UNIVERSITY OF WASHINGTON

Decision-Theoretic Crowdsourcing

Mausam
Associate Professor
Computer Science and Engineering
Indian Institute of Technology, Delhi

(Joint Work with Jonathan Bragg, Peng Dai,
Andrey Kolobov, Chris Lin and Dan Weld)

30,000 ft View

- Crowdsourcing is huge & growing rapidly
  - Exciting application area for AI
- Artificial Intelligence has incredible potential
  - Reduces costs/errors by 30-85%
  - … and this is just the tip of the iceberg

Outline

- Motivation for Crowdsourcing
- Crowdsourcing Challenges: design of Clowder
- Control of Crowdsourced Workflows
  - Simple consensus multiple-choice tasks
  - Simple consensus tasks with infinite answer choices
  - Tasks with complex workflows
  - Tasks with multiple workflows
  - Tasks with multi-class categorization
- Future Work

Crowdsourcing Challenges

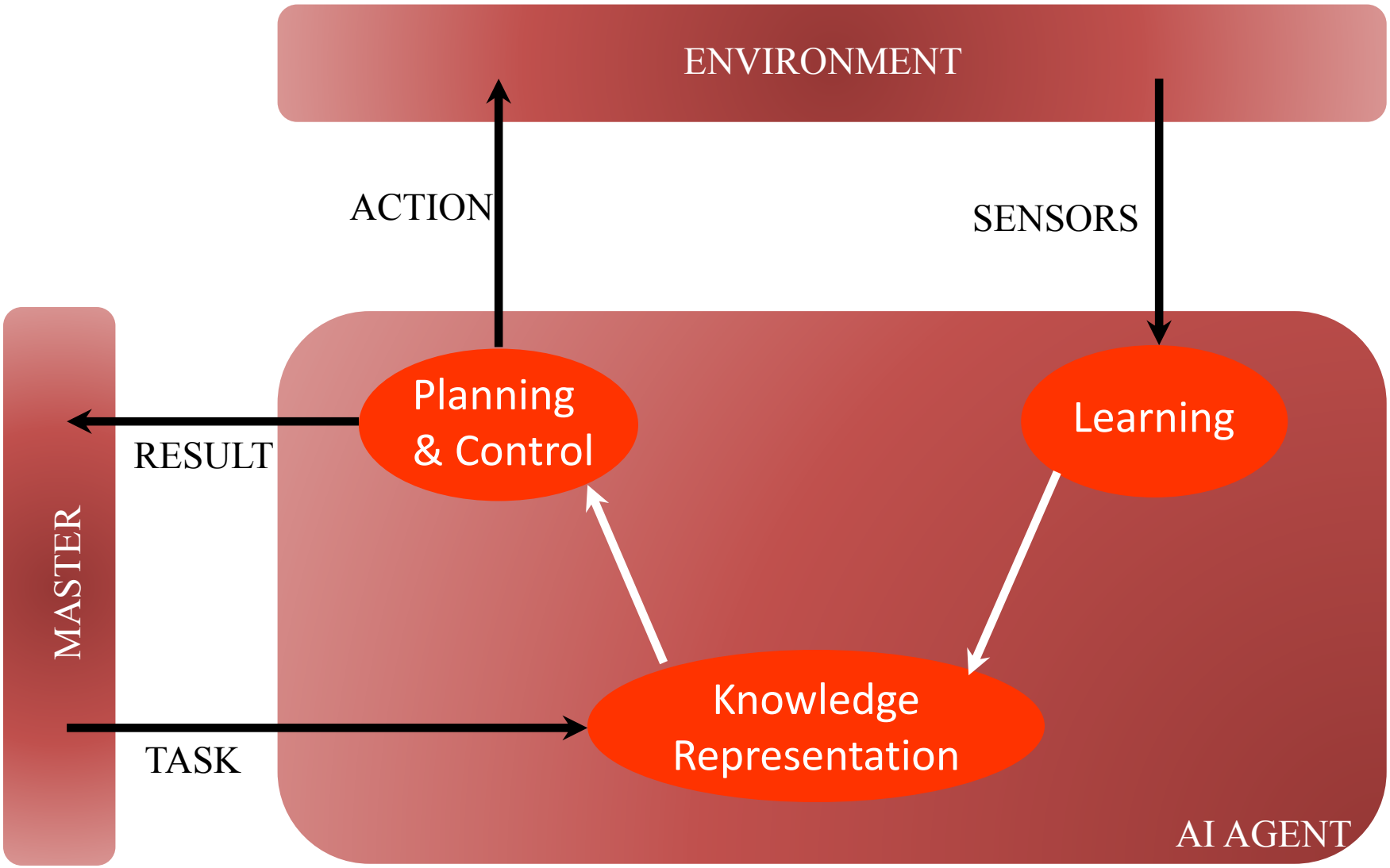
- High variance in worker quality
  - how to track quality of output?
- Complex task → small work-unit workflows
  - support for workflow testing & optimization?
- Requesters need strong computational skills
  - create general-purpose tools?
- Requester-Worker matching
  - auto-learn worker strengths/weaknesses?
- A great application area for AI techniques!

Our Vision

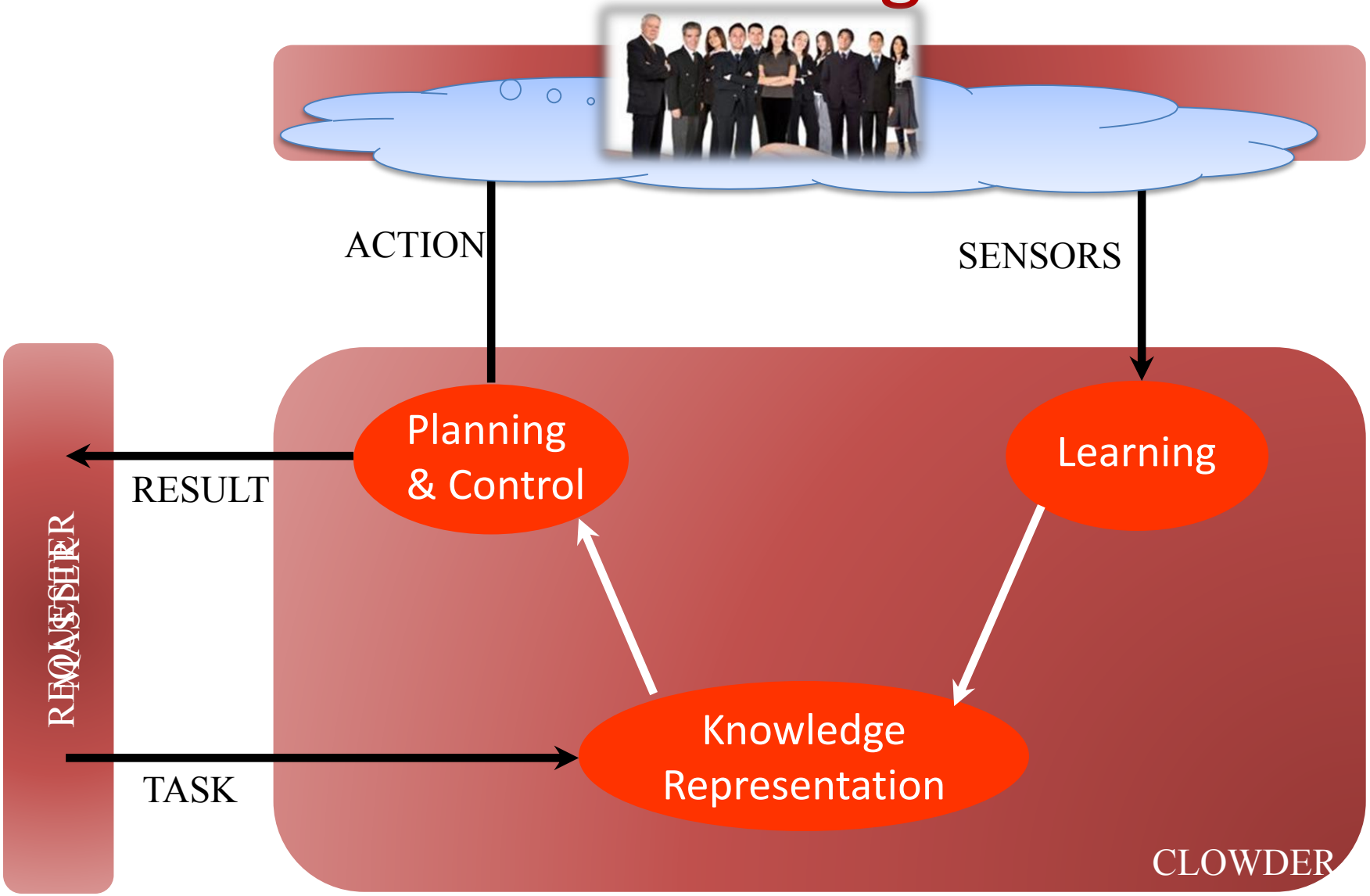
[Weld, Mausam, Dai HCOMP'11]

- Explore & demonstrate value of AI to crowdsourcing
- A unified crowdsourcing tool to aid requesters & manage crowd
  - Declarative language to specify workflows
  - Shared models for common tasks
  - Integrated modeling of workers
  - Comprehensive decision-theoretic control
- Clowder: an integrated AI agent

A General AI Agent



The Clowder Agent



The Clowder Agent

Jobs

Work

- Which workflow to select? low Parameters
- Which job to post? r Abilities/Skills
- When to submit? ifficulties

Output

REQUESTER

Workflow(s)
Tasks

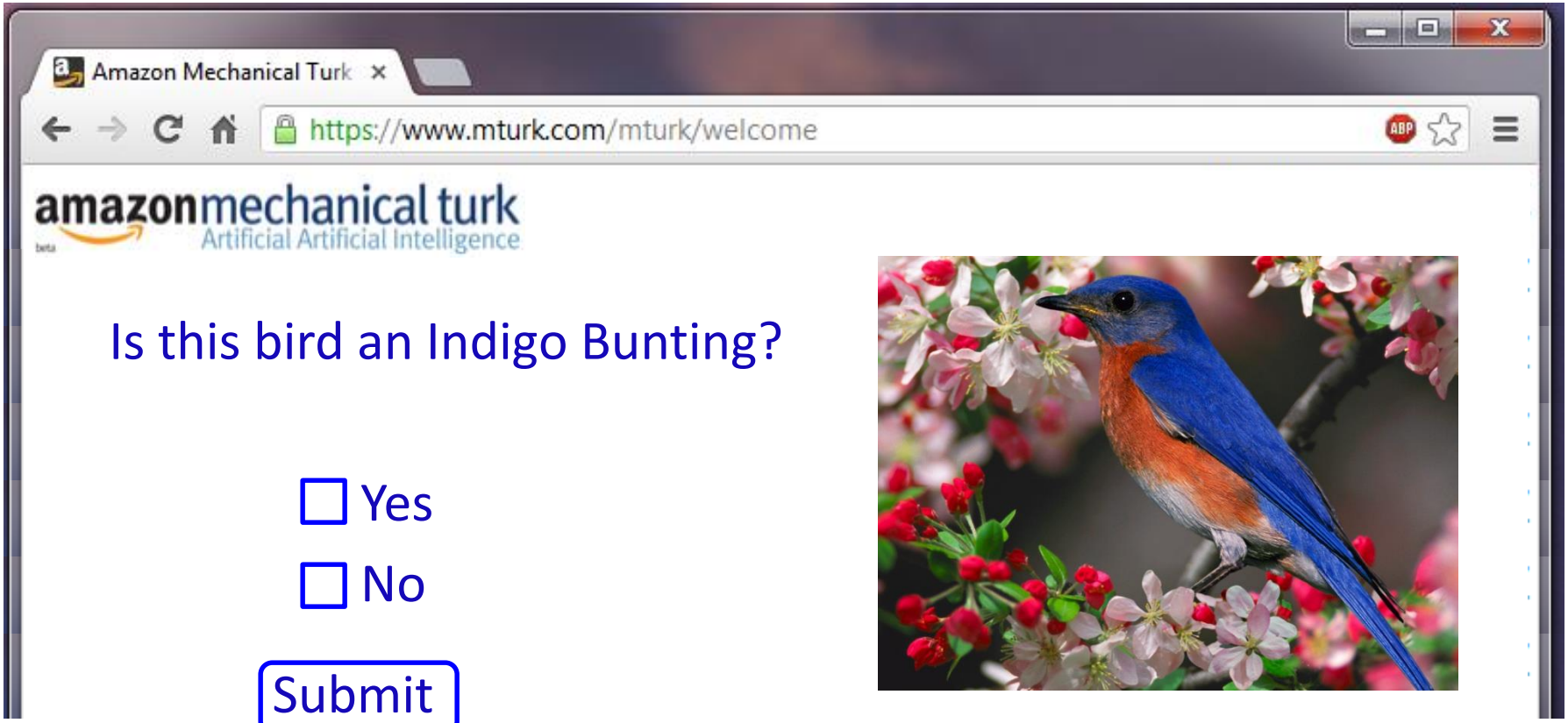
- A language to specify workflows
- Knowledge about diff. workers
- Knowledge about different tasks

CLOWDER

Outline

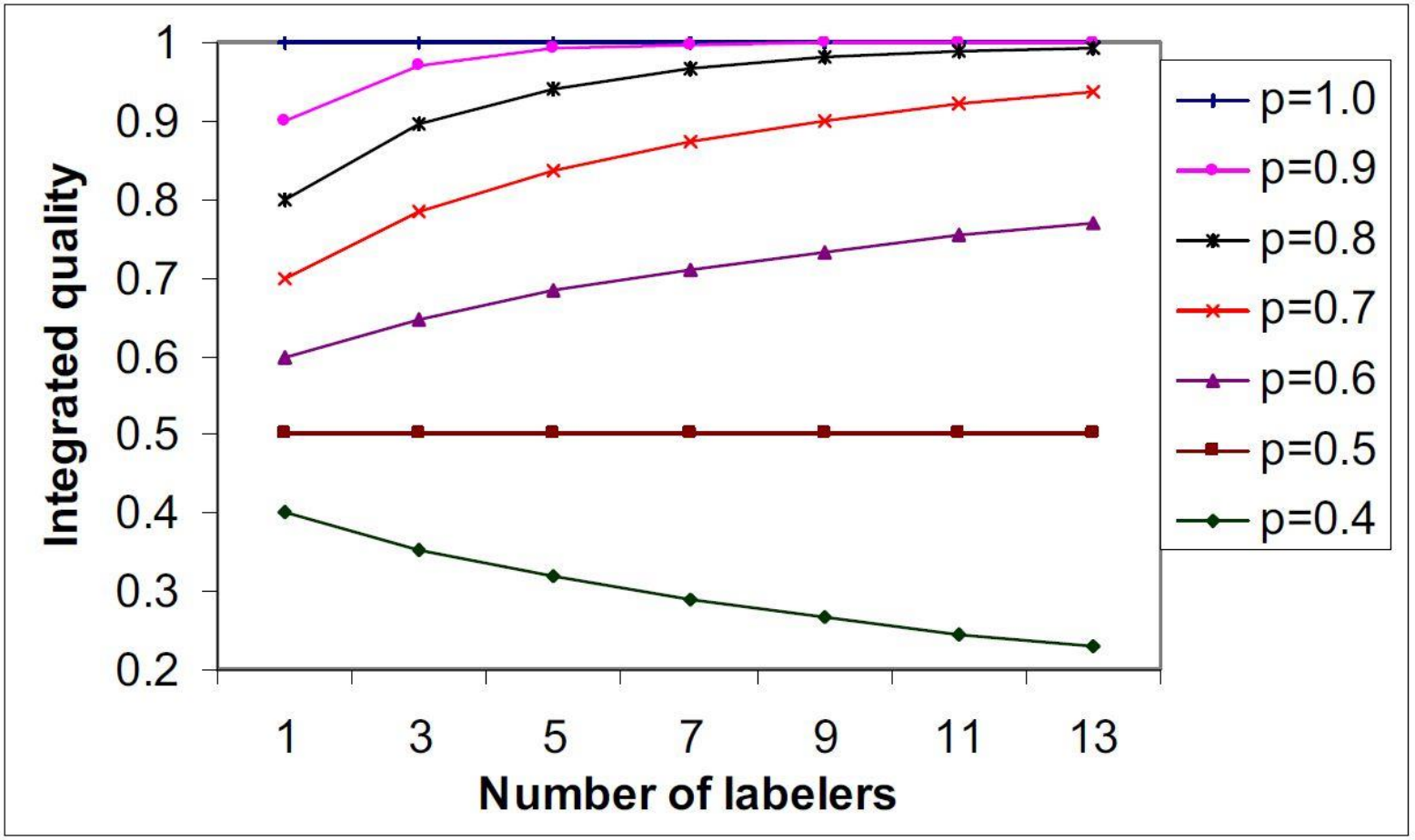
- Motivation for Crowdsourcing
- Crowdsourcing Challenges: design of Clowder
- Control of Crowdsourced Workflows
  - Simple consensus multiple-choice tasks
  - Simple consensus tasks with infinite answer choices
  - Tasks with complex workflows
  - Tasks with multiple workflows
  - Tasks with multi-class categorization
- Future Work

Multiple Choice/Consensus Tasks



Majority Voting

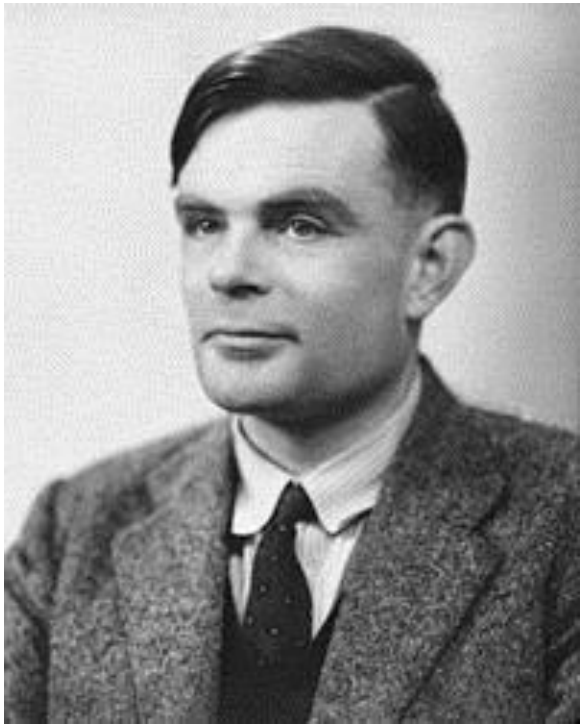
[Sheng et al, 2008; Snow et al, 2008]



Majority vote of 8 Turkers better than expert labeling

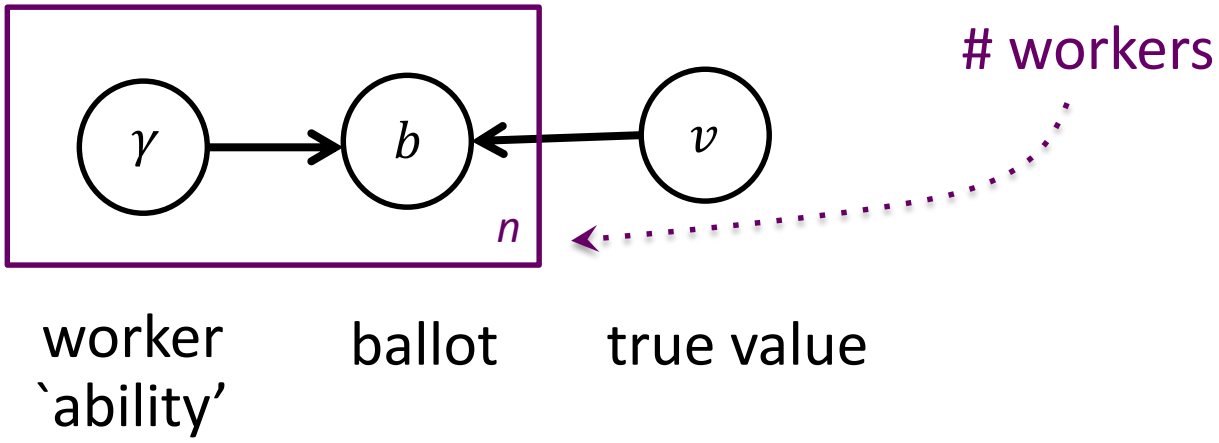
Worker Ability

$\gamma = 0$



$\gamma = \infty$

Quality-Corrected Voting



$$P(v \mid b_1,\ldots,b_n, \gamma_1,\ldots, \gamma_n) \sim P(b_1,\ldots,b_n \mid v, \gamma_1,\ldots, \gamma_n)P(v)$$
$$= P(v)\, \Pi_i P(b_i \mid v, \gamma_i)$$

Assumption: workers independent of each other

Quality-Corrected Voting 2

Are workers really independent?

Easy

Hard

Intrinsic difficulty (d∈[0,1]) measures problem hardness

Conditional Independence

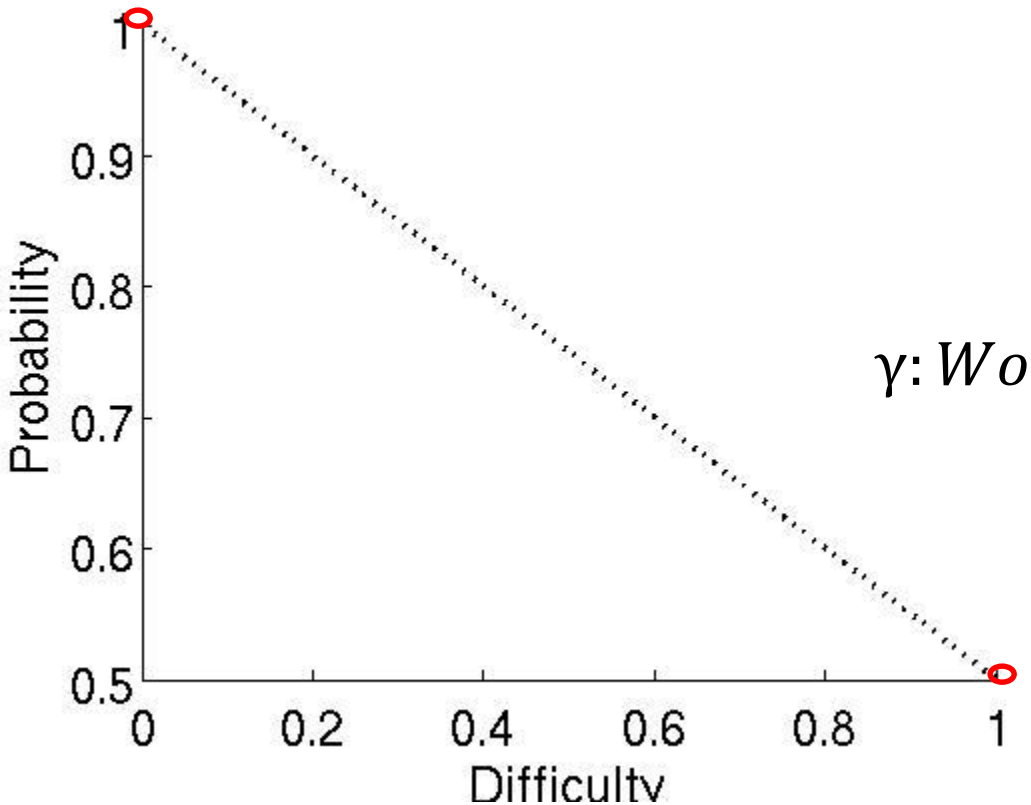
- workers independent given intrinsic difficulty

Probability of a Correct Answer

[Dai, Mausam, Weld AAAI'10]

$$a_w(d) = \frac{1}{2}[1 + (1-d)^{1/\gamma_w}]$$

Assume: no malevolence



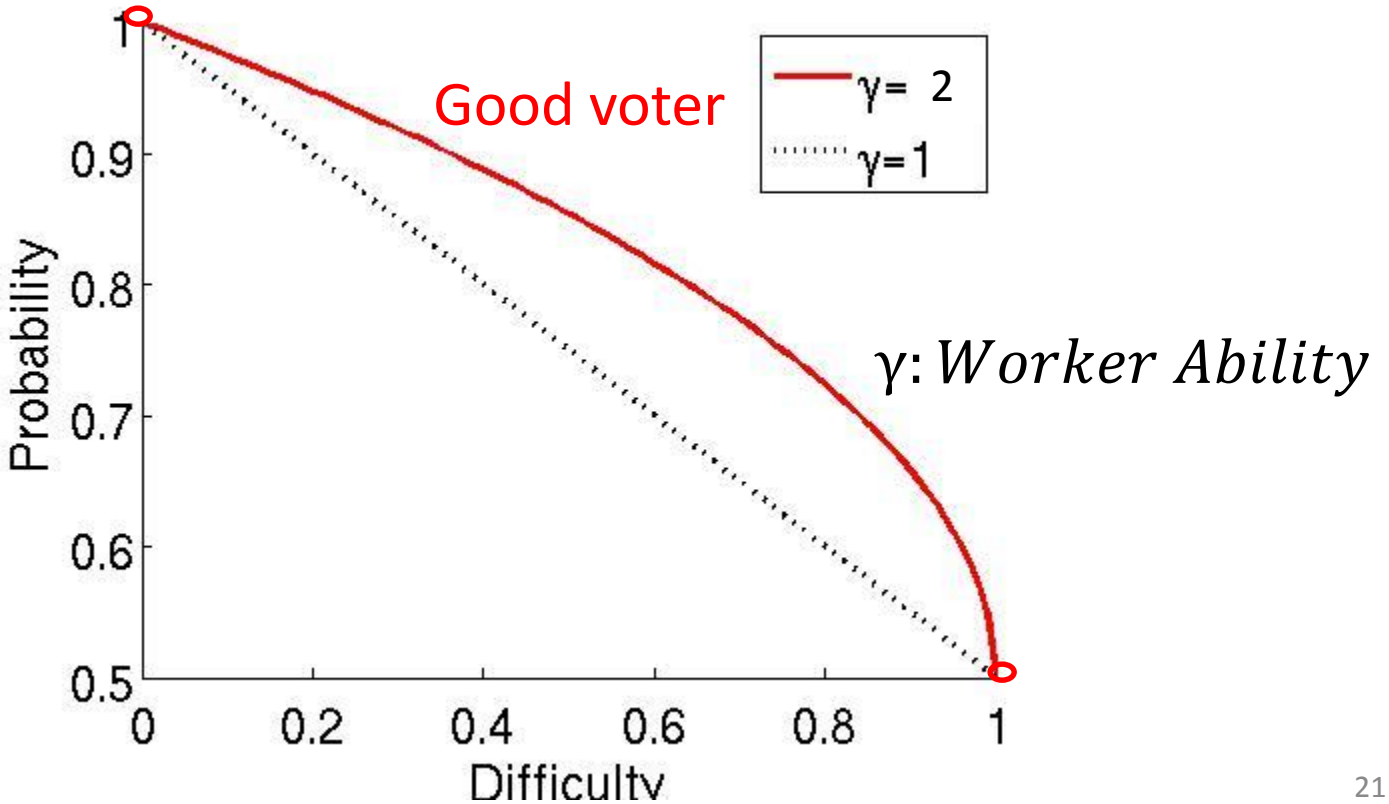
γ: *Worker Ability*

Probability of a Correct Answer

[Dai, Mausam, Weld AAAI'10]

$$a_w(d) = \frac{1}{2}[1 + (1-d)^{1/\gamma_w}]$$

Assume: no malevolence



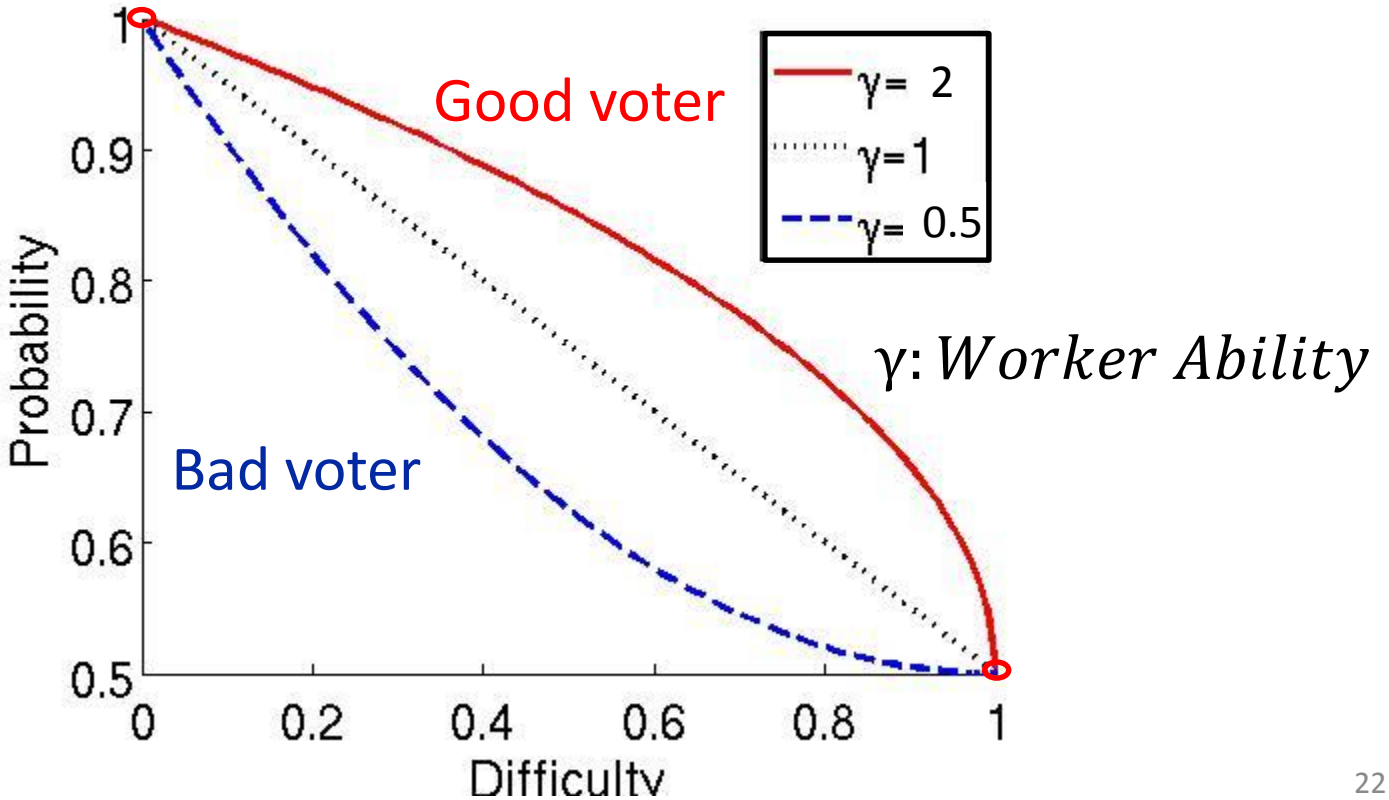
γ: *Worker Ability*

Probability of a Correct Answer

[Dai, Mausam, Weld AAAI'10]

$$a_w(d) = \frac{1}{2}[1 + (1-d)^{1/\gamma_w}]$$

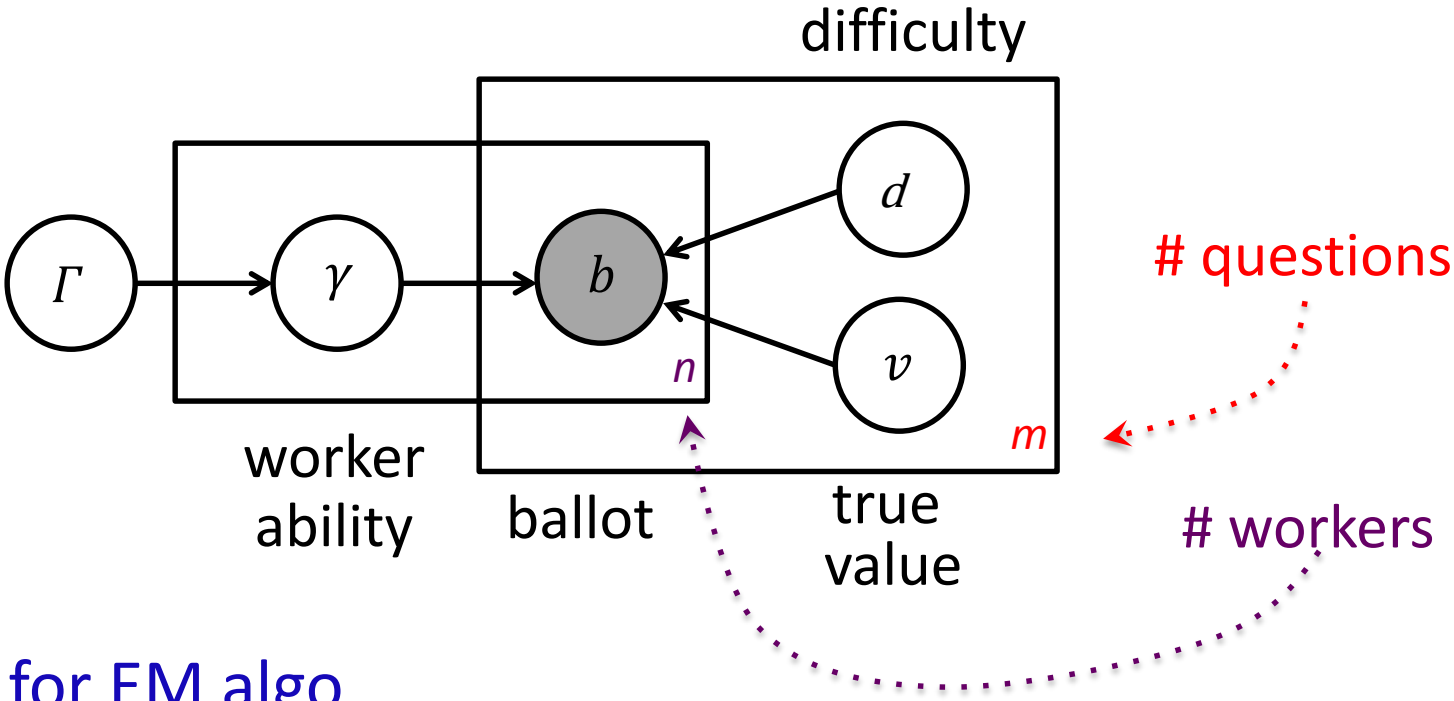
Assume: no malevolence



Unsupervised Learning

[Dai, Lin, Mausam, Weld AIJ' 13]

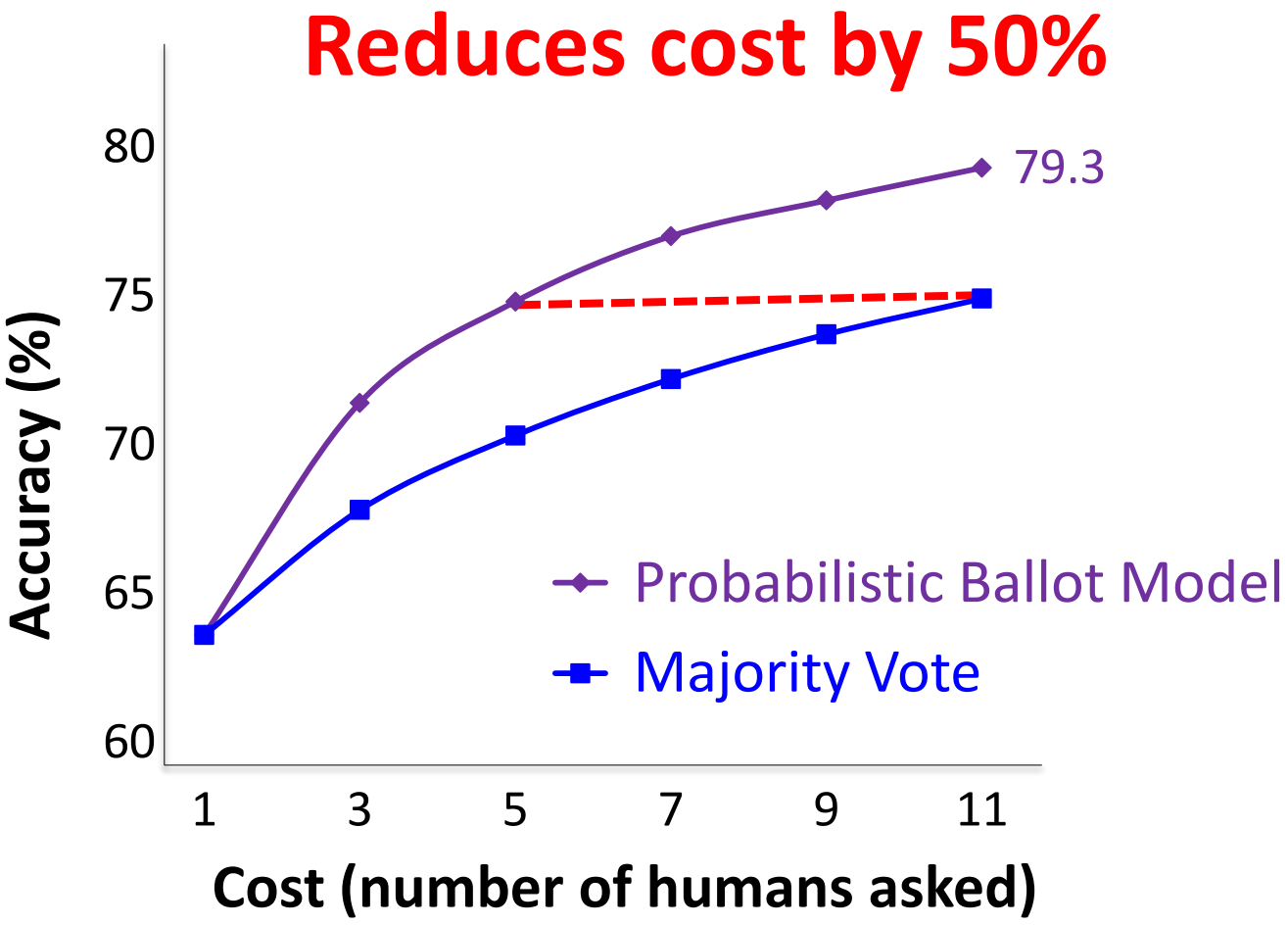
- No labeled data
- Joint estimation of all parameters: EM algorithm



- Intuitions for EM algo
  - one who commonly disagrees with others: ~spammer
  - one who usually agrees with others: ~good worker
  - if many workers disagree: ~hard question
  - as we identify some good workers, we trust them more…

Probabilistic Ballot Model

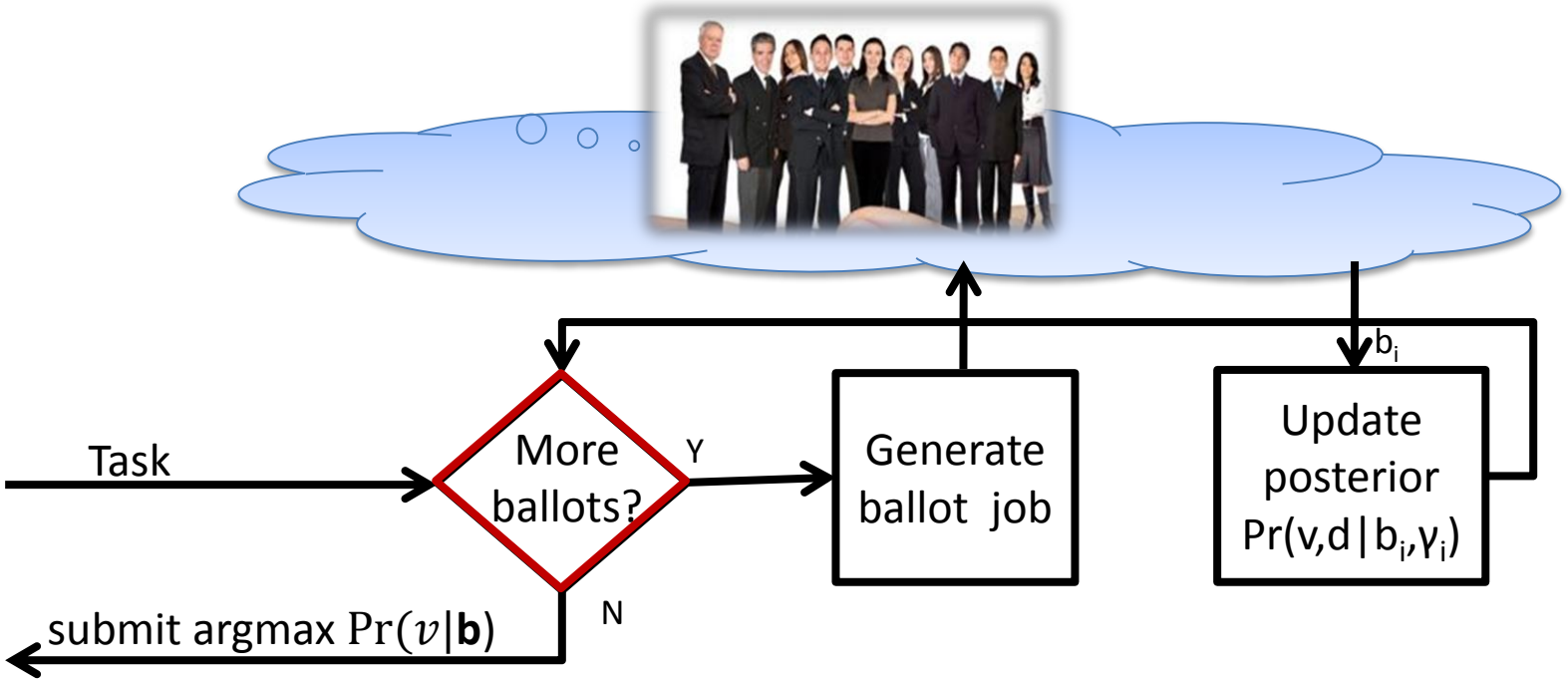
weight votes by learned accuracy



Planning & Control

- How many workers to ask per question?
- Simple Strategy
  - learn #workers that gets desired avg accuracy
- Is this good enough?
  - shouldn't difficult questions get more ballots?
  - if a good worker answers, do we need many more?
- *Dynamic* Control of Crowdsourced Workflows

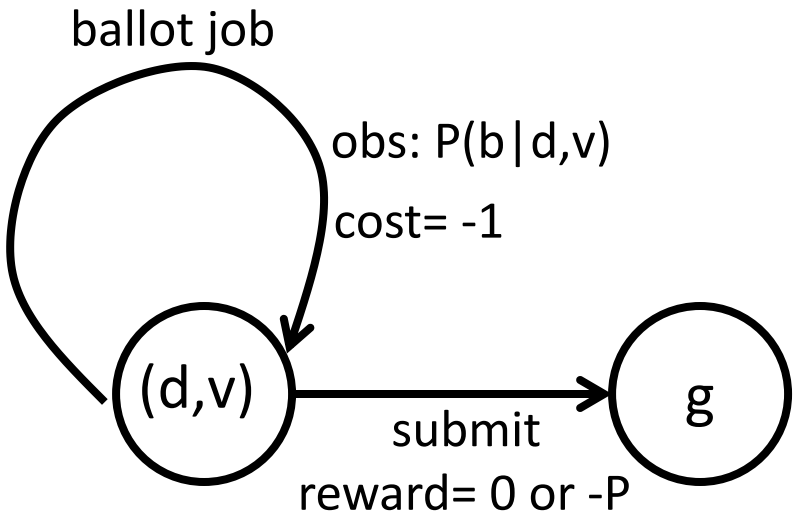
Clowder$_{simple}$



Decision Making under Uncertainty (POMDP)

Partially Observable Markov Decision Process

- States (hidden): (d, v) and goal
- Actions:
  - Generate ballot job
  - Submit the best answer
- Transitions
- Observations



- Cost: money spent per ballot job
- Rewards: User-defined penalty for incorrect answer

- Maximize E[Reward-Cost]

POMDP Solutions

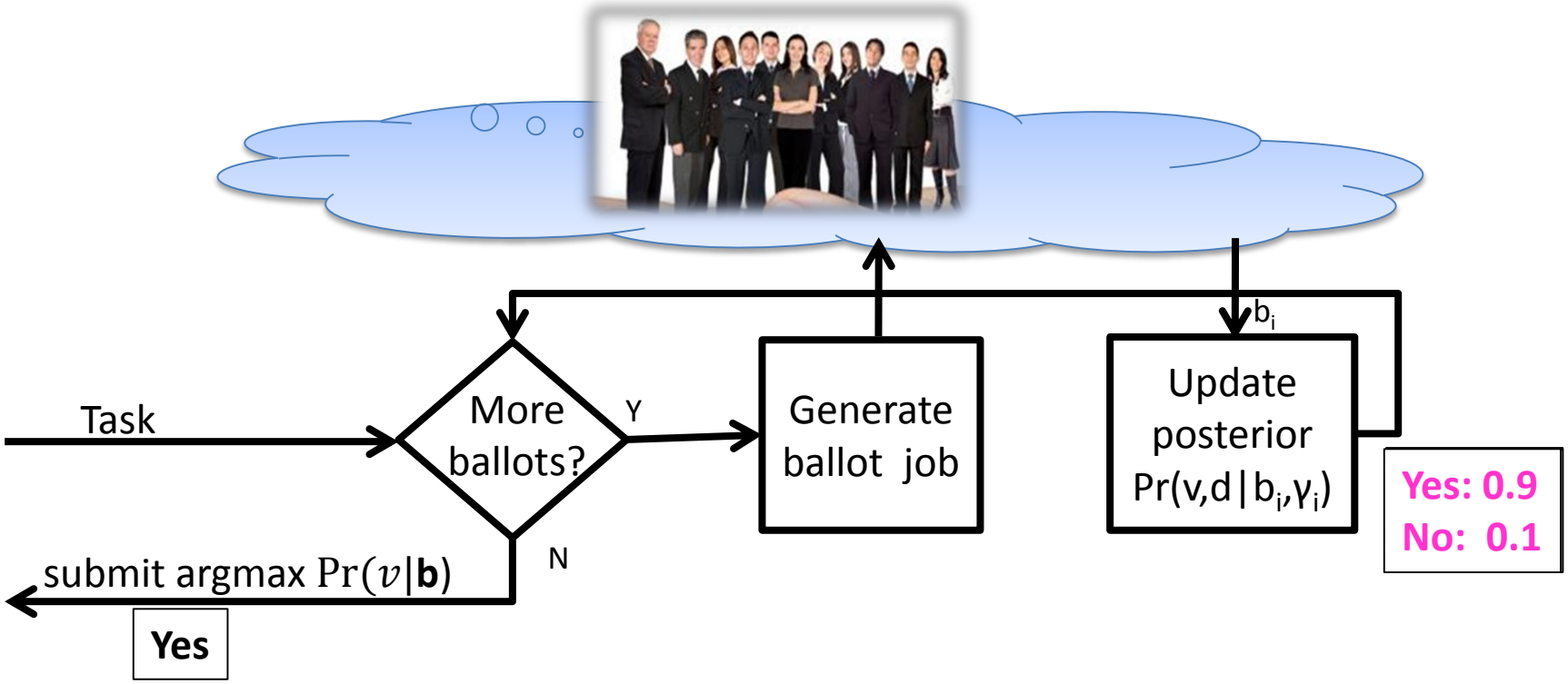
- Because state is hidden the agent maintains a *belief* = probability distribution over world states

- Bellman Equations

$$V^*(b) = \max_a \left[\sum_s R(s,a)b(s) + \gamma \int \sum_z P(b' \mid b, a, z) V^*(b') db' \right]$$

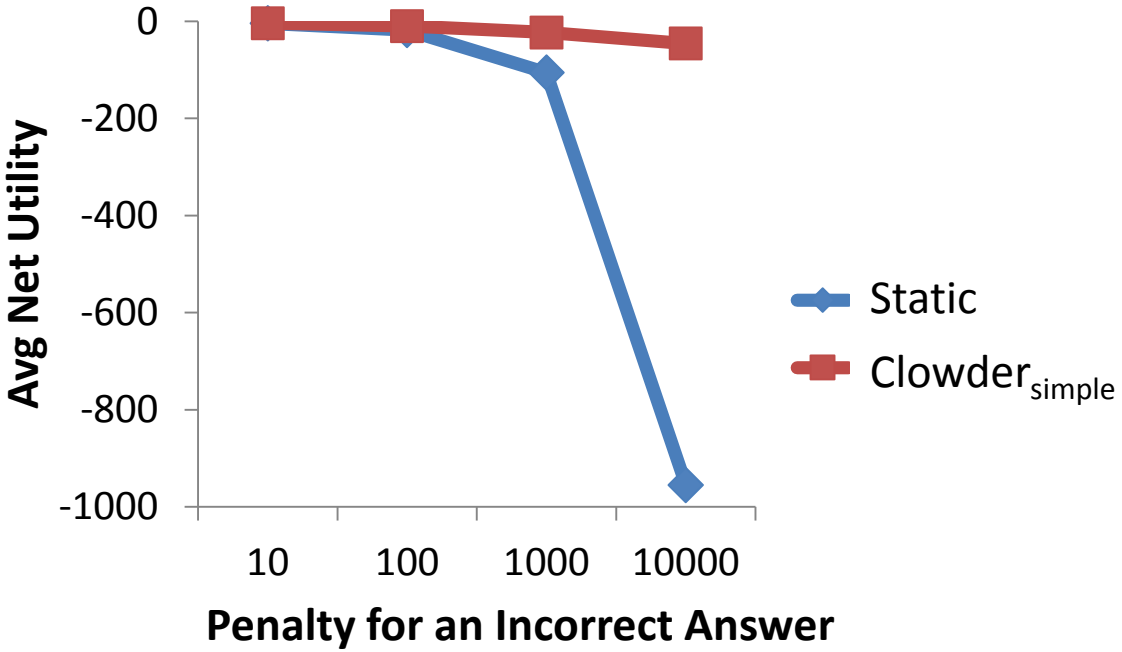
- Dynamic Programming
- Piece-wise linear value function representation
- Approximated by reducing # of linear components

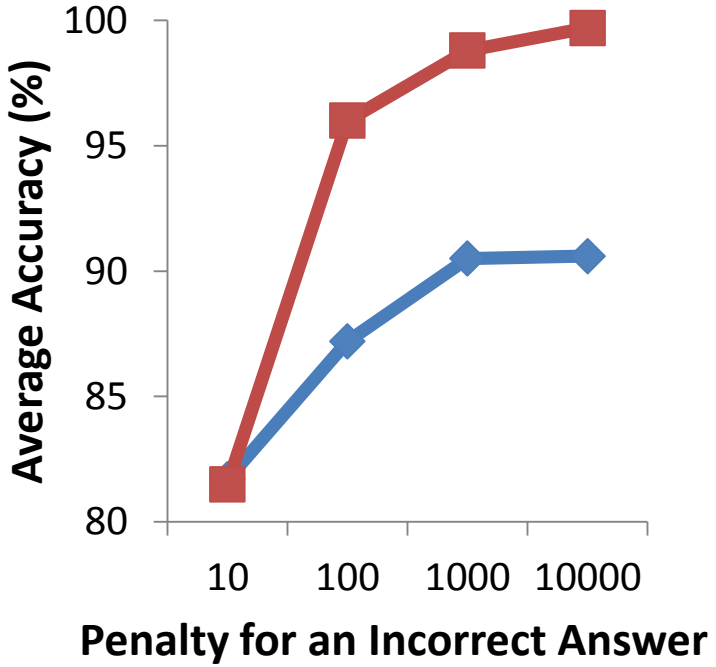
Clowder$_{simple}$



Experiments

[Dai, Lin, Mausam, Weld AIJ'13]





Intelligent Decisions

- Less money spent on confusing (hard) questions
- Better output due to worker tracking
- May be able to handle tricky questions

Summary: Simple Consensus Tasks

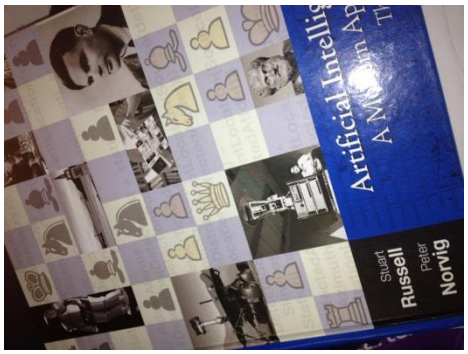
- A generative model for worker response
  - based on question difficulty and worker ability
- An unsupervised approach for joint learning
- POMDP-based control
  - for superior quality output for the same price

Outline

- Motivation for Crowdsourcing
- Crowdsourcing Challenges: design of Clowder
- Control of Crowdsourced Workflows
  - Simple consensus multiple-choice tasks
  - Simple consensus tasks with infinite answer choices
  - Tasks with complex workflows
  - Tasks with multiple workflows
  - Tasks with multi-class categorization
- Future Work

Tasks with Infinite Answer Choices

Cut/paste the Amazon page for the product in this picture



Call this restaurant and find an email address of the manager

Restaurant Name:	**Danny's Tobacco**
Address:	**615 N Main St** **Elizabethtown, Kentucky 42701**
Phone Number:	**(270) 737-1333**
Email:	

Which Canon product is closest to Nikon D40?

Challenges

- Inference
  - Infinitely growing answer space
  - Chinese Restaurant Process for inference
- Control
  - Trick: model probability that answer is unseen
  - Infinitely growing state/action space
  - Algorithm: limited lookahead search

Experiments

[Lin, Mausam, Weld UAI'12]

- Task: 134 S.A.T. Math Questions
- Penalty of incorrect answer: -$1

	Clowder$_{infinite}$	MajorityVote (7)
Avg Accuracy (%)	**99.25**	**95.52**
Avg Cost	**5.17 cents**	**5.46 cents**
Avg Net Utility	**-5.92**	**-9.94**

- 83% error reduction
- More experiments needed on tougher tasks

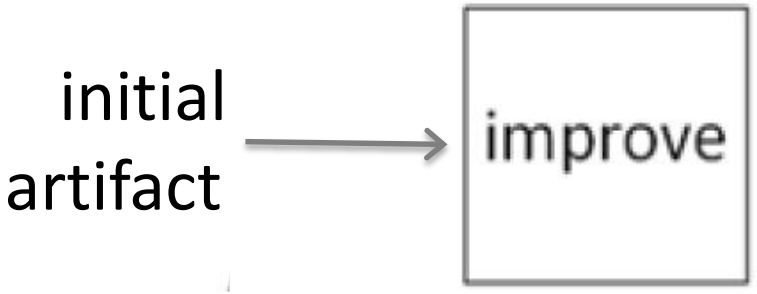
Outline

- Motivation for Crowdsourcing
- Crowdsourcing Challenges: design of Clowder
- Control of Crowdsourced Workflows
  - Simple consensus multiple-choice tasks
  - Simple consensus tasks with infinite answer choices
  - Tasks with complex workflows
  - Tasks with multiple workflows
  - Tasks with multi-class categorization
- Future Work

Simple → Complex Tasks

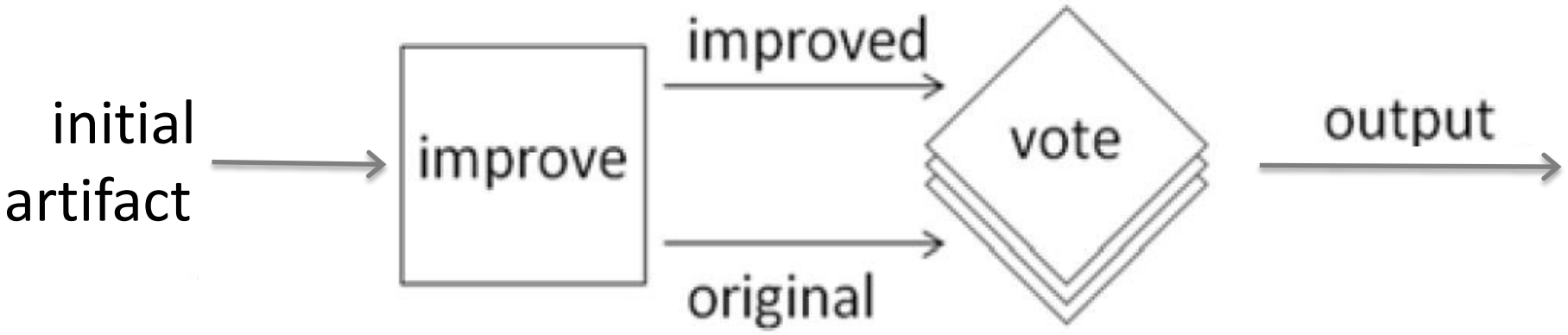
- Workflows change the game
- Dividing a complex task into smaller jobs
  - information flow between these jobs
- Examples
  - audio transcription (CastingWords proprietary)
  - generating articles (Iterative Improvement)
  - handwriting recognition (Iterative Improvement)
  - Soylent: intelligent word processor (Find-Fix-Verify)
  - …

Iterative Improvement Workflow



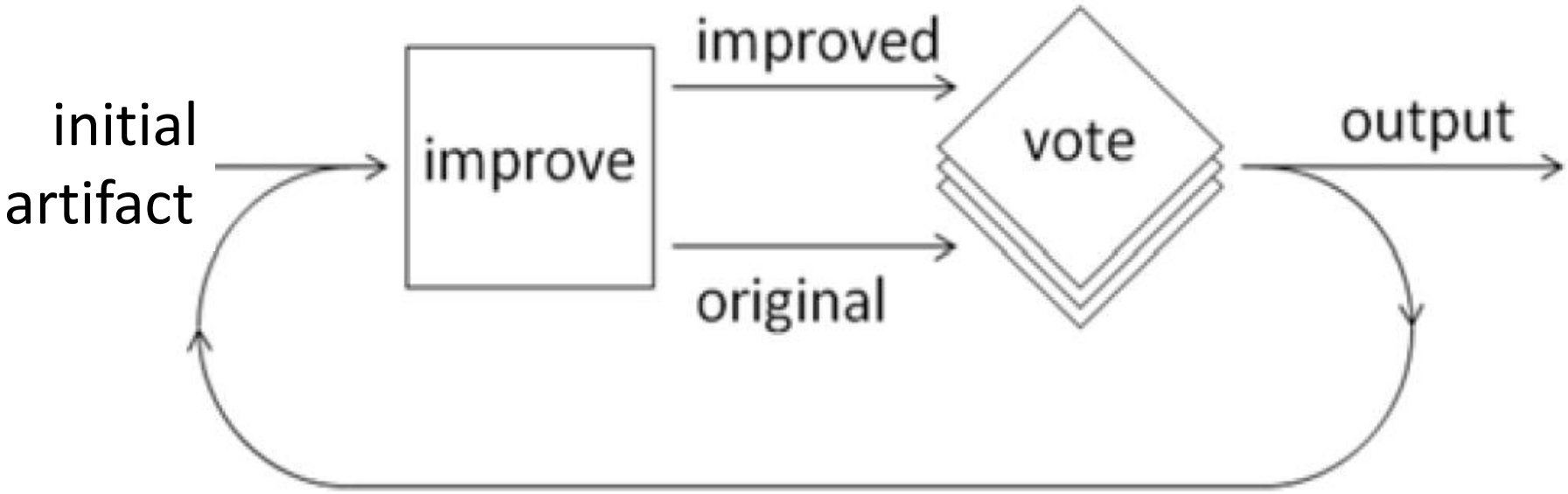
[Little et al, 2010]

Iterative Improvement Workflow



[Little et al, 2010]

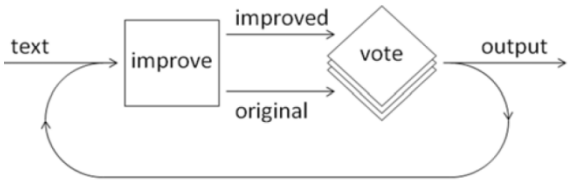
Iterative Improvement Workflow



[Little et al, 2010]

Iterative Improvement Workflow

[Little et al, 2010]



First version

A parial view of a pocket calculator together with some coins and a pen.

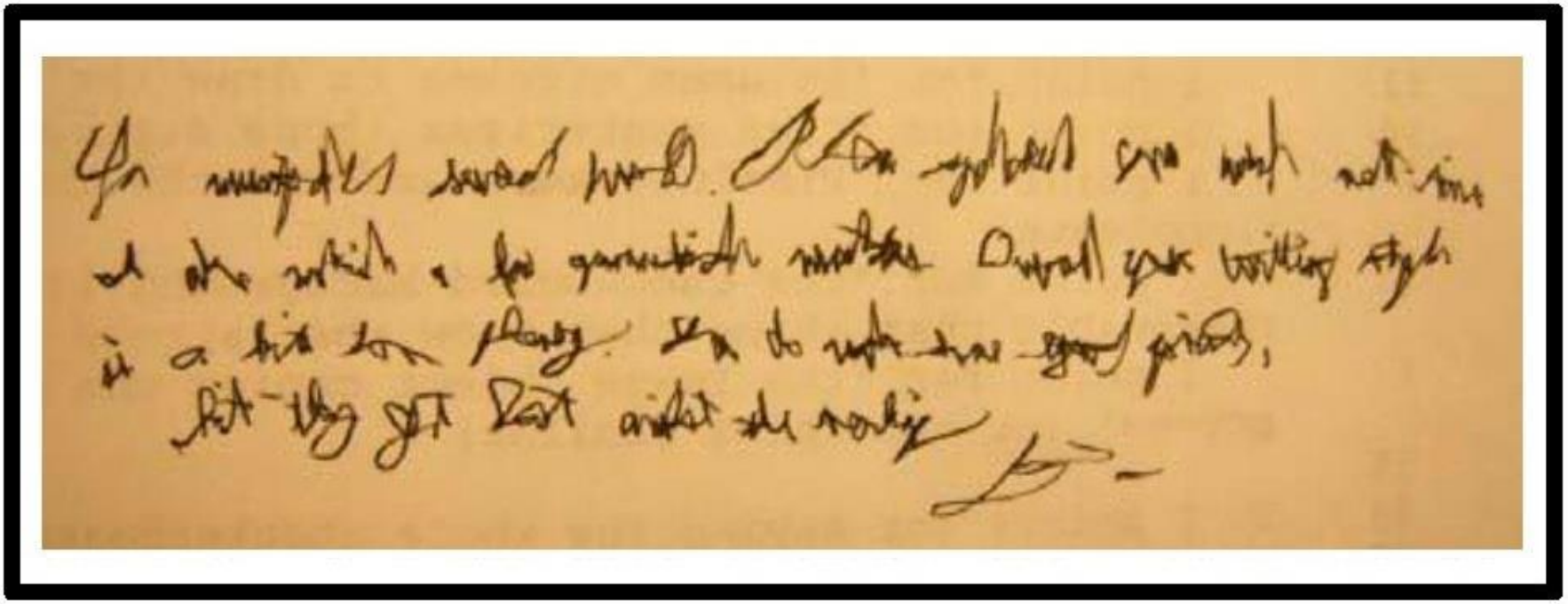
Version after 8 iterations

A CASIO multi-function, solar powered scientific calculator.

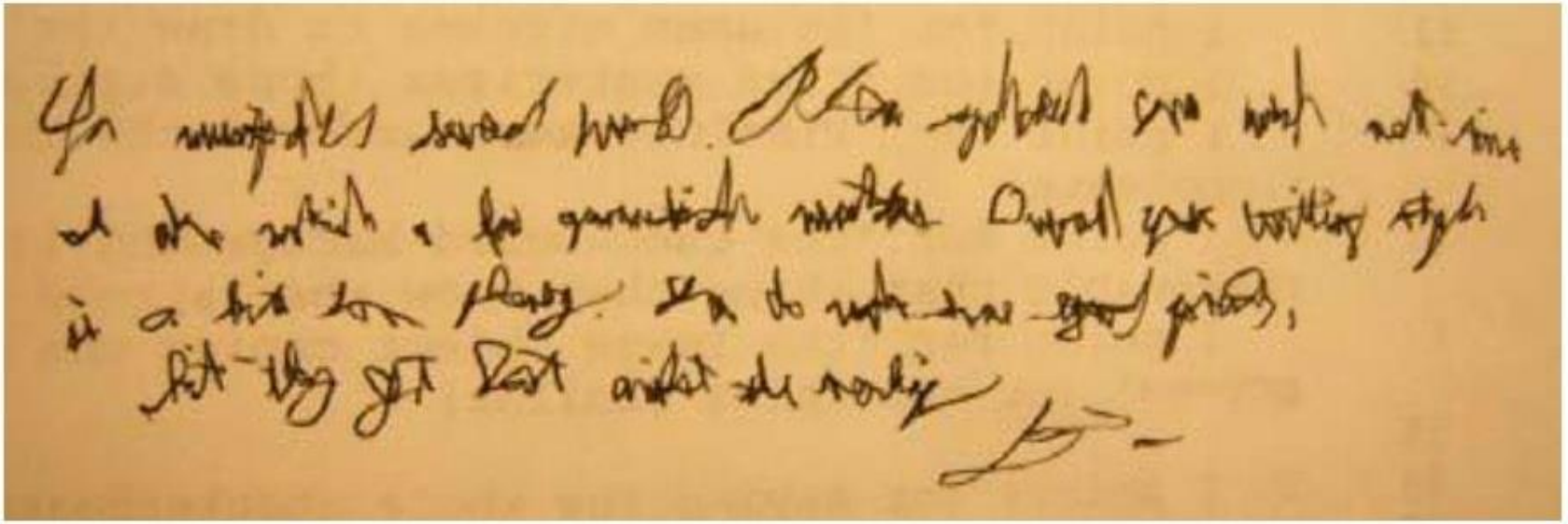
A blue ball point pen with a blue rubber grip and the tip extended.

Six British coins; two of £1 value, three of 20p value and one of 1p value.

Seems to be a theme illustration for a brochure or document cover treating finance - probably personal finance.



[Little et al, 2010]



"You (misspelled) (several) (words). Please spellcheck your work next time. I also notice a few grammatical mistakes. Overall your writing style is a bit too phoney. You do make some good (points), but they got lost amidst the (writing). (signature)"

According to our ground truth, the highlighted words should be "flowery", "get", "verbiage" and "B-" respectively.

[Little et al, 2010]

Quality Control of Complex Tasks

- Challenge: there are infinite answers.
- Challenge: there is ***no*** correct answer!
  - what to track?

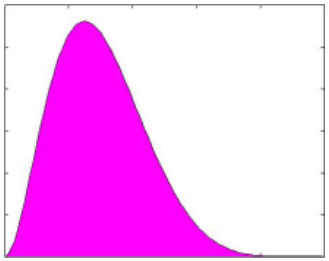
- Define Quality ***q***: a measure of answer goodness
- Track quality as more artifacts are created
- $q \epsilon [0,1]$

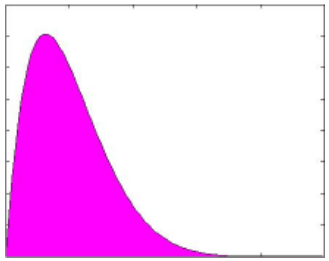
Iterative Improvement POMDP

[Dai, Mausam, Weld AAAI'10]

- **World State:** Quality of artifact(s)
- **Belief State:** Probability distribution over world states
- **Actions:** Submit jobs to labor mkt & observe results
  - Eg, improve job — prob distribution on new artifact
  - Eg, vote job — Bayesian update on quality
    EM update on difficulty, worker diligence
- **Objective:** Maximize $\mathbf{E}[R(w) - \sum c]$

Belief States





Artifact qualities

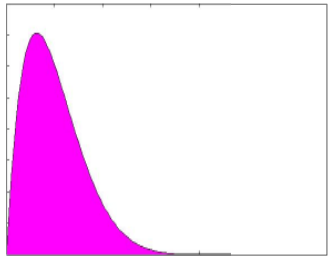
$Q_1 \in [0, 1]$

$Q_2 \in [0, 1]$

Approximate with Beta distribution,
Truncated normal or
Discretized approximation

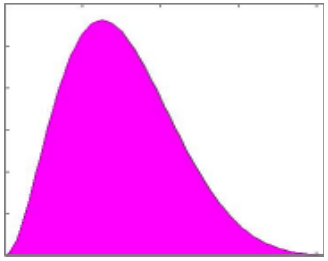
Transition Model of Improve Action

Prior(quality$_{\alpha 1}$)



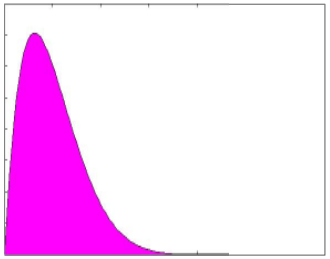
Worker creates new artifact

Prior(quality $_{\alpha 2}$)

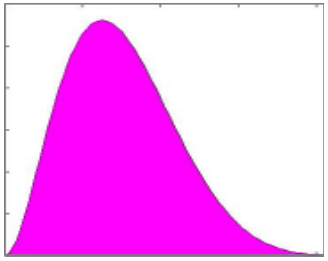


Transition Model of Ballot Action

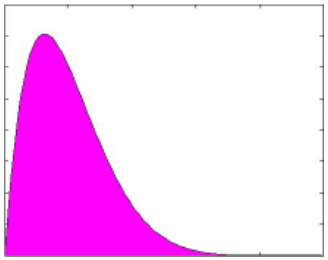
Prior(quality $_{\alpha 1}$)



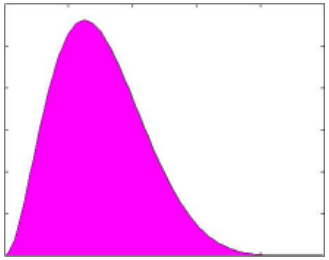
Prior(quality $_{\alpha 2}$)



Worker votes that artifact 1 is better

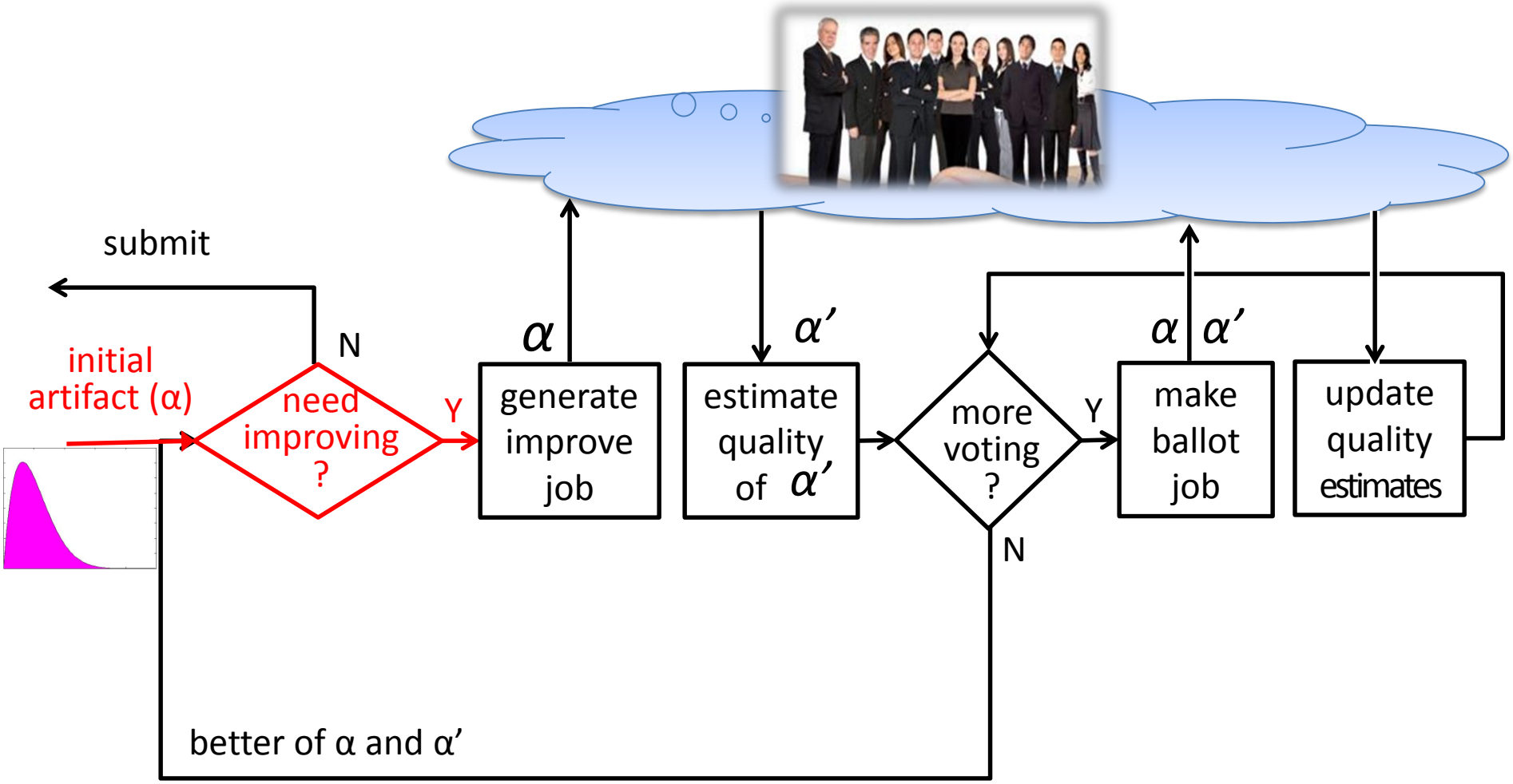


Posterior(quality $_{\alpha 1}$)

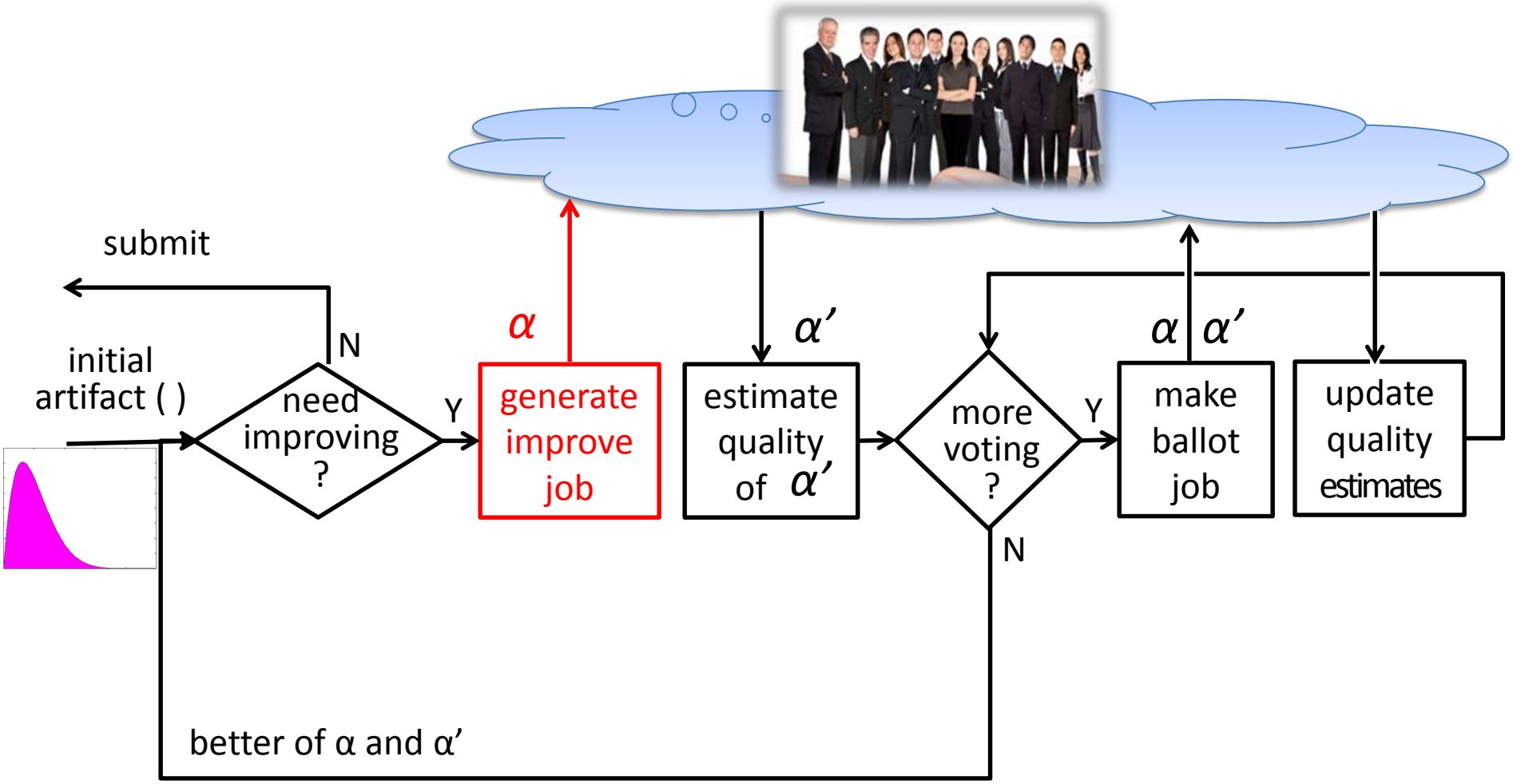


Posterior(quality $_{\alpha 2}$)

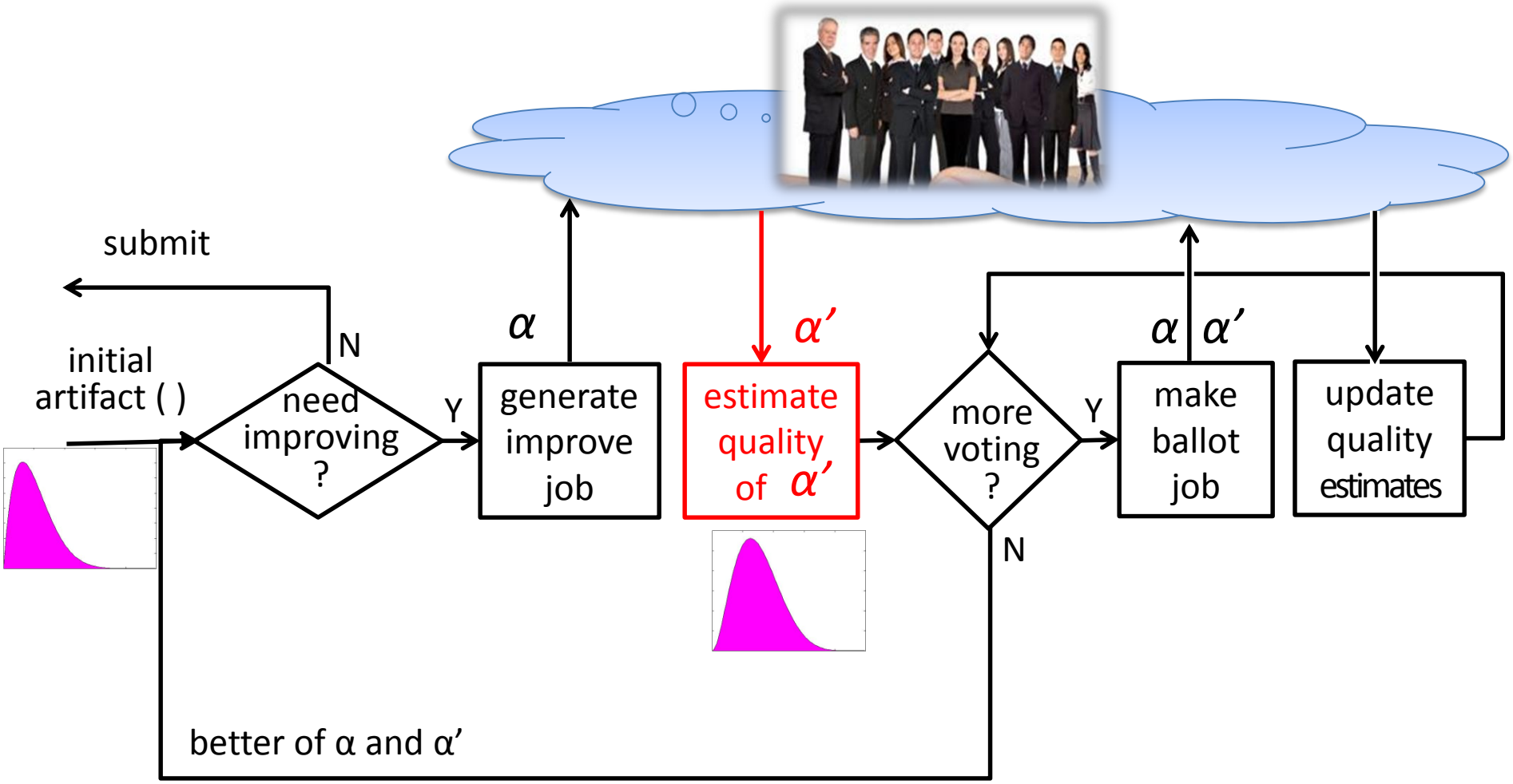
Clowder$_{iterative}$



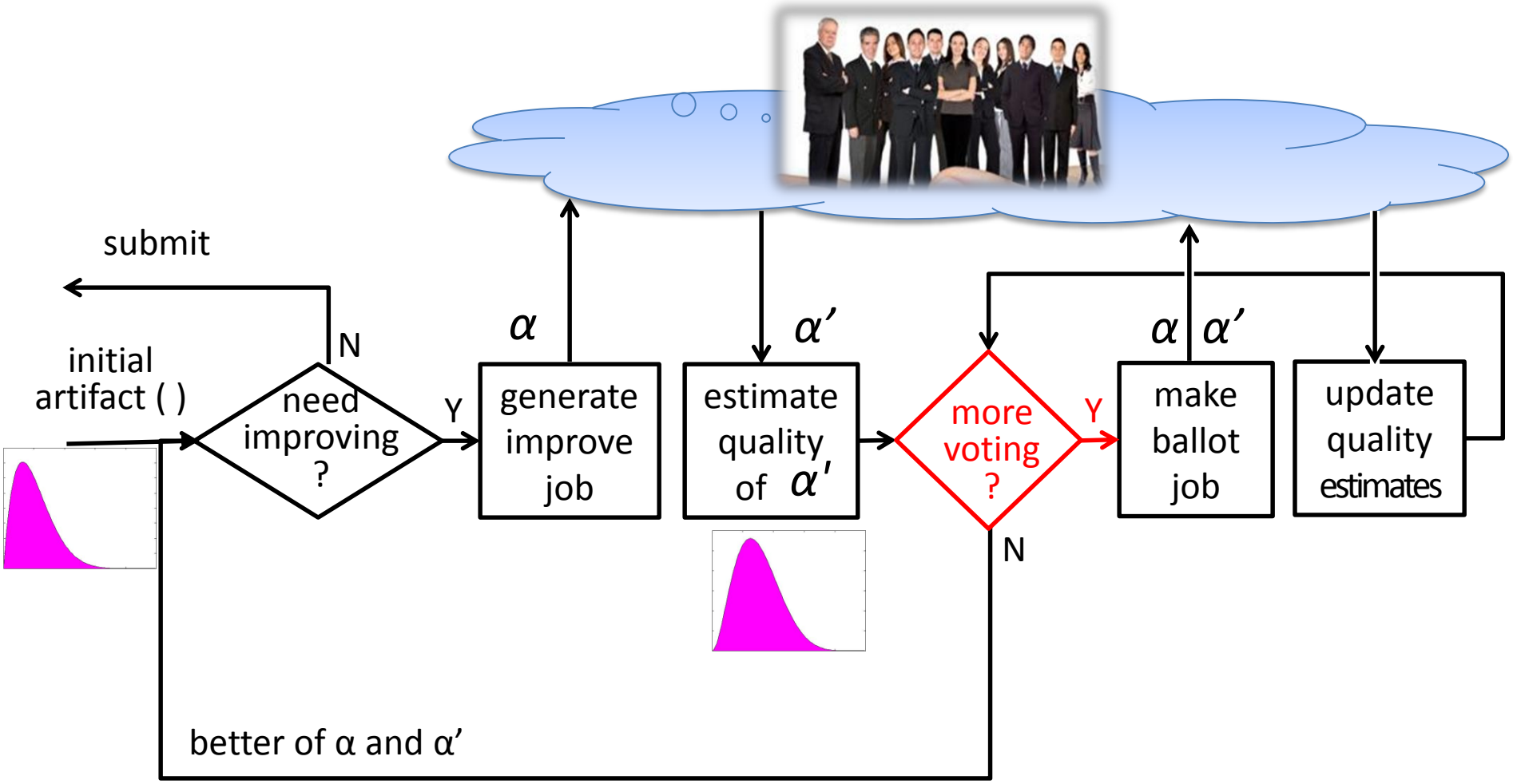
Clowder$_{iterative}$



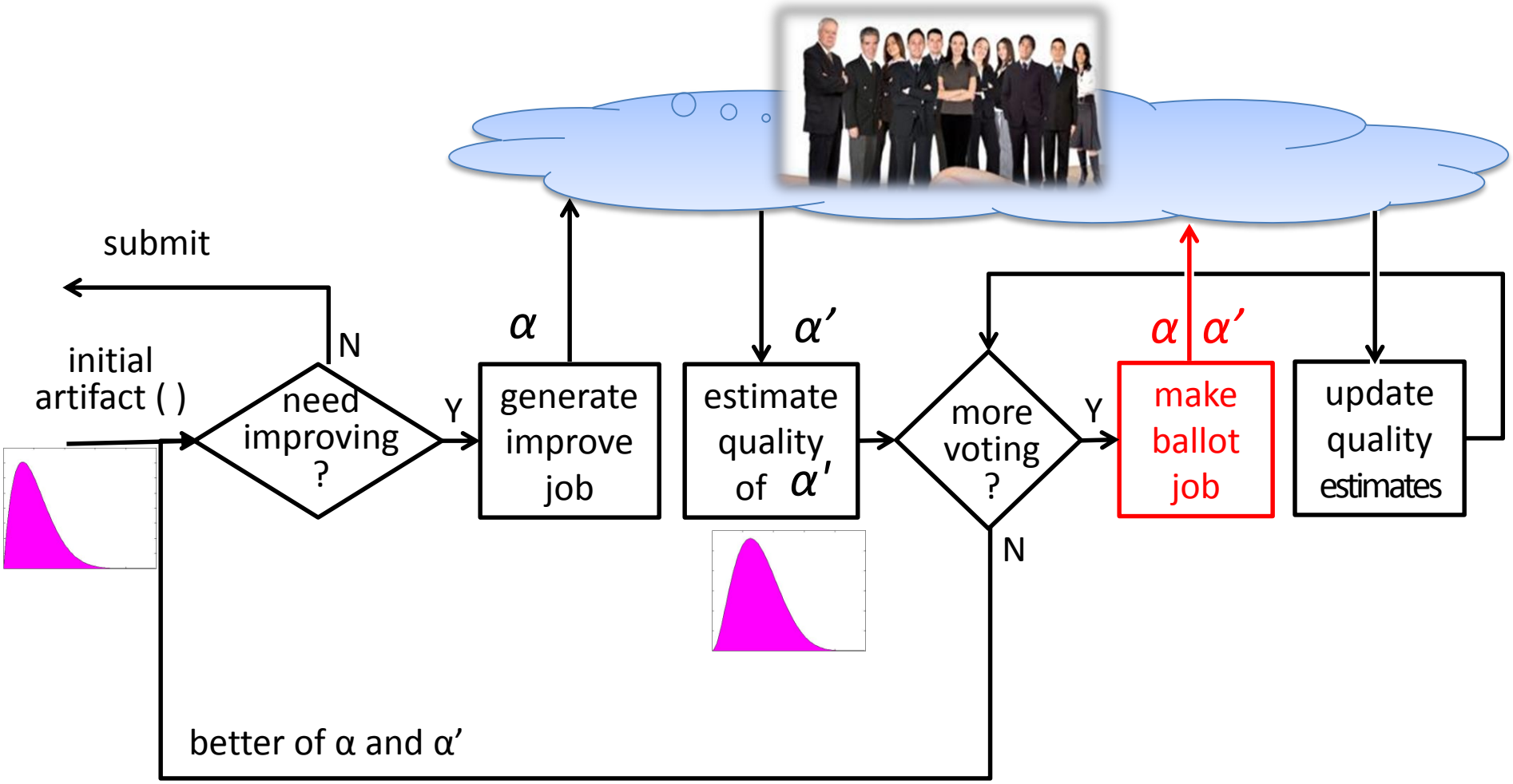
Clowder$_{iterative}$



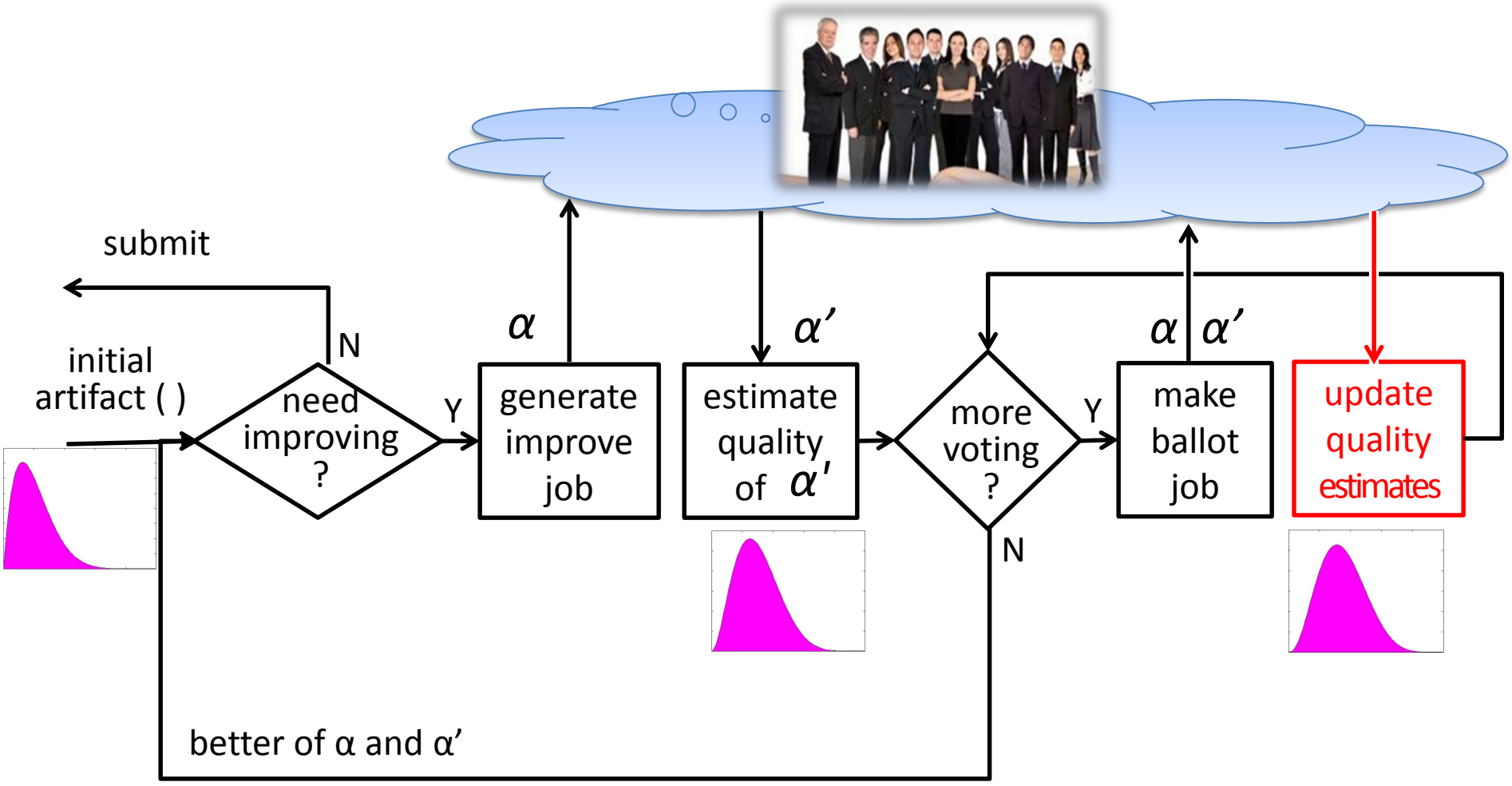
Clowder$_{iterative}$



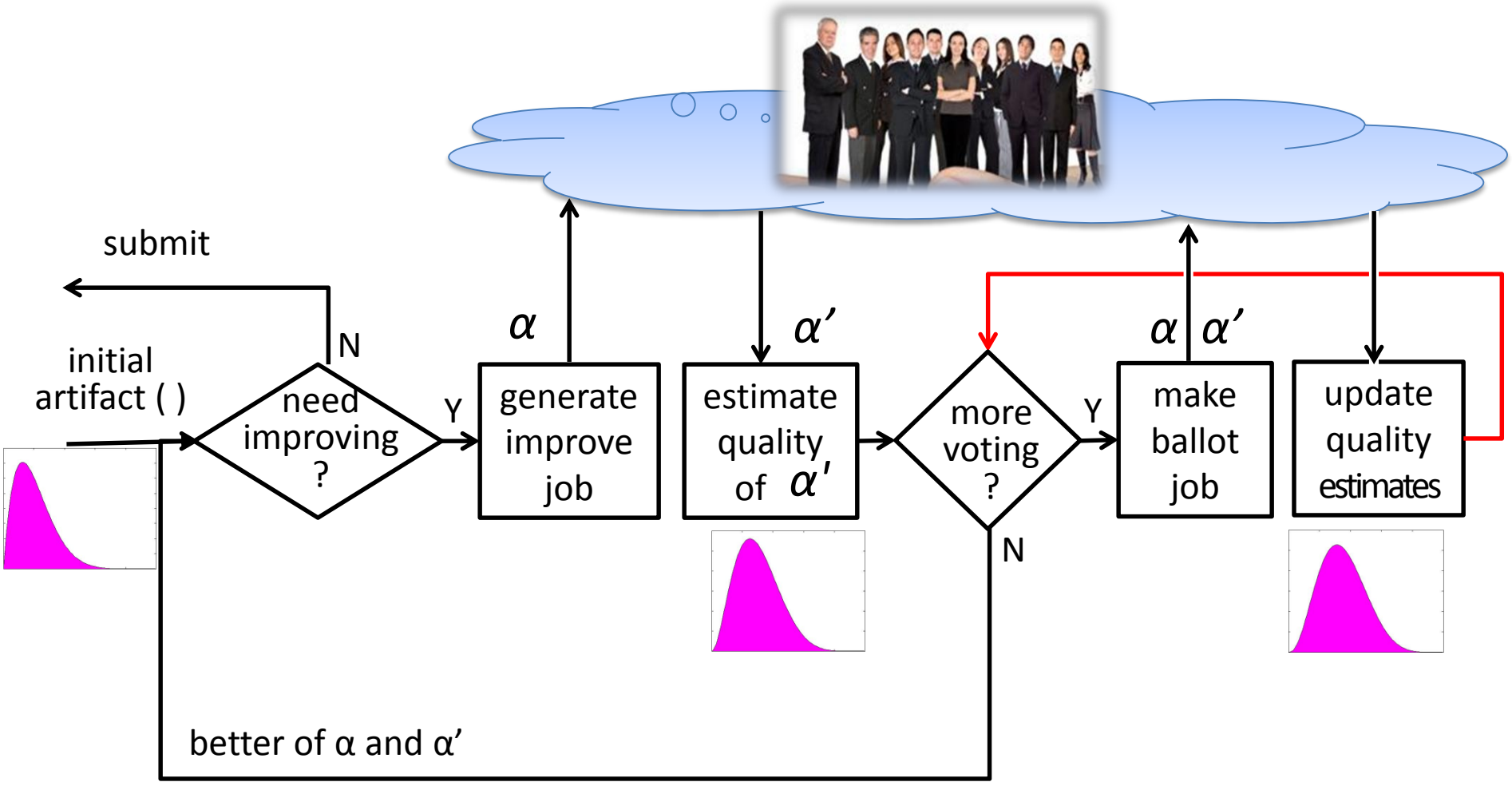
Clowder$_{iterative}$



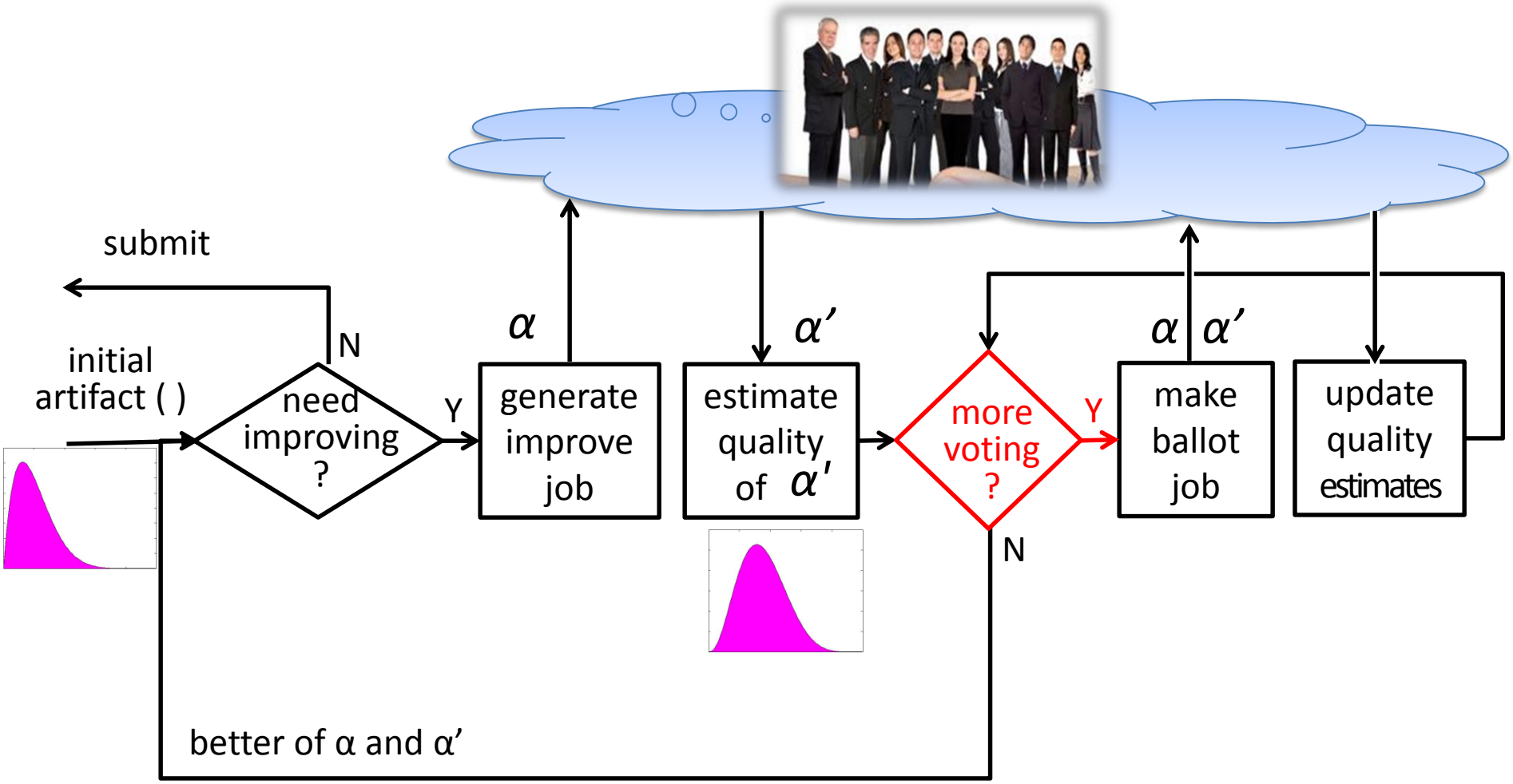
Clowder$_{iterative}$



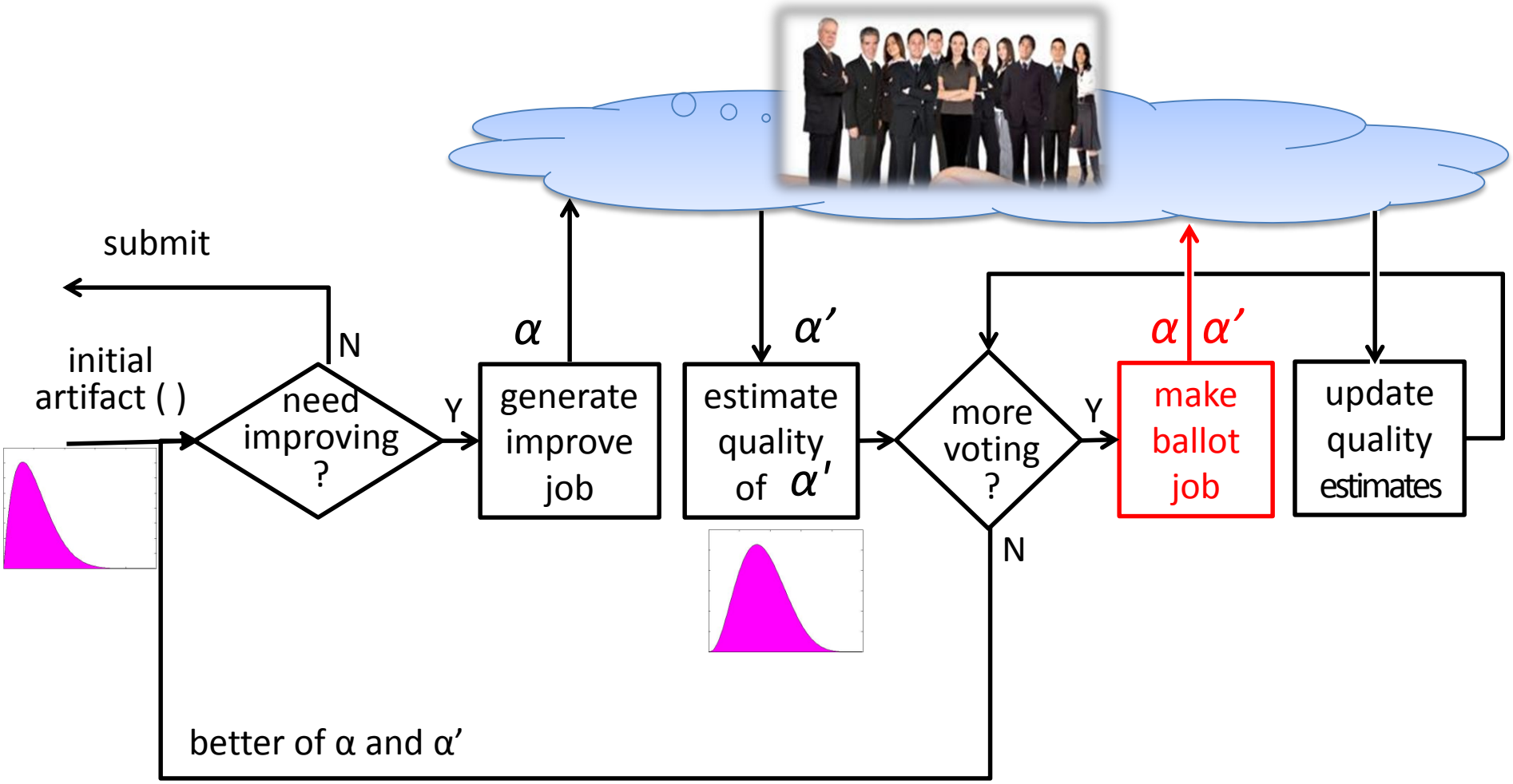
Clowder$_{iterative}$



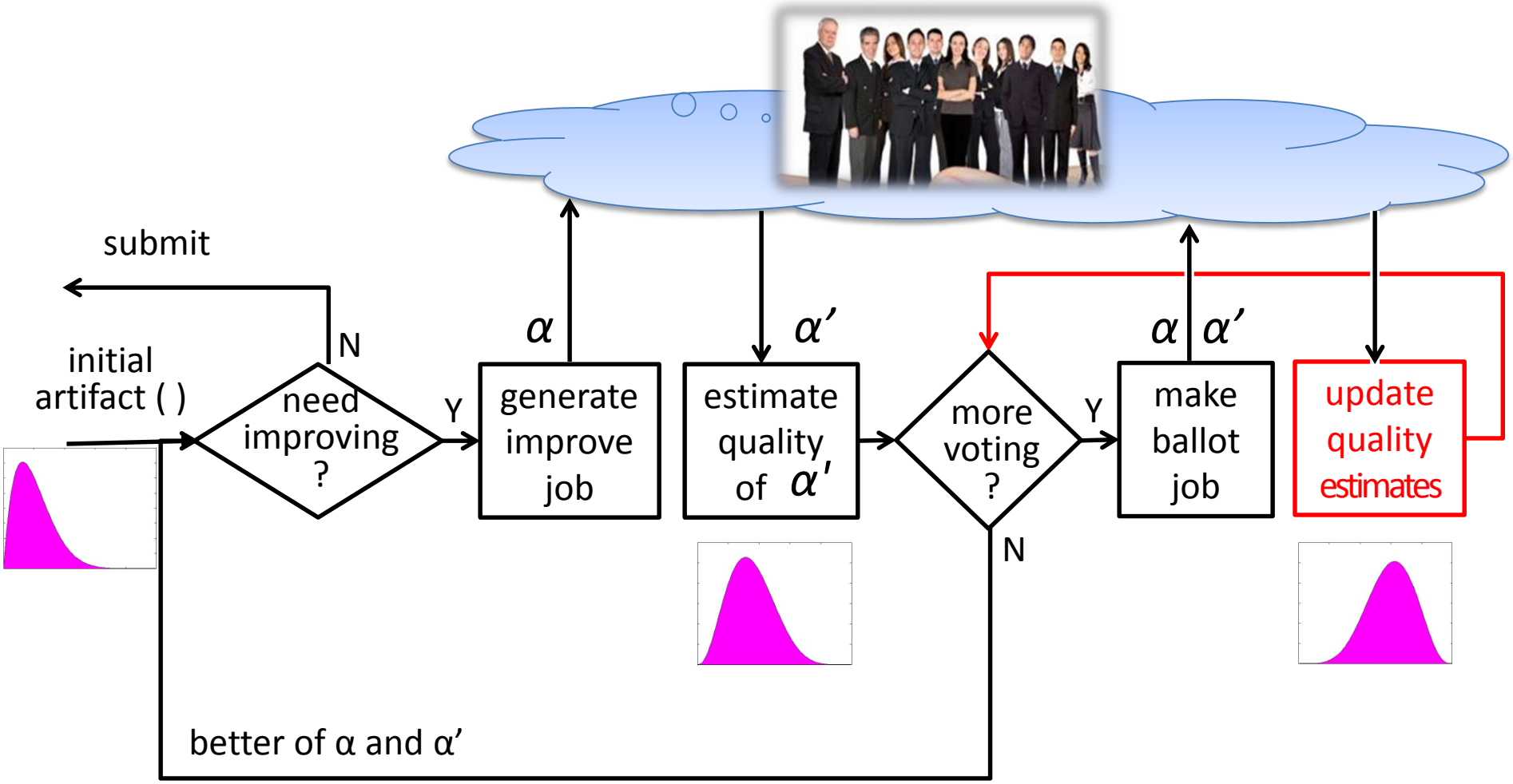
Clowder$_{iterative}$



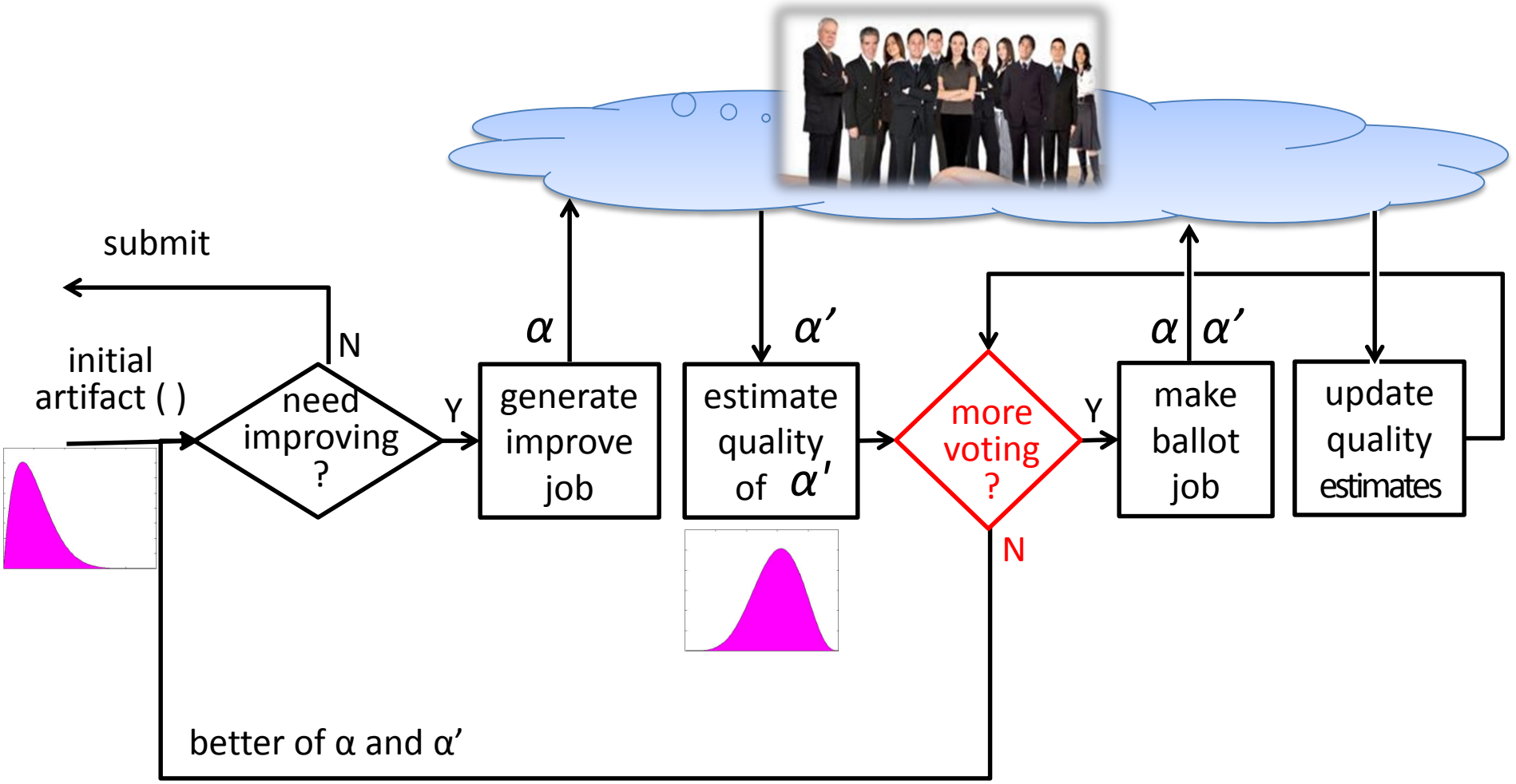
Clowder$_{iterative}$



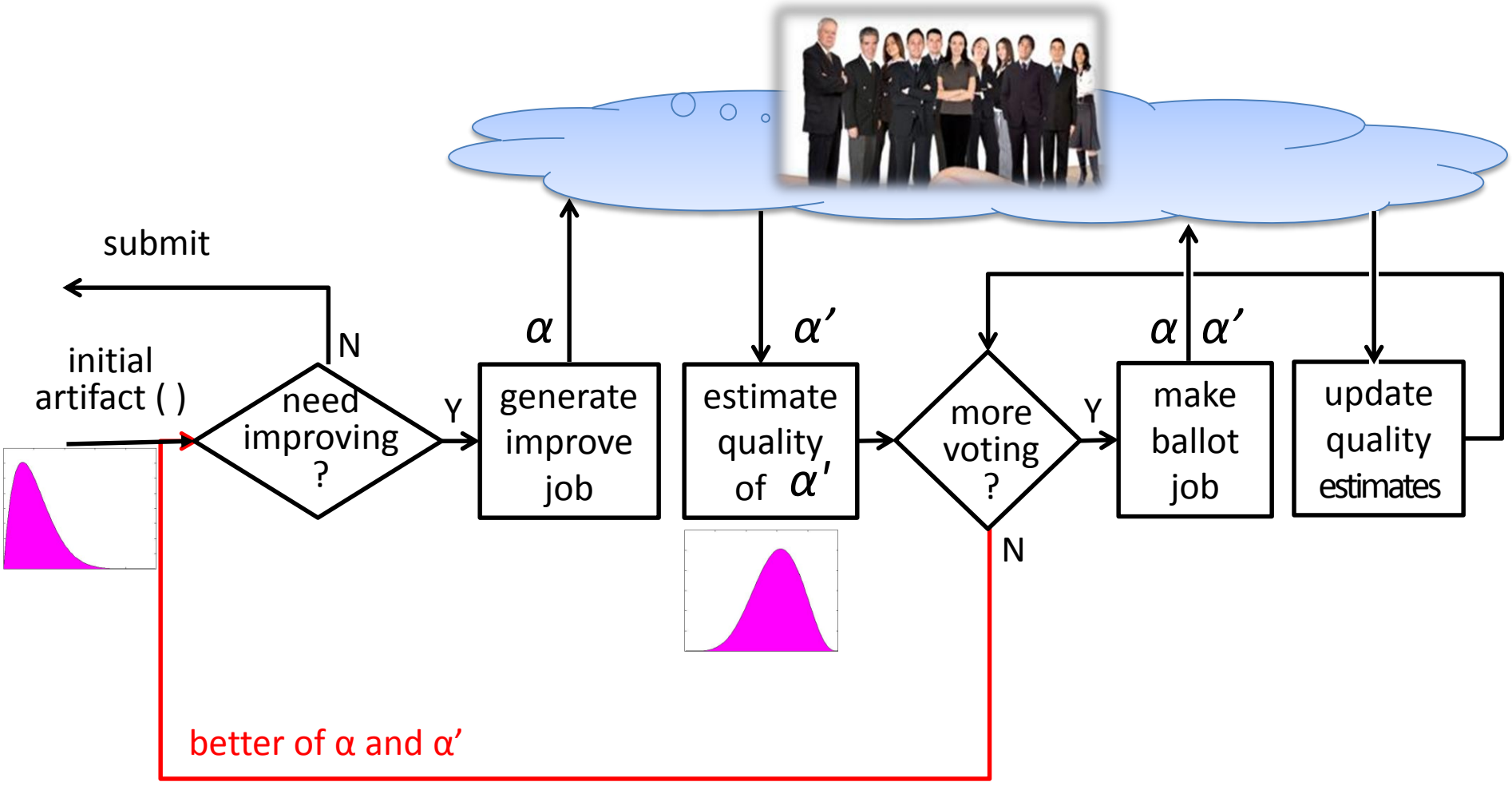
Clowder$_{iterative}$



Clowder$_{iterative}$



Clowder$_{iterative}$



Clowder$_{iterative}$

submit

initial artifact ( )

need improving ?

N

Y

generate improve job

α

estimate quality of α'

α'

more voting ?

Y

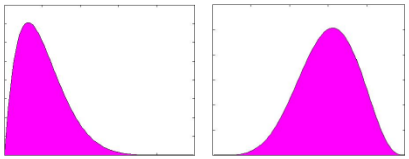
N

make ballot job

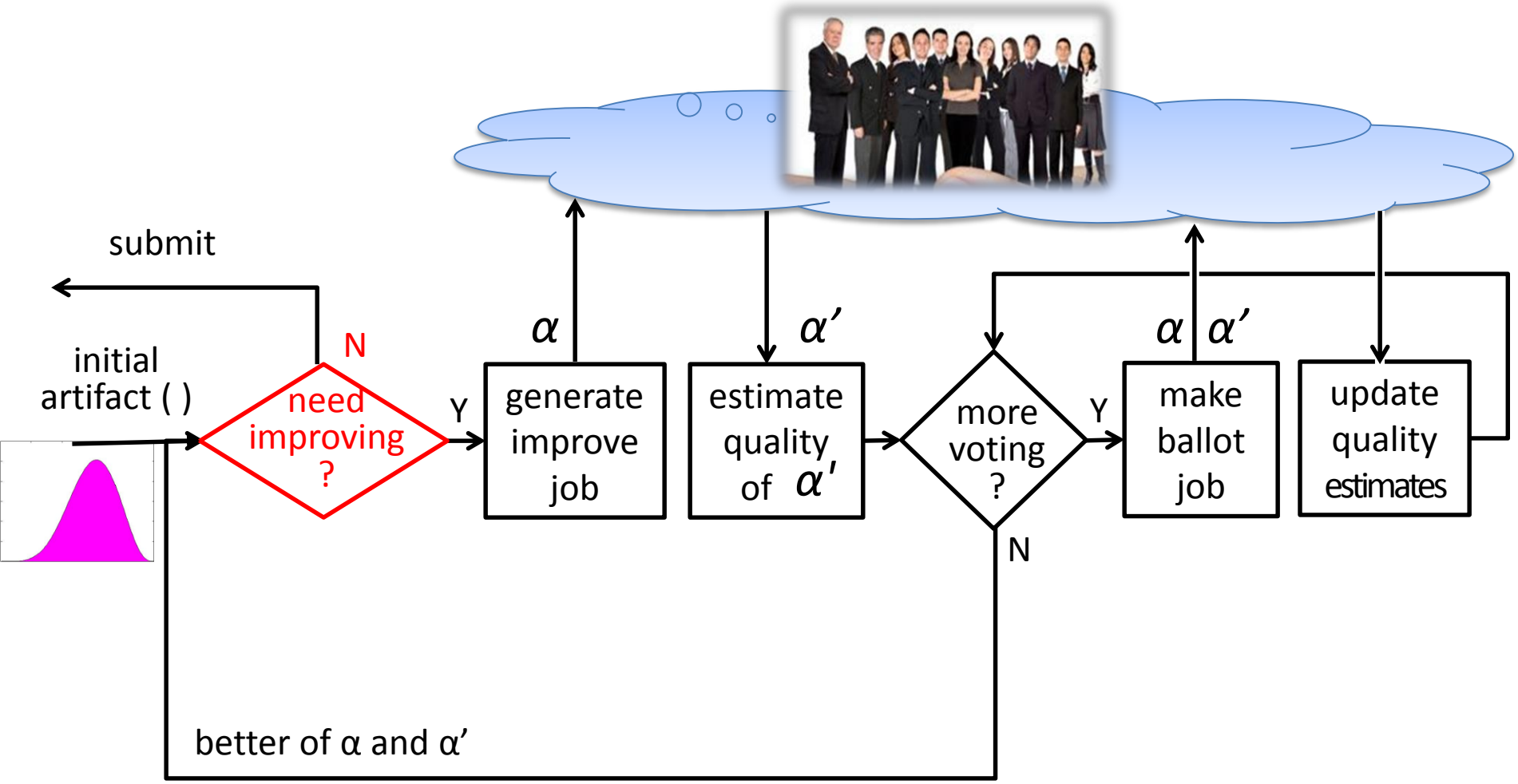
α α'

update quality estimates

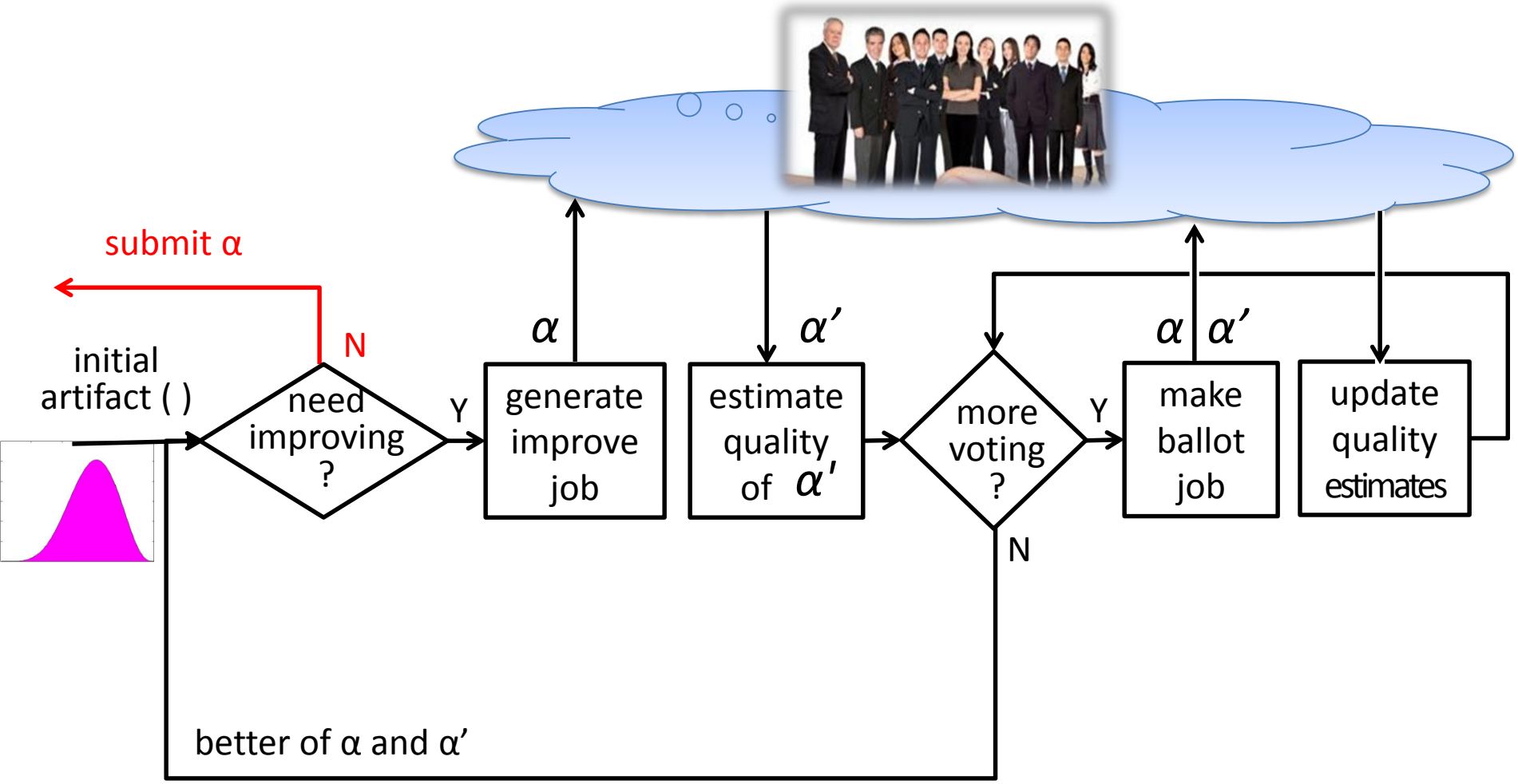
better of α and α’



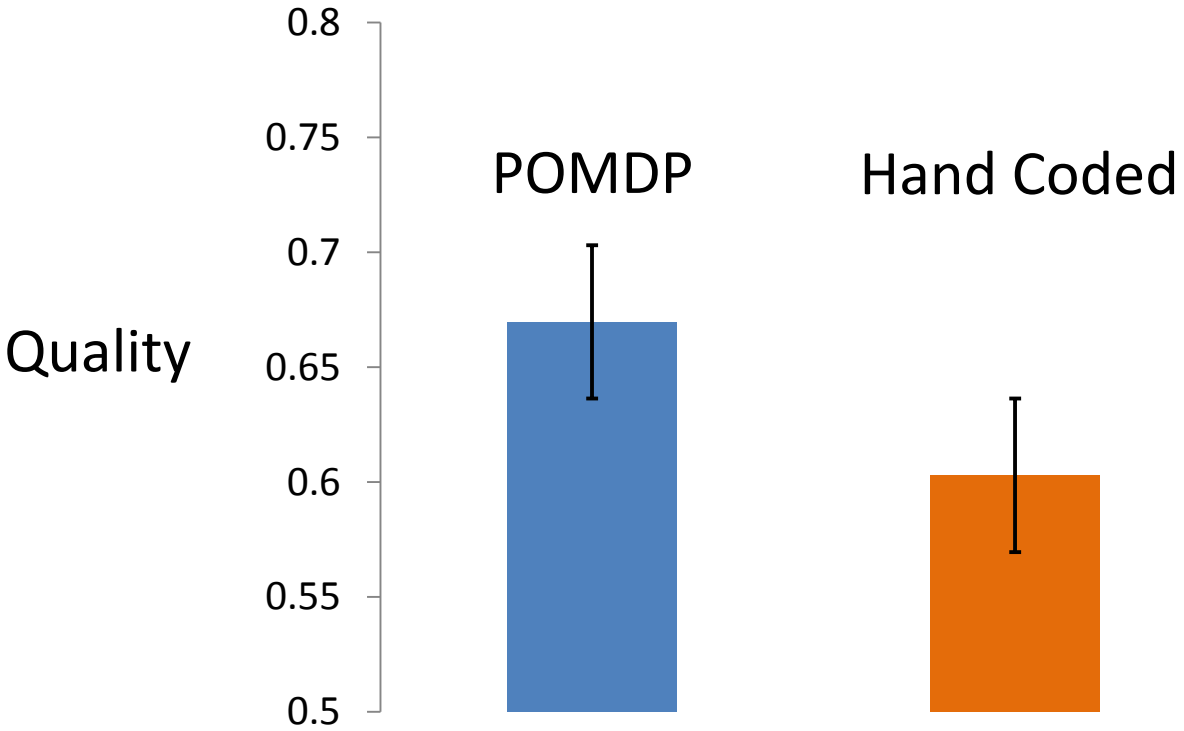
Clowder$_{iterative}$



Clowder$_{iterative}$



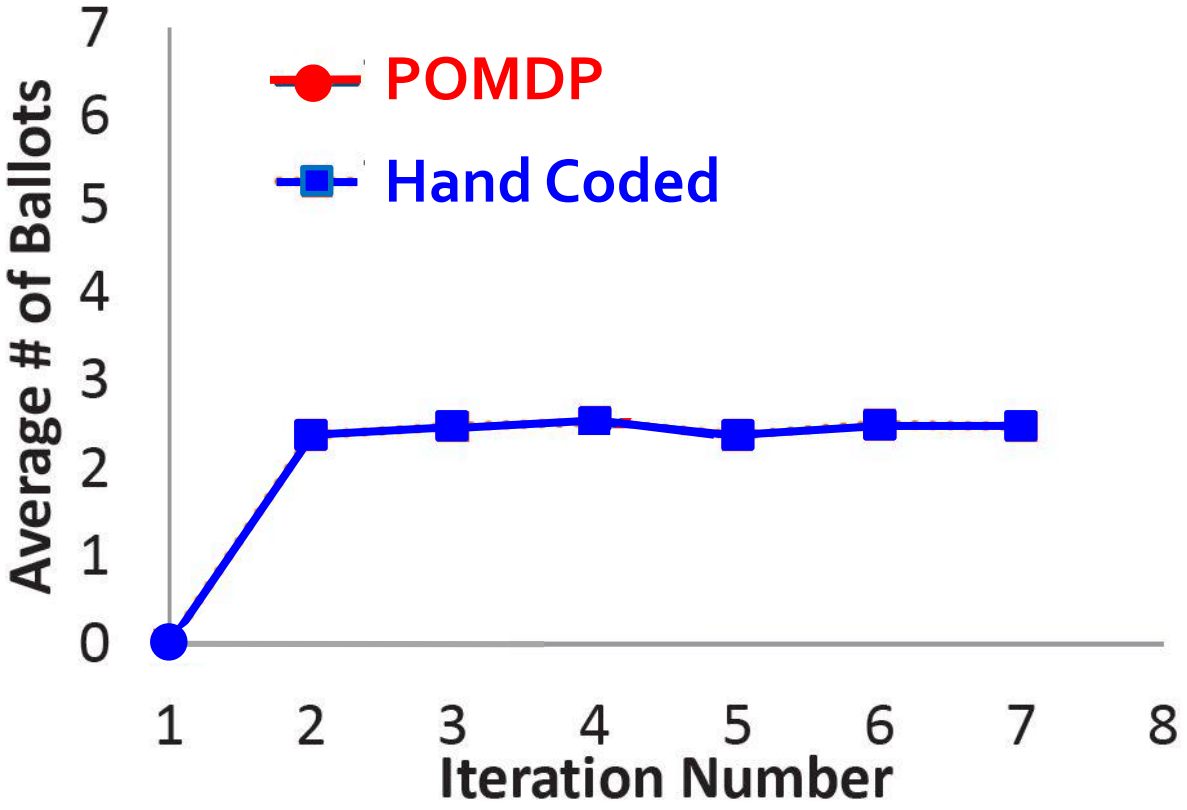
Comparison



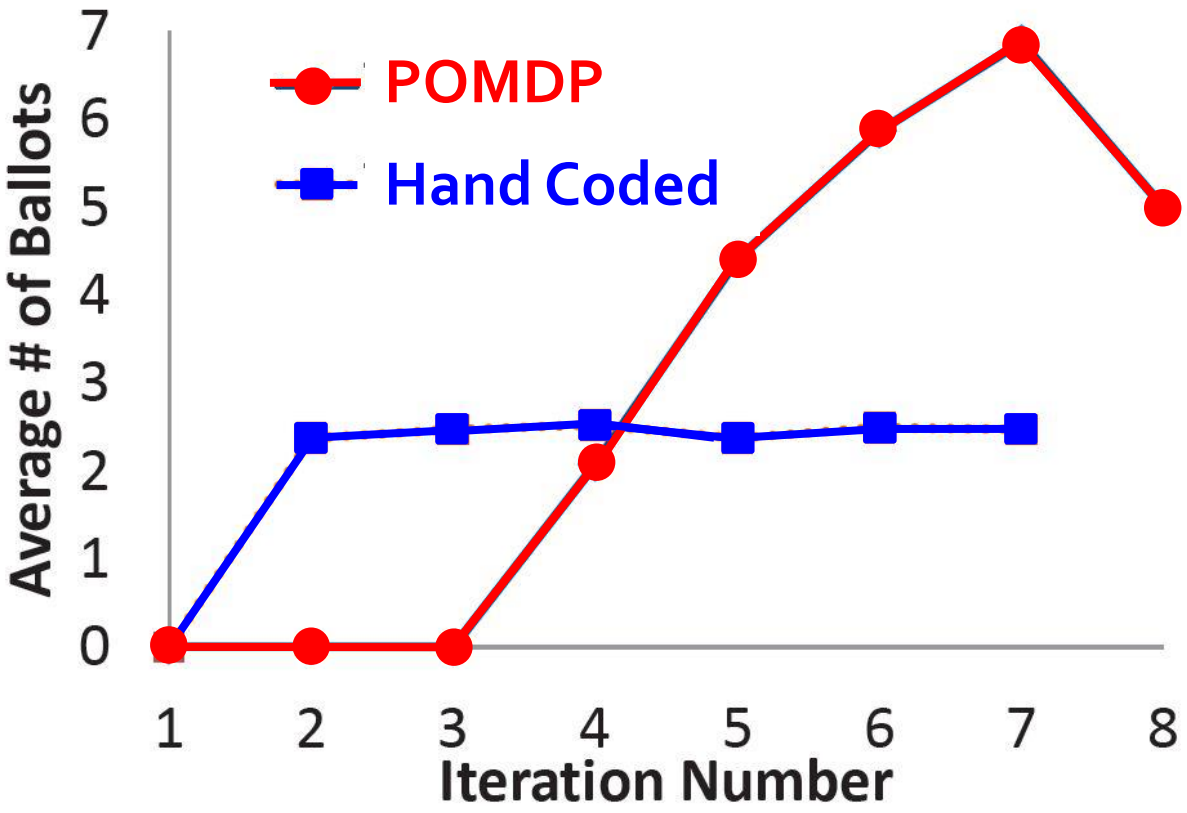
40 images, same average cost

Controlling quality: ***POMDP 30% cheaper***

Allocation of Human Labor



Human Labor Redirected



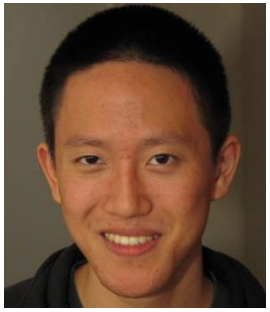
Summary

- Clowder can control much more complex tasks
  - Bayes net for representing and learning knowledge
  - POMDP for long-term control
- Unforeseen intelligent behavior
- Suggests that ideas may extend to generic workflows

Thanks

Dan Weld

Peng Dai



Chris Lin

Jonathan Bragg

Andrey Kolobov