3 until both the maximum likelihood estimates and the missing data estimates converge

Adaptivity

Explore-exploit approaches

Quality-cost tradeoff

Adaptive exploration

How many workers?

When to stop?

{facebook, www.facebook.com}

{solar storms, www.solarstorms.org}

Stopping rules

Anonymous and non-anonymous workers

Automatic honey pots creation

I. Abraham, O. Alonso, V. Kandylas, R. Patel, S. Shelford, A. Slivkins. “How Many Workers to Ask? Adaptive Exploration for Collecting High Quality Labels”. SIGIR 2016

Behavioral features

Focus on the way workers work instead of what they produce

Task fingerprinting

High correlation with work quality

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

Practical considerations

What to use?

Depends on complexity and infrastructure access

Voting and honey pots

Cheap and easy to implement

EM-based approaches

Assumes historical performance

Worker verification

More HIT development

Recommendation

Start with a simple quality algorithm

Collect more data and then try more advanced techniques incrementally.

The human side

Demographics

Several studies

U.S. and India (47%, 34%; 80%, 20%)

Number of workers at any time is ~ 2,500

Incentives

Main incentive is money

Fair payment

Minimum wage in the U.S.

96% of workers on MTurk earn below minimum wage

Increased financial incentives increase quantity but not quality

Worker experience

Intrinsic motivation is a predictor of job satisfaction

Workers are not independent workers

Communities

Legal and ethics

Very little research

Be mindful of the law

Be open and honest about expectations and data collected

Discussion

Worker

Instructions are not clear

I'm not a spammer – I just don't get what you want

Boring task

A good pay is ideal but not the only condition for engagement

Requester

Attrition

Balancing act: a task that would produce the right results and is appealing to workers

I want your honest answer for the task

I want qualified workers; system should do some of that for me

Managing crowds and tasks is a daily activity

More difficult than managing computers

Putting all things together

So far …

This is all good but looks like a ton of work

The original goal: good labels

Data quality and experimental designs are *preconditions* to make sure we get the right stuff

Labels will be used for rankers, ML models, evaluations, etc.

Don’t cut corners

Development process

Prototype development

Small data set
Internal team
Testing and debugging

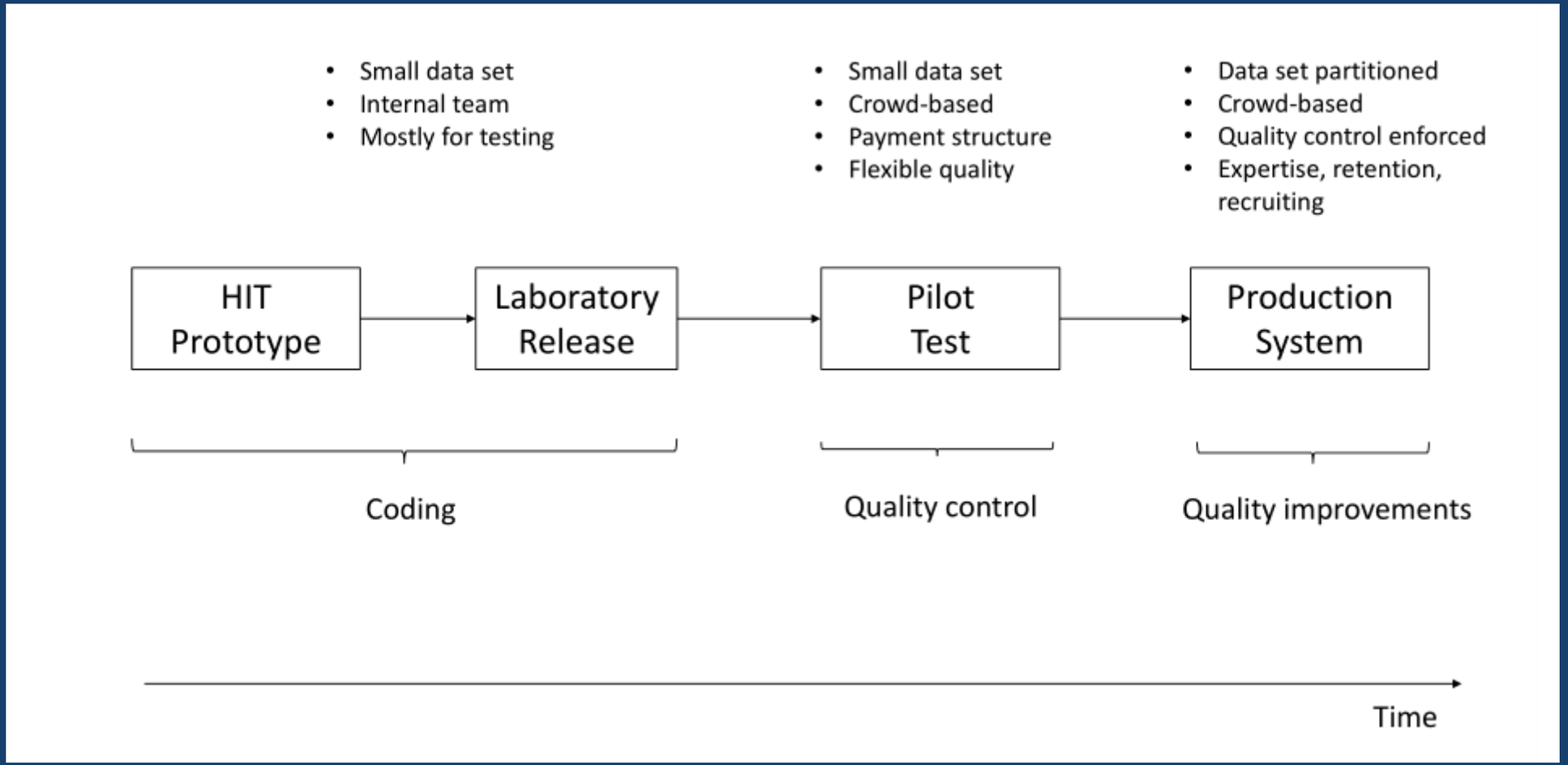
Early stage production

Small data set
Crowd-based
Calibration

Continuous production

Large data sets (partitioned)
Crowd-based
Enforced quality control

Incremental development process



The programming side

Coding

Patterns

Modularization

Testing and debugging

Maintenance

Monitoring

Program structure

Design HITs that humans can do well

Think in terms of data pipelines and workflows

Combine humans and machines

Design patterns

Iterative refinement

Find-fix-verify

Do-verify

Partition-Map-Reduce

C. Callison-Burch. “Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk”, EMNLP 2009.

M. Bernstein et al. “Soylent: A Word Processor with a Crowd Inside”, UIST 2010

Testing and debugging

Testing

Attempt to break a program

Debugging

You know the program is broken

How do we test & debug a HIT?

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

Testing a HIT

Follow same methodology for code

Test input for validity and plausibility

Don’t patch a bad HIT -> rewrite it

Version all templates and metadata

Implementation details

Phase	Recommendation
Coding	One language for extracting data from clusters and compute metrics. Avoid moving data from different tools; encoding, data formats, etc.
Design	Use patterns as much as possible. Examples: iterative refinement, find-fix-verify, do-verify, partition-map-reduce, price-divide-solve. Get ready to throw away HITs and results.
Modularization	Design HITs that humans can do well. Think in terms of pipelines and workflows
Testing and debugging	Don’t patch a bad HIT: rewrite it. Identify problems with data, workers, and task design.
Maintenance	Version all templates and metadata including payment structure.
Monitoring	Dashboard and alerts.
Documentation	Document the essence of the HIT and its mechanics/integration points.

Summary of practices

Don’t ignore established software engineering principles

Eliminate errors early

Be skeptical of one size fits all solutions

Incremental approach for quality control

Data dependencies and technical debt

Data collection and protection

Other practical tips

Sign up as worker and do some work

Eat your own dog food

Address feedback (e.g., poor guidelines, payments, passing grade, etc.)

Everything counts

Systems and data pipelines

Evaluation

IR, NLP, Machine translation, etc

ML models

Beat the machine

Find cases where the machine is wrong

Error framework for predictive models

J. Attenberg, P. Ipeirotis, F. Provost. “Beat the Machine: Challenging Humans to Find a Predictive Model’s ‘Unknown Unknowns’”. ACM JDIQ 2015

Taxonomy creation

Cascade

Combining different HITs

HIT primitives

Generate

SelectBest

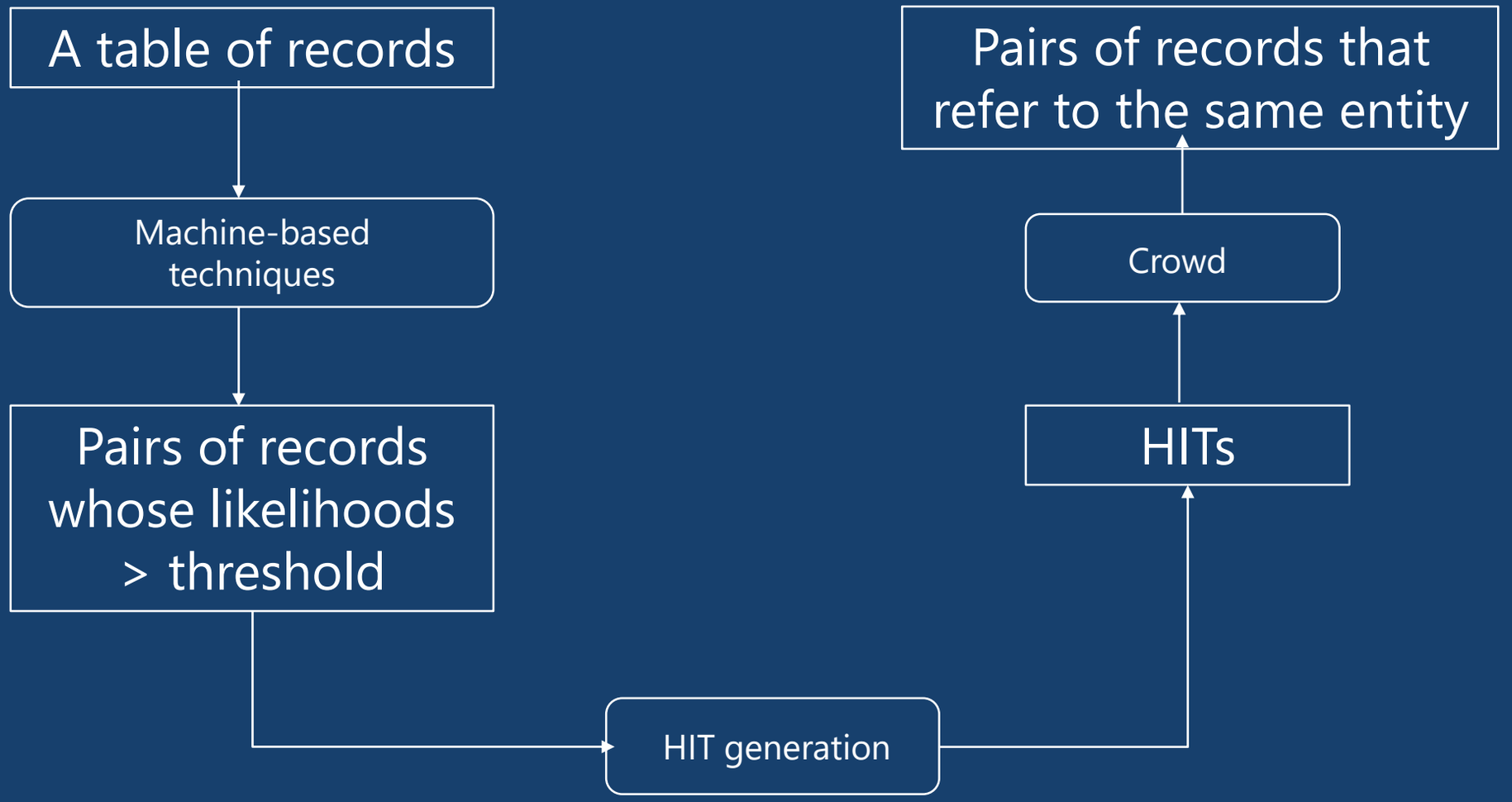
Categorize

Global structure inference

Competitive output in quality and price

L. Chilton, G. Little, D. Edge, D. Weld, J. Landay. “Cascade: Crowdsourcing Taxonomy Creation”. CHI 2013.

CrowdER – Entity resolution



J. Wang, T. Kraska, M. Franklin, J. Feng. CrowdER: Crowdsourcing Entity Resolution. PVLDB 2012

Near-dupes evaluation

User two platforms to generate labels

UHRS (MS internal)

Amazon Mechanical Turk

Two phase approach

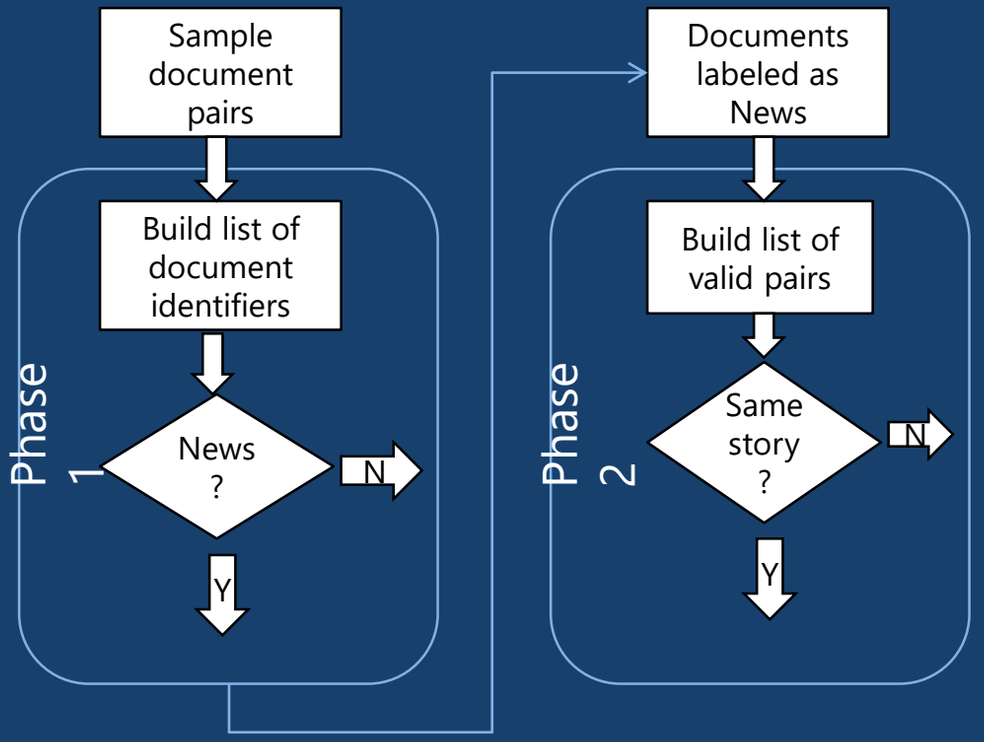
Assess documents in sampled pairs as news/not news

Duplicate or non-duplicate assessment for news pairs

Advantages

Separate workforce

Parallelize



O. Alonso, D. Fetterly, M. Manasse. “Duplicate News Story Detection Revisited”. AIRS 2013.

Near-dupes: design templates

Please read the following document and let us know if the article is a news article in English

The following are **considered** news:

- Generic news articles (e.g., finance, sports, politics, etc.)
- Press releases by a company or institution
- Government data reports (example) or election returns (example)

The following are **not considered** news:

- Weather reports (conditions)
- Headlines with a single or no sentences/paragraph or photo captions (example)
- Event calendars (example)
- Article or blog post that describes another article with a teaser quoting a small amount of content from the original article.
- Library collections (example) or course listings (example)

Politics Home

Obama Sticks to the Script in First Week of Presidency

In the highly scripted first days of Obama's administration, the flurry of activity was intended to show that he was making good on his promise to bring change.

FOXNews. com

Saturday, January 24, 2009

- Photos

President Obama

Does the document above contain a **news** article written in English?

- **Yes**. It is a news article.
- **No**. It is not a news article.
- **I don't know**. I can't tell if the document is a news article.
- **Other**. Web page didn't load/error message/etc.
- **Non English**. This document is not in English.

1. Are these 2 news articles about the same event/topic?

- **Yes**. These news articles are about the same.
- **No**. These news articles are not the same.
- **I don't know**. I can't tell if the news articles are the same or not.
- **Other**. Web page didn't load/error message/etc.
- **Non English**. This document is not in English.

2. Does one document cover more detail than the other?

- Document A covers more detail than document B.
- Document B covers more detail than document A.
- No

Document A

Obama Sticks to the Script in First Week of Presidency

In the highly scripted first days of Obama's administration, the flurry of activity was intended to show that he was making good on his promise to bring change.

FOXNews. com

Saturday, January 24, 2009

- Photos

President Obama

Jan. 21: President Obama places his first round of phone calls to Middle East leaders inside the Oval Office(White House photo by Pete Souza).

Document B

Put a Seed Newsvine link on your own site. Here's how...

- 6 Votes
- 19 Comments
- Print

Obama breaks from Bush, avoids divisive stands

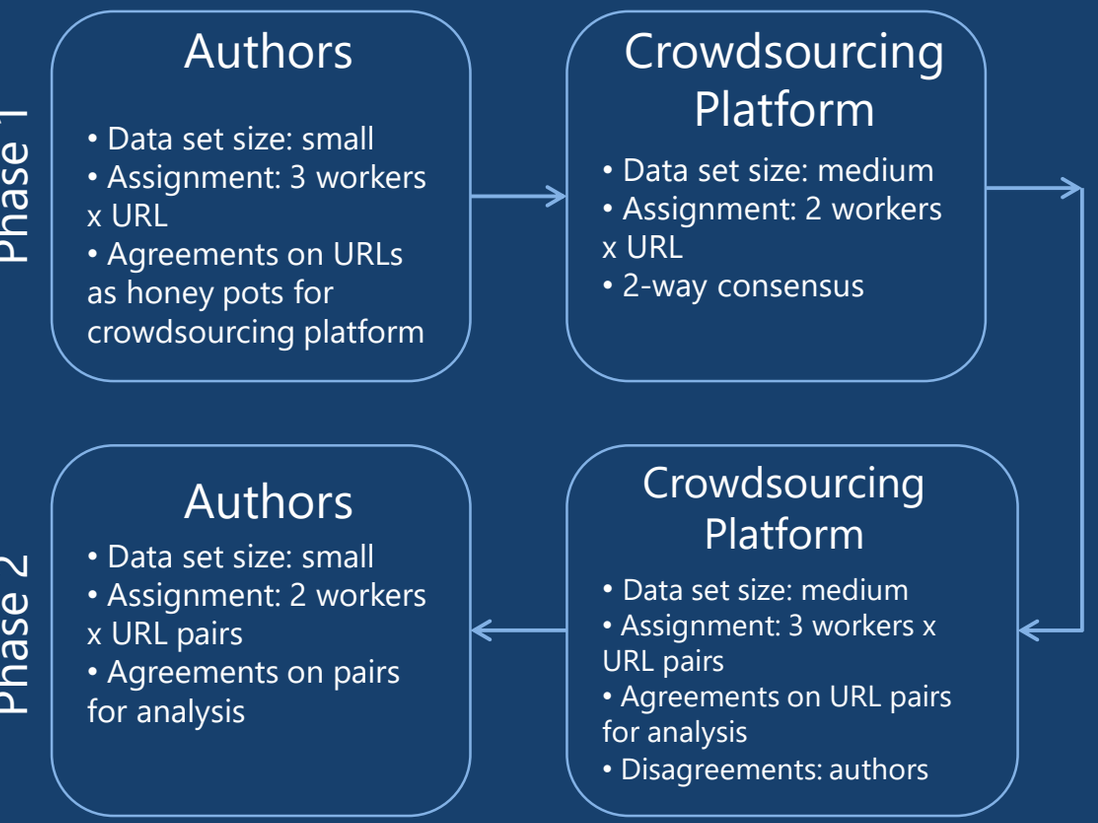
Sat Jan 24, 2009 10:29 AM EST
politics, obama, barack-obama, week, first
Liz Sidoti, Associated Press Writer

WASHINGTON& mdash; Barack Obama opened his presidency by breaking sharply from George W. Bush's unpopular administration, but he mostly avoided divisive partisan and ideological stands. He focused instead on fixing the economy, repairing battered

Source code: http://research.microsoft.com/en-us/downloads/a0200ed2-b0d9-4a9d-9170-bf1b448dcb67/

Near-dupes: quality control

- Use different assessors at each phase
- Generate known good labels for each phase
- Used as honey pots for crowd labels
- Phase 1 documents with consensus as news are used in Phase 2 pairs
- Phase 2 disagreements are resolved by the authors
- Authors used same platform



Information extraction

Higgins

Wernicke

Weighted majority voting

Behavioral features outperform performance-based methods

S. Kondreddi, P. Triantafillou, G. Weikum. “Combining Information Extraction and Human Computation for Crowdsourced Knowledge Acquisition”, ICDE 2014

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

Looking ahead

Machines and humans in sync

Delicate balance but lots of potential

When to use a machine or human for computation

Labels for the machine may not be labels for humans

Debugging code executed by humans

Debugging machine model by humans

Best algorithms for the machine may not be the best choices for humans

Social networks

People are more than Human Processing Units

Our social networks also embody additional knowledge about us, our needs, and the world

The social dimension complements computation

Some research

aaRdvark

CrowdStar

D. Horowitz, S. Kamvar. The anatomy of a large-scale social search engine. WWW 2010

B. Nushi, O. Alonso, M. Hentschel, V. Kandylas. CrowdSTAR: A Social Task Routing Framework for Online Communities. ICWE 2015

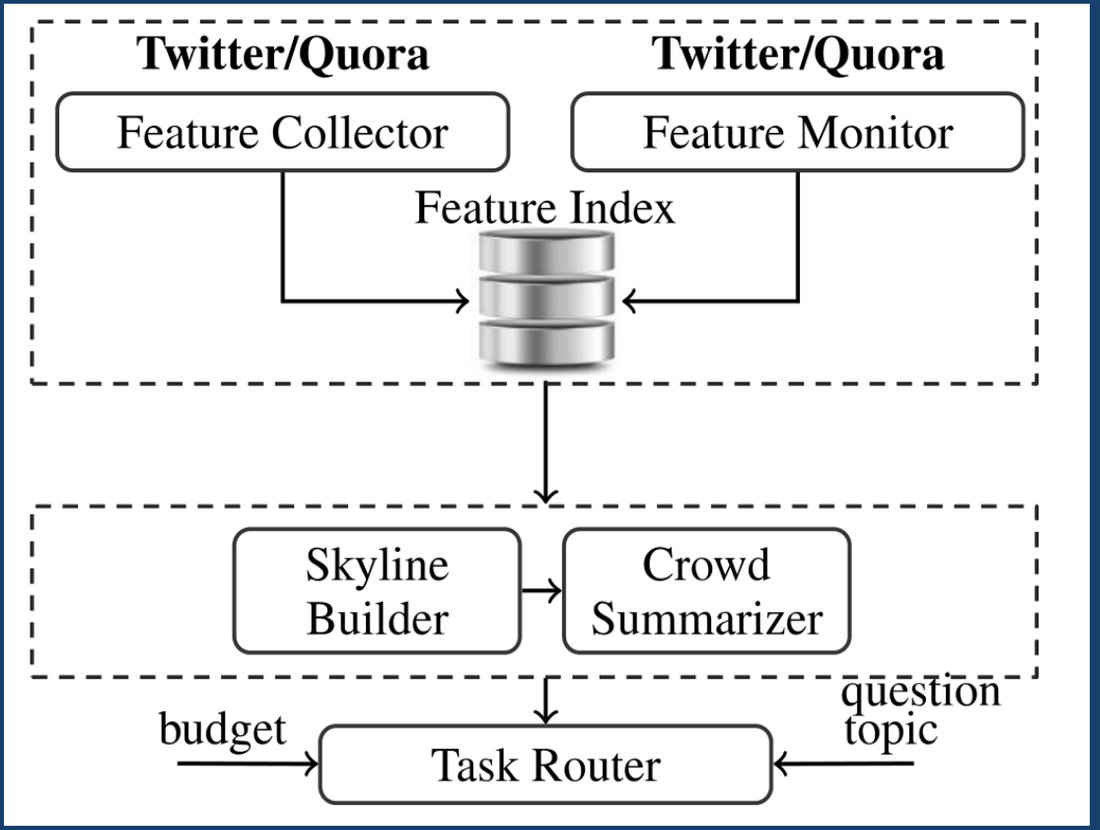
Task routing

Finding competent people in online crowds to route and answer questions

Expertise detection

Availability

Social load balancer



Programming languages

Powerful breakthrough for software engineering

Little research

TurKit

AutoMan

VoxPL

G. Little, L. Chilton, M. Goldman, R. Miller. “TurKit: Tools foriterative tasks on mechanical turk”. In Proc. of HCOMP Workshop, 2009

D. Barowy, E. Berger, D. Goldstein, S. Suri. “VoxPL: Programming with the wisdom of the crowd”, CHI 2017

D. Barowy, C. Curtsinger, E. Berger, A. McGregor. “AutoMan: a platform for integrating human-based and digital computation”. OOPSLA, 2012

Other topics

Real-time and interactivity

Database and crowd-powered algorithms

Fairness and bias

Challenges & opportunities

Lots of potential for systems that combine humans and machines

Current platforms are very rudimentary

Little information on shared practices

Takeaways

Repeatable label quality at scale works but requires a solid framework

Programming principles

Programing machines is hard, programming applications that involves computations by machines + humans is *harder*

Labels for humans != labels for the machine

Takeaways - II

Data management

Biases, configurations

Three aspects that need attention

Workers
Work
Task design

Lots of different skills and expertise required

Social/behavioral science, human factors, algorithms, economics, distributed systems, statistics

HCOMP 2019

www.humancomputation.com

Thank you!

Email: omalonso@microsoft.com
Twitter: @elunca