

# Experiments with TurkKit

**Crowdsourcing and Human Computation  
Lecture 6**

**Instructor: Chris Callison-Burch  
TA: Ellie Pavlick**

**Website: [crowdsourcing-class.org](http://crowdsourcing-class.org)**

stop

Run

Run Repeatedly

will run again in 21 seconds

Reset

input

properties

```
...e improve me..."
5; i++) {
...om HTML page on S3
...e to improve a description for an image
...reateWebpageFromTemplate(<div>
...tps://www.cis.upenn.edu/~ccb/images/reed.jpg" width="500" alt="description not availa
...
...se improve the description for this image.</li>
...yle="width:500px;height:170px" name="newText">{text}</textarea>
...submit" value="Submit"></input>
...
...on MTurk using the webpage
...k.createHIT({
...rove text about an image",
...rove a small paragraph describing an image.",
...e,
...,
...5
...
...HIT to be done,
...rogram will probably stop here
...e it is executed,
...will not have had
...to complete the HIT)
...waitForHIT(hitId)
...
...text from the hit, and display it
...t.assignments[0].answer.newText
...-----")
...-----")
...
...highlighting the differences between the texts
...ightDiff(text, newText)
...
...HTML page on S3
...e to vote between the two versions
...reateWebpageFromTemplate((" + <div>
...="https://www.cis.upenn.edu/~ccb/images/reed.jpg" width="500" alt="description not av
...
...Please choose the better description for this image.</li>
...
...class="random">
...
...<td><input name="voteA" type="submit" value="&gt;" style="width:30px;height:50px"></i
...
...<td><pre style="width:500px;border:thin solid; white-space: pre-wrap; white-space: -m
...ace: -o-pre-wrap;">__DIFF_A__</pre></td></tr>
...
...class="random">
```

output

```
vote = keep
disabled HIT: 2CF2Q4UUKTQAM5SN08GM0KBTW
deleted S3 object at: http://s3.amazonaws.com/AKIAIESZ7MDFJWECM4MA.TurKit/frjxl
deleted S3 object at: http://s3.amazonaws.com/AKIAIESZ7MDFJWECM4MA.TurKit/l76pm
S3 object put at: http://s3.amazonaws.com/AKIAIESZ7MDFJWECM4MA.TurKit/xy0dvqq6t
created HIT: 2OKJENLVWJ5RRN6E51Q86MWBTD9S02
url: https://www.mturk.com/mturk/preview?groupId=2B526DG67D10576VUL3OS2
hit completed: 2OKJENLVWJ5RRN6E51Q86MWBTD9S02
-----
Adorable baby with deep blue eyes, wearing light blue and white elephant pajama
Baby Cool Looking and smooth skin,very bright eyes,attractive dressing wearing
-----
S3 object put at: http://s3.amazonaws.com/AKIAIESZ7MDFJWECM4MA.TurKit/m5ne0slmq
created HIT: 2KZVWAVKI62QL28HHNQ9HQH6XJ5QI6
url: https://www.mturk.com/mturk/preview?groupId=2AE0XULRGNTELF2NIWFX7
stopped
```

HITs / S3

database

HITs:

1 <https://www.mturk.com/mturk/preview?groupId=2B526DG67D10576VUL3OS2JE1XL8FT>

1 <https://www.mturk.com/mturk/preview?groupId=2AE0XULRGNTELF2NIWFX7N62VV>

S3 Objects:

<http://s3.amazonaws.com/AKIAIESZ7MDFJWECM4MA.TurKit/xy0dvqq6t3t1qh2ye2x87eydae>

<http://s3.amazonaws.com/AKIAIESZ7MDFJWECM4MA.TurKit/m5ne0slmqflebrt2l55w43vhcx>

TurKit in action



Adorable baby with deep blue eyes, wearing light blue and white elephant pajamas and a floppy blue hat.

Baby Cool Looking and smooth skin, very bright eyes, attractive dressing wearing light blue and white elephant pajamas and a floppy blue hat. Overall impression very sweet and also funny.



# What are the basic units of collecting work?

- Human computation is a new field
- Writing algorithms that involve people as function calls is relatively unexplored
- How can we characterize the types of work that we can do, or the processes that yield the best results?

# Iterative v. Parallel Processing

- Basic distinction in the workflow
- Should crowd workers do tasks independently in parallel?
- Or should they work together in an iterative fashion and build off of each other's work?

# Tradeoffs

- Iterative process shows each worker the results from previous workers
  - Must collect contributions serially
- Parallel processes asks each worker to solve a problem alone
  - no workers depend on the results of other workers, so can be parallelized

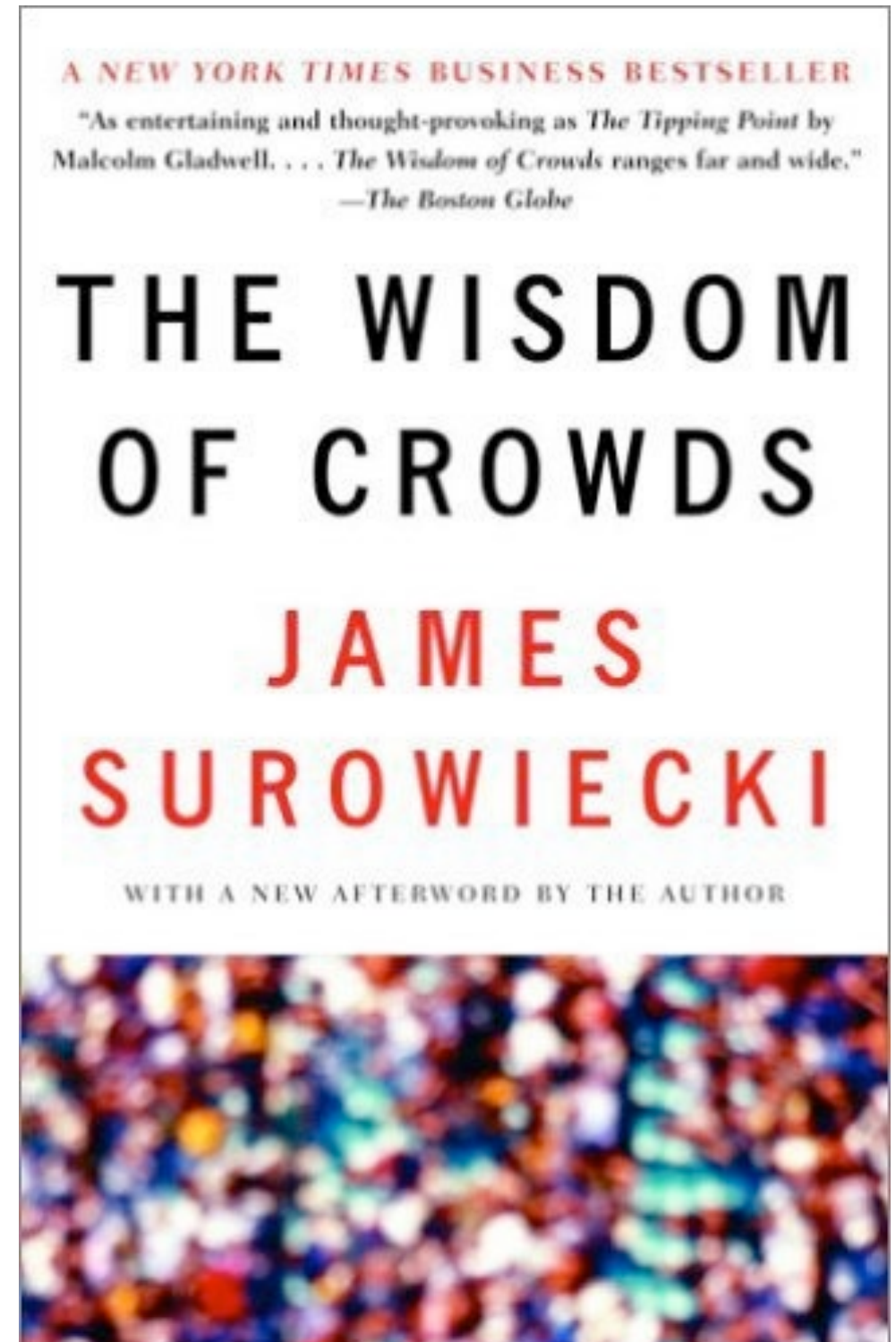
# Threadless v. Wikipedia

- One person starts an article, and then other people iteratively improve it by looking at what people did before them and adding information, correcting grammar, creating a consistent style, etc.
- t-shirts are created in parallel. People submit ideas independently, and then others vote to determine the best ideas that will be printed.

# Wisdom of Crowds

Requirements for a crowd to be wise

- Diversity of Opinion
- Independence
- De-centralization
- Aggregation





# Wisdom of Crowds: Independence

Surowiecki argues that aggregating answers from a decentralized, disorganized group of people, all thinking independently yields more accurate answers than from individuals.

Individual errors need to be uniformly distributed, so individual judgments must be made independently.

# Does this hold empirically on MTurk?

- Greg Little, Lydia Chilton, Max Goldman, and Rob Miller verify it through a set of experiments
- Exploring tradeoffs between iterative v. parallel processing in writing, brainstorming, and transcription.

# Writing



# Transcription

transcription is the process of copying a segment of DNA into RNA

					.			
--	--	--	--	--	---	--	--	--

transcription is the process of copying a segment of DNA into RNA

		.		.					
--	--	---	--	---	--	--	--	--	--

transcription is the process of copying a segment of DNA into RNA

							.				.
--	--	--	--	--	--	--	---	--	--	--	---

# Brainstorming

- Our company sells headphones. There are many types and styles available. They are useful in different circumstances. Our site helps users assess their needs, and get the pair of headphones that is right for them.
- Please suggest 5 new company names for this company.



# Higher level goals

- Establish models and design patterns for human computation processes
- Figure out how best to coordinate small contributions from many people to achieve larger goal
- Focus is on *aggregation* dimension from taxonomy of human computation

# Model

dependently  
(iteratively)

independently  
(in parallel)

creation tasks

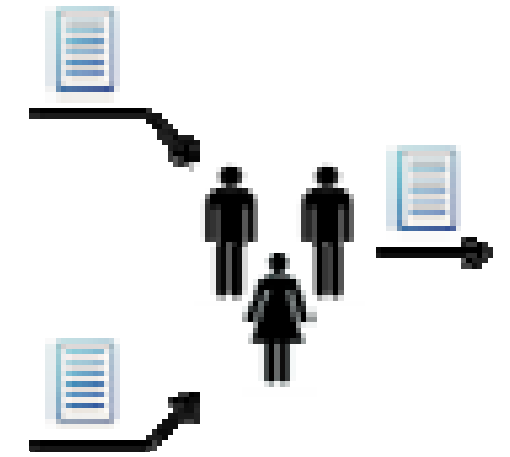
decision tasks


# Creation tasks



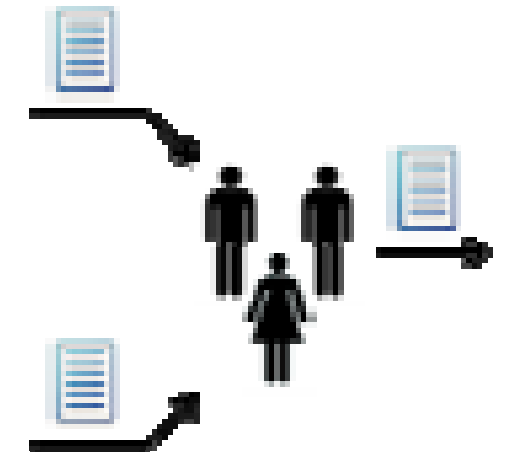
- Goal is to produce new high quality content
- Example creation tasks: writing, ideas, imagery, solutions
- Few constraints on worker inputs to the system
- Computer doesn't understand the input

# Decision tasks



- Decision tasks solicit opinions about existing content
- Example: choose between two descriptions of the same image
- User input is constrained because the computer has to interpret the responses

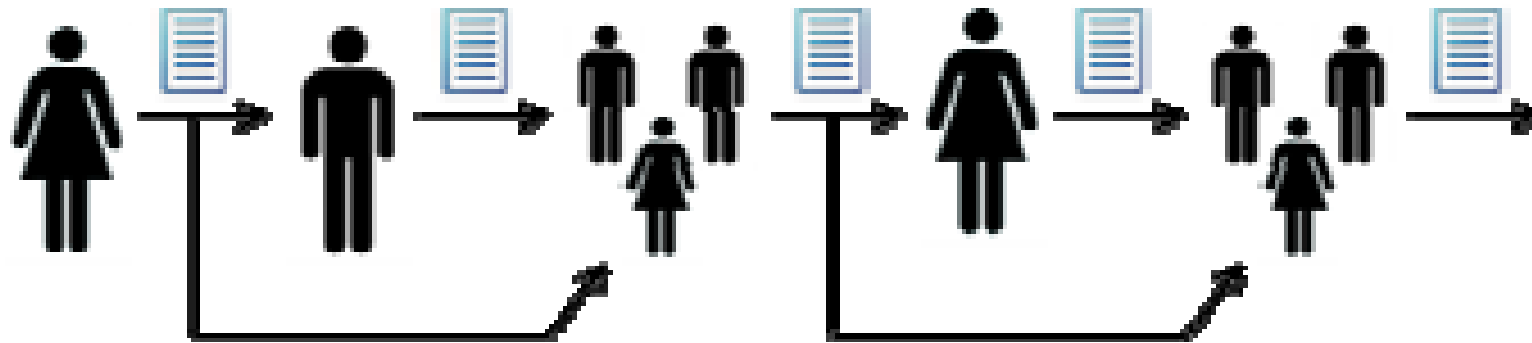
# Decision tasks



- Goal of decision tasks is to solicit *accurate* responses
- Solicit multiple responses and aggregate them
- Mechanisms:
  - **comparisons**: is image description A better than image description B?
  - **ratings**: Rate the quality of this description on a scale from 1-10



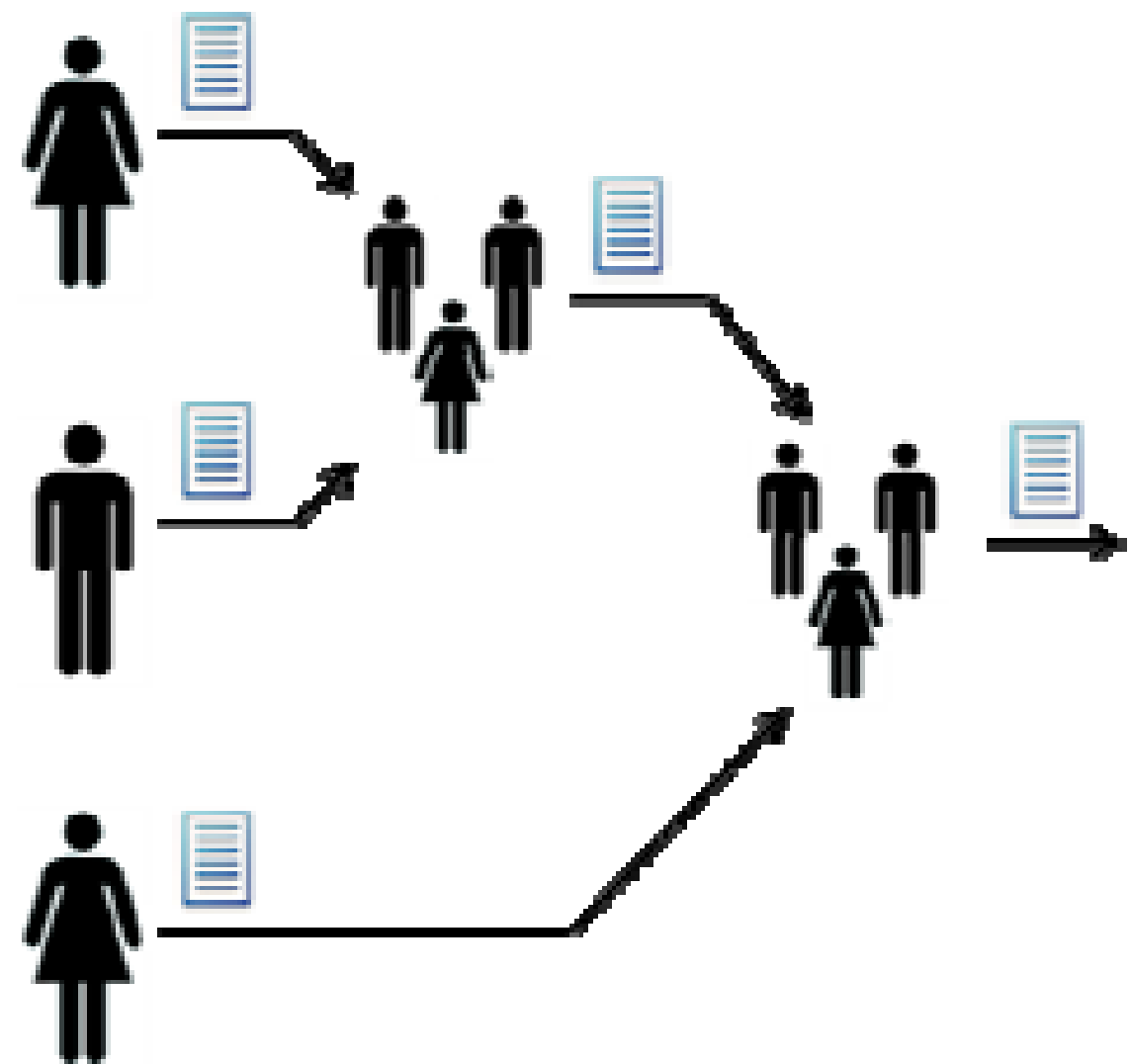
# Pattern #1: Iterative Combination



- Workers are shown the content generated by previous workers
- Computer optionally tracks the best content, shows it or all previous content

# Pattern #2: Parallel Creation

- Creation tasks are executed in parallel
- Workers do not see each others outputs
- Outputs can be compared via decision tasks, as before
- May be difficult to merge content



# Experiments

- Little, Chilton, Goldman, and Miller performed 3 experiments on MTurk to compare iterative v. parallel patterns
- Writing image descriptions
- Transcribing obscured texts
- Brainstorming company names

# Image description experimental setup

- Selected 30 engaging images from <http://www.publicdomainpictures.net>
- Each image went through 6 creation tasks, and 5 comparison tasks (with 5 people voting on the comparisons)
- Run on MTurk. Paid \$0.02 for creation, and \$0.01 for comparison.

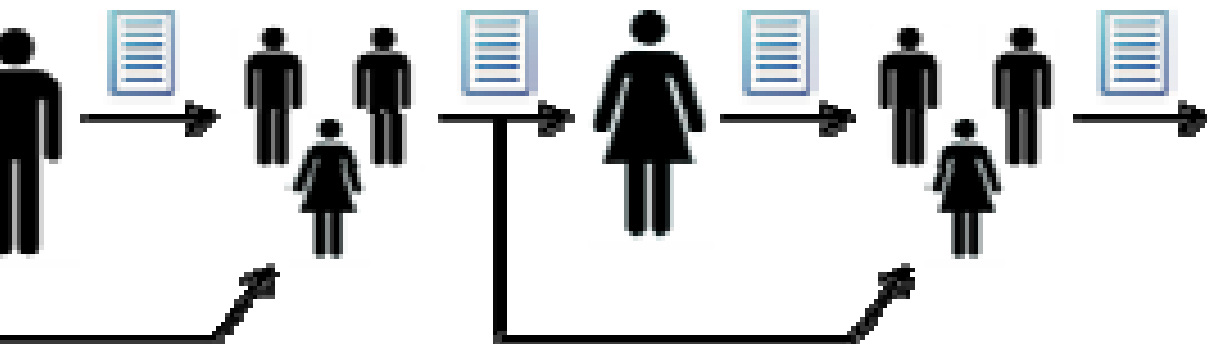


- Please describe the text factually
- (You may use the provided text as a starting point, or delete it and start over)
- Use no more than 500 characters

Lightening strike in a blue sky near a tree and a building.

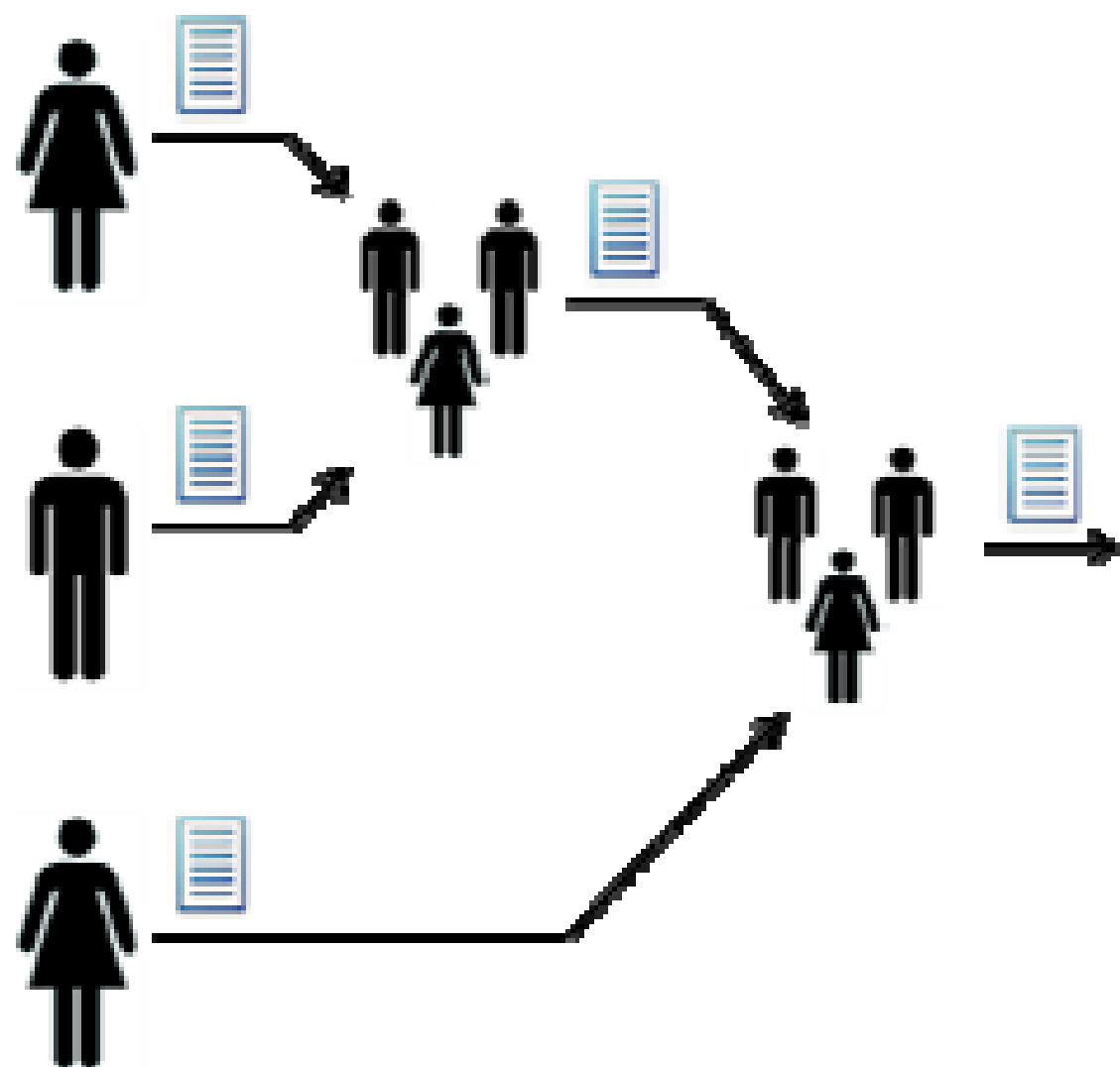


- Iteration 1: Lightening strike in a blue sky near a tree and a building.
- Iteration 2: The image depicts a strike of fork lightening, striking a blue sky over a silhouetted building and trees. (4/5 votes)
- Iteration 3: The image depicts a strike of fork lightning, against a blue sky with a few white clouds over a silhouetted building and trees. (5/5 votes)
- Iteration 4: ~~The image depicts a strike of fork lightning, against a blue sky wonderful capture of the nature.~~ (1/5 votes)
- Iteration 5: This image shows a large white strike of lightning coming down from a blue sky with the tops of the trees and rooftop peaking from the bottom. (3/5 votes)
- Iteration 6: This image shows a large white strike of lightning coming down from a blue sky with the silhouettes of tops of the trees and rooftop peeking from the bottom. The sky is a dark blue and the lightening is a contrasting bright white. The lightening has many arms of electricity coming off of it. (4/5 votes)



This image shows a large white strike of lightning coming down from a blue sky with the silhouettes of tops of the trees and rooftop peeking from the bottom. The sky is a dark blue and the lightening is a contrasting bright white. The lightening has many arms of electricity coming off of it.

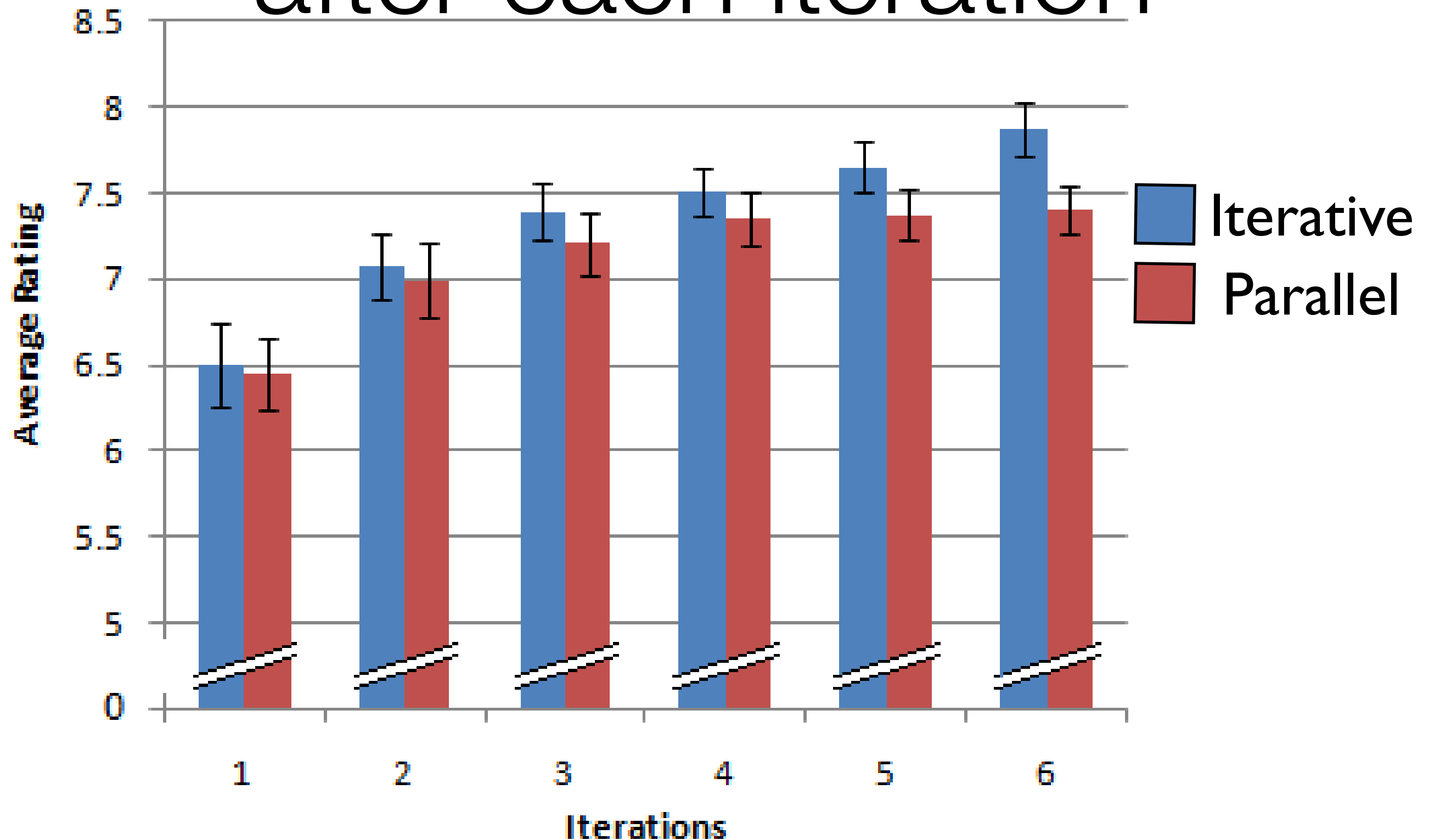
**Average Rating: 8.7**



White lightning n a root-like formation shown against a slightly wispy clouded, blue sky, flashing from top to bottom. Bottom fifth of image shows silhouette of trees and a building.

**Average Rating: 7.2**

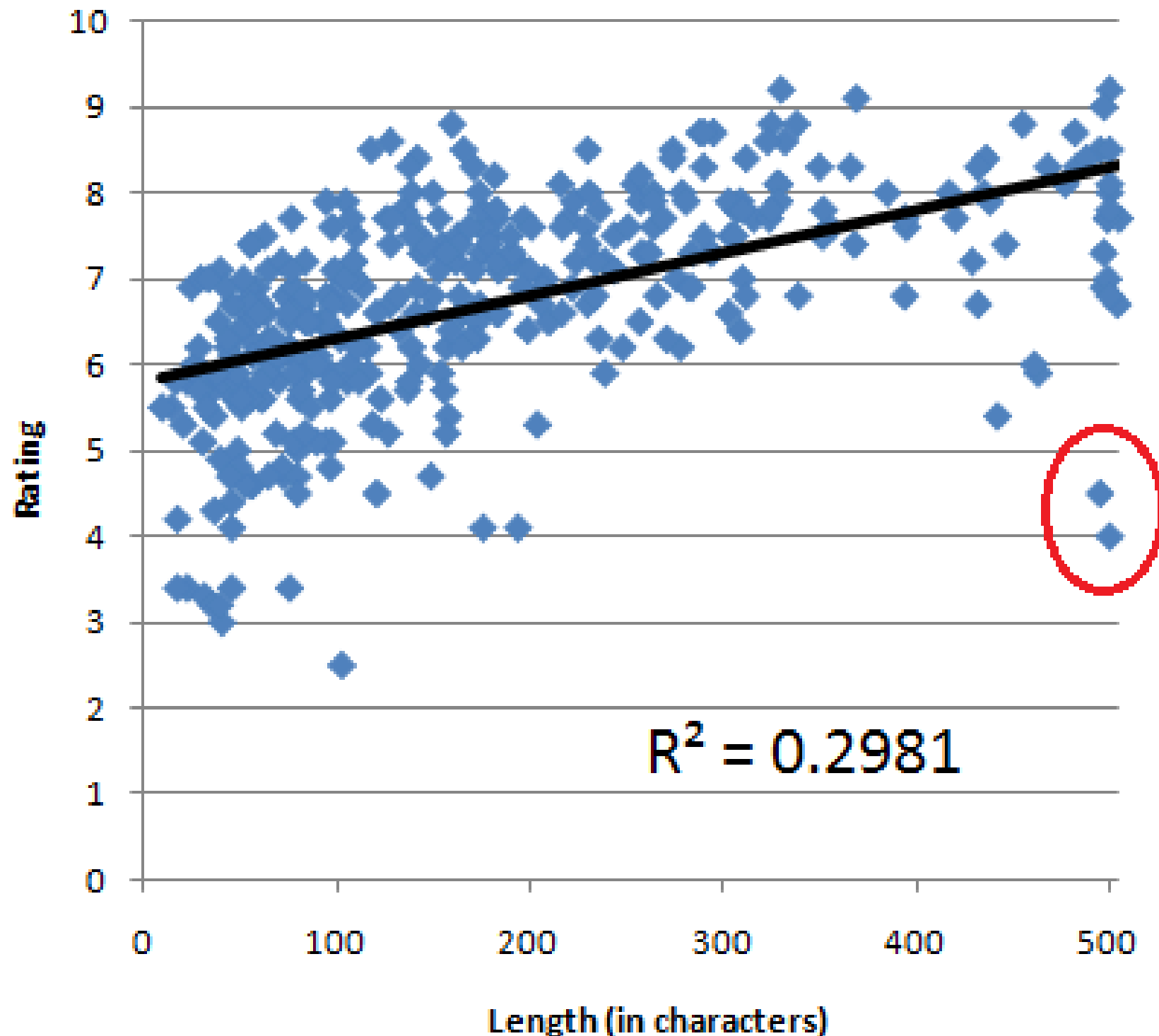
# Relative improvements after each iteration



# What do Workers do at each iteration

- **31%** mainly append content at the end, make only minor modifications (if any) to existing content
- **27%** modify/expand existing content, but it is evident that they use the provided description as a basis
- **17%** seem to ignore the provided description entirely and start over
- **13%** mostly trim or remove content
- **11%** make very small changes (adding a word, fixing a misspelling)

# Correlation with description length and rating





# Experiment 2:

## Brainstorming Names

- Presented descriptions of 6 fictional companies
- Asked Turkers to list 5 names each
- Iteration had 6 tasks for each company, Turkers are shown the names so far
- Parallel had 6 independent Turkers for each company

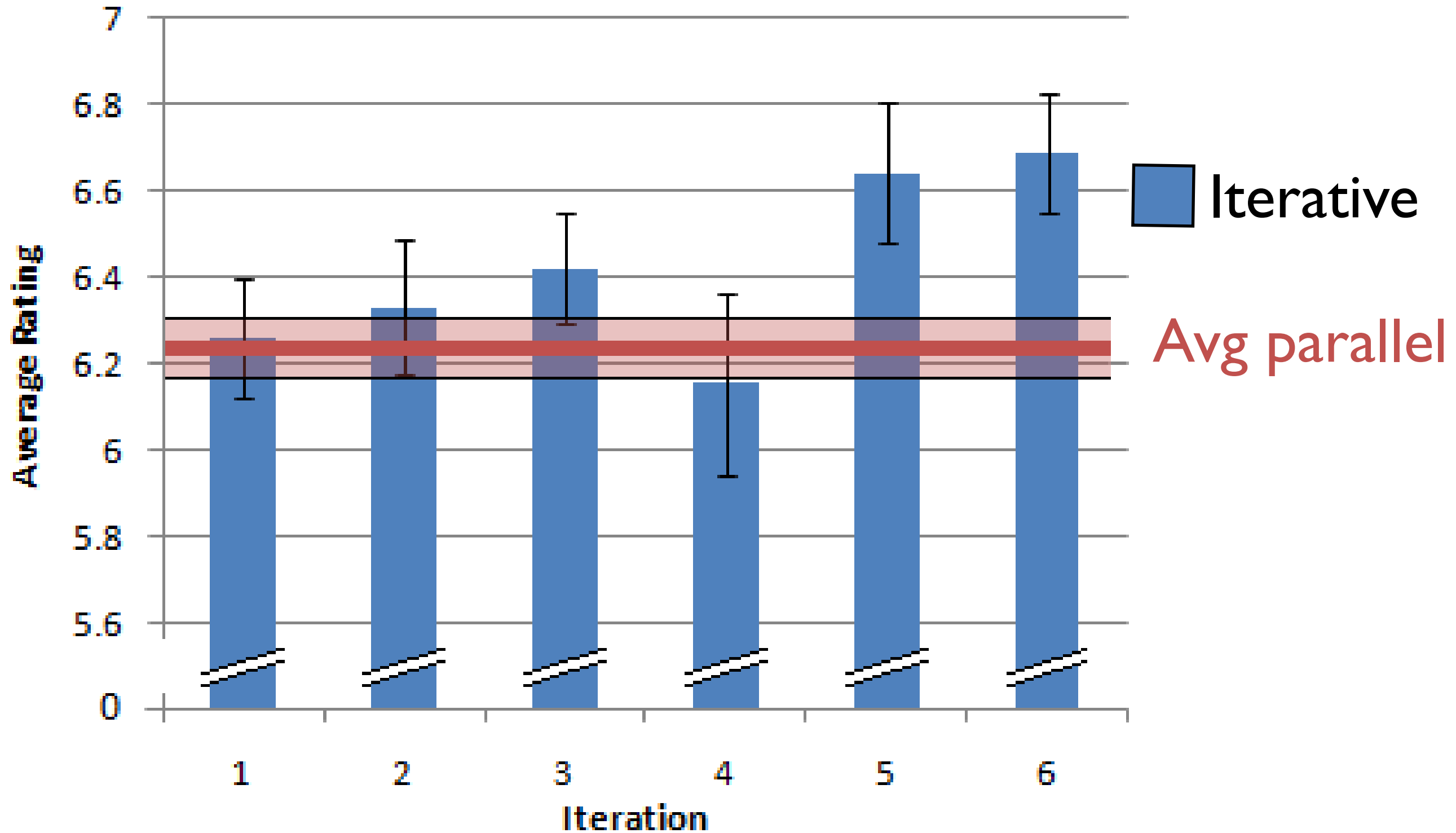
# Brainstorming

- Our company sells headphones. There are many types and styles available. They are useful in different circumstances. Our site helps users assess their needs, and get the pair of headphones that is right for them.
- Please suggest 5 new company names for this company.

# Example names

<b>Iterative</b>		<b>Parallel</b>	
Easy on the Ears	7.3	music brain	8.3
Easy Listening	7.1	Headphone House	7.4
Music Explorer	7.1	Headshop	7
Right Choice Headphone	7.1	Talkie	6.8
...		...	
Least noisy hearer	5.1	company sell	4.3
Headphony	4.9	head phones r us	4.2
Shop Headphone	4.8	different circumstances	3.7

# Iterative improvements



# Getting the best name

- Iteration seems to increase the average rating of new names
- Not clear that iteration is the right choice for generating the best rated names
- Iterative process has a lower variance: 0.68 compared with 0.9 for the parallel process
- Showing turkers suggestions may cause them to riff on the best ideas they see, but makes them unlikely to think too far afield from those ideas

# Experiment 3: Blurry text recognition

- Human OCR, inspired by reCAPTCHA
- “We considered other puzzle possibilities, but were concerned that they might be too fun”
- 16 creation task in both iterative and parallel processing

# Blurry Text Transcription

**Abstract**

[illegible][illegible][illegible][illegible][illegible]

# Choosing the best result

- If a particular word is guessed a plurality of times, then choose it
- Otherwise pick at random from the words that tied for best



- Please transcribe as many words as you can.
- Put a \* in front of words you are unsure about.

It is important to be sure you , are in the

If a \*festival . \*two \*me . \*but \*is

you to . I am sure to have a

if . \*two If

, are in the to be sure .

\*festival . \*festival

Submit

- Please transcribe as many words as you can.
- Put a \* in front of words you are unsure about.

TV is supposed to be bad for you, but I am watching some TV shows. I think some TV shows are really entertaining, and I think it is good to be watched.

If a \*festival . \*two \*me . \*but \*is

If . \*two If

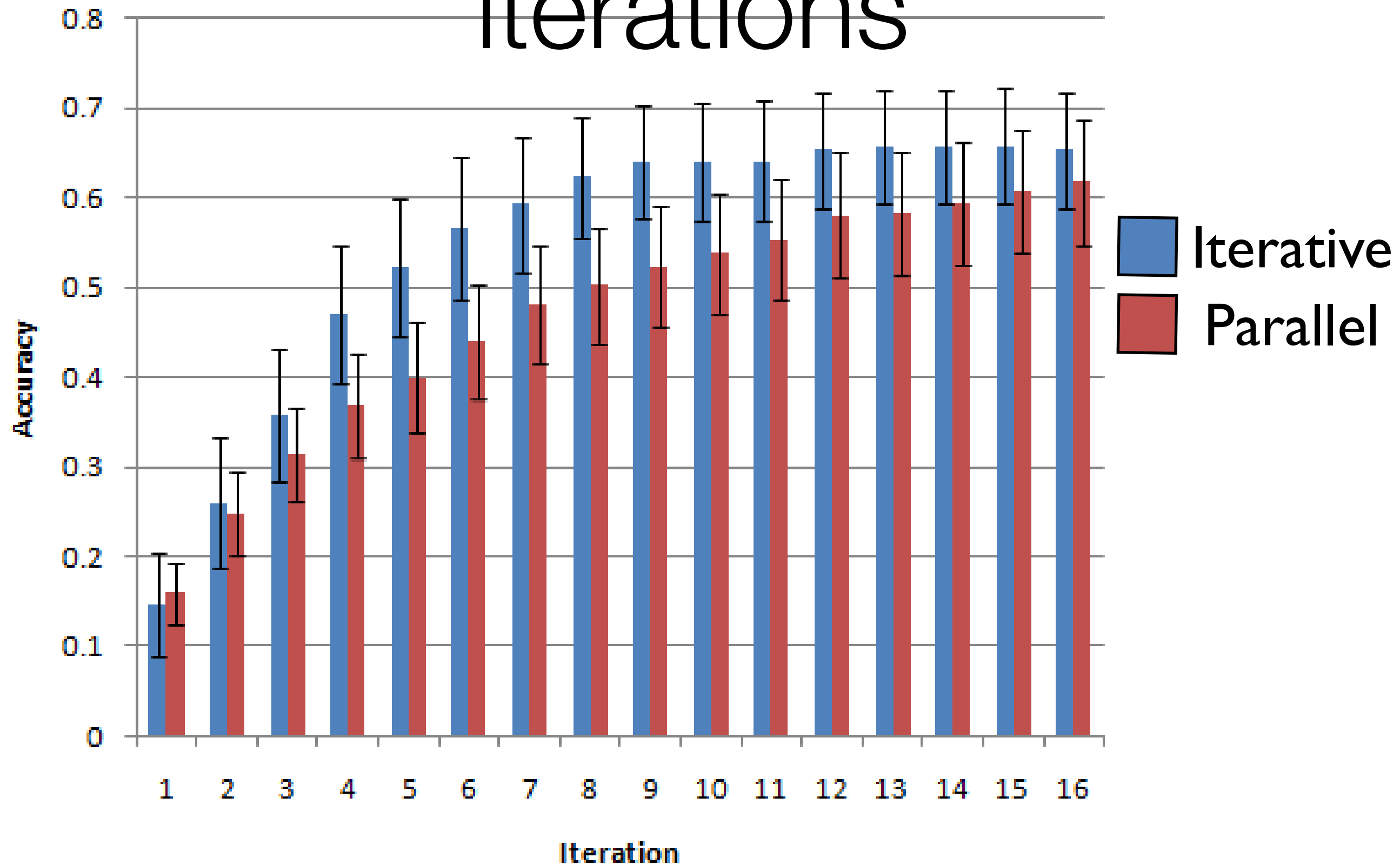
\*festival . \*festival

Submit

**Iterative:** TV is supposed to be bad for you, but I ~~am~~ watching some TV shows. I think some TV shows are really entertaining, and I think it is good to be ~~watched~~.  
(94% correct)

**Parallel:** TV is supposed to be bad for you, but I like watching some TV shows. I think some TV shows are really ~~advertising~~, and I think it is good to be entertained. (97% correct)

# Accuracy after several iterations



# Sometimes poor initial guesses cause problems

- **8th iteration:** “Please do ~~ask~~ \*anything ~~you need~~ \*~~me~~. Everything is ~~going fine, there~~ \* \*, ~~show me then~~ \* \* anything you ~~desire~~.”
- **16th iteration:** “Please do ~~ask~~ \*~~about~~ anything ~~you need~~ \*~~me~~. Everything is going fine, there \*were \* , show me then \*~~bring~~ \* anything you ~~desire~~.”
- Several of the workers doing the task in the parallel condition got it 100% correct

# Discussion

- What do these results tell us about iterative versus parallel processing in human computation?
- Are the experiments well formulated?
- Is James Surowiecki right?

# Tradeoff between Average and Best

- The brainstorming task showed tradeoff between increasing the average quality v. increasing the chance of finding the best
- Showing previous work increased quality, but decreased variance

# Leading people astray

- The blurry text task showed initially bad guesses can lead to poorer quality later
- Suggests that a hybrid approach may be better: start multiple iterative jobs in parallel

# Future Work



# Recap: Model

dependently  
(iteratively)

independently  
(in parallel)

creation tasks

decision tasks


# What factors affect Creation Tasks?

- How much does the reward affect quality?
- How much work is expected? Is it better to break the task down into smaller pieces?
- Are examples are shown? Is prior work shown?

# What factors affect Decision Tasks?

- Goal is to determine the best items in a set
- What's the best way to achieve this?
  - Absolute ratings?
  - Pair-wise comparisons?
  - Sorting multiple items in a single task?

# New building blocks

- What other building blocks exist?
- What paradigms and metaphors should we use to think about human computation?

# Office Hours

Written assignment #2:  
Write a company profile

# Things to address in profile

- What does the company do? How does this company use crowdsourcing?
- How do users contribute? Could you participate?
- Use the analytic framework that we developed earlier (Motivation? Human Skill? Aggregation? Quality Control?)
- Does it involve a computer algorithm in addition to human computation?
- What is the size of its market?

Example company  
presentation from Ellie