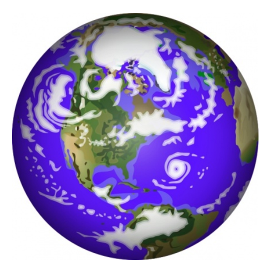
Planning to Control Crowd-Sourced Workflows

Daniel S. Weld
University of Washington

UW CrowdLab

dub

30,000' View



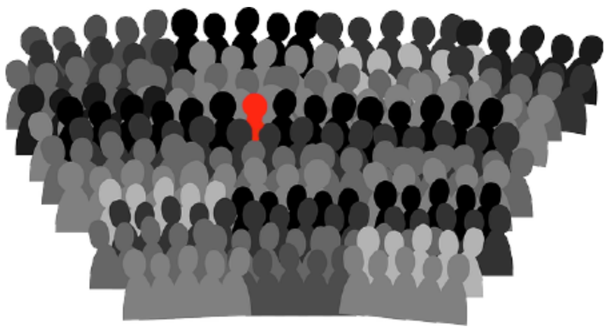
- Crowdsourcing is huge & growing rapidly
  - Virtual organizations
  - Flash teams with mixed human & machine members
- Automatic organization of work
  - Reduce labor required by 30-85%

UW CrowdLab

2 dub

Crowdsourcing

- Performing work by **soliciting effort** from many people
- **Combining the efforts** of volunteers/part-time workers (each contributing a small portion) to produce a large or significant result



Crowdsourcing Successes

tripadvisor — 190 M reviews of 4.4 M businesses

stackoverflow — Answers to 7.1 M prog. questions

Universal reference for anything

Citizen Science

ZOONIVERSE REAL SCIENCE ONLINE

800,000 volunteers – Hubble images
Discovered "Hanny's Voorwerp" black-hole
"Pea galaxies"

eBird

Audubon

Crowdsourced bird count & identification
Migration shift -> effect of climate change

Game to find 3D structure of proteins.
Solved 15 year outstanding AIDS puzzle

5

dub

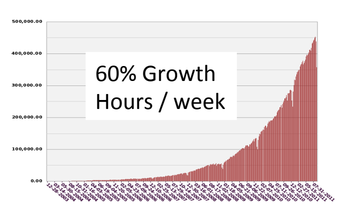
Labor Marketplaces

Will Grow to $5B by 2018 [Staffing Industry Analysts]

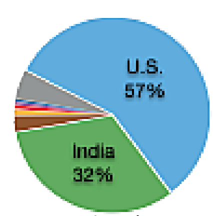
oDesk
Changing How the World Works.

- 2.7 million workers
- 540,000 requestors
- 35M hours worked in 2012

spare5



amazonmechanical turk
Artificial Artificial Intelligence



Nationality

UW CrowdLab

Charts from Panos Ipierotis' blog; phone from pixabay

dub

Example Job on Mechanical Turk

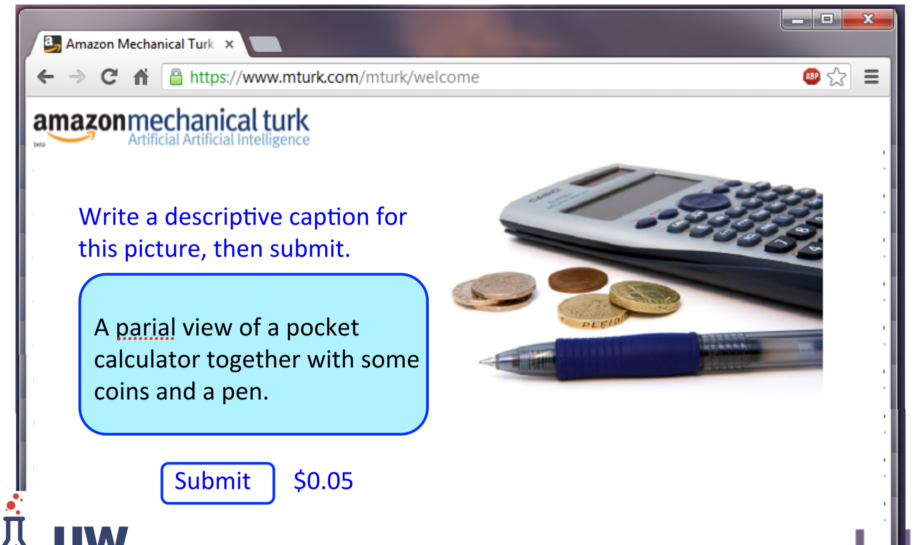


Figure from [Little et al. 2010]

Big Work from Micro-Contributions

- Challenges
  - Small work units
  - Reliability & skill of individual workers vary
- Therefore
  - Use a ***workflow*** to aggregate results & ensure quality
  - Manage workers with (unreliable) workers

Ex: Iterative Improvement

initial caption

UW CrowdLab

[Little et al, 2010] 9

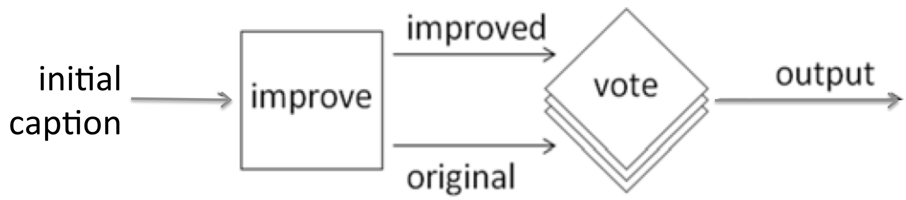
Ex: Iterative Improvement

initial caption → improve

UW CrowdLab

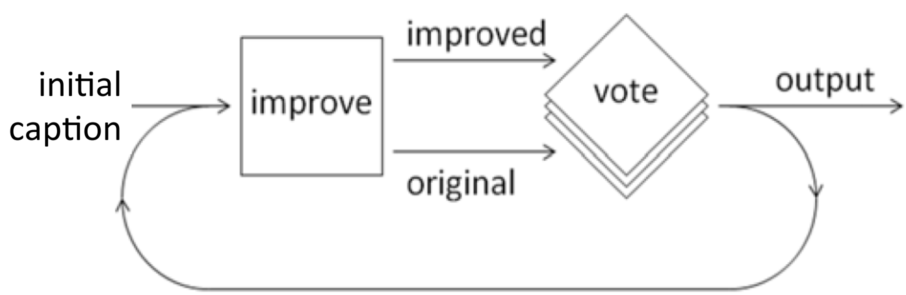
[Little et al, 2010] 10

Ex: Iterative Improvement



[Little et al, 2010] 11

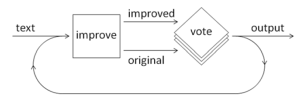
Ex: Iterative Improvement



[Little et al, 2010] 12

Iterative Improvement

[Little et al, 2010]



First version

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

Figure from [Little et al. 2010]

After 8 iterations

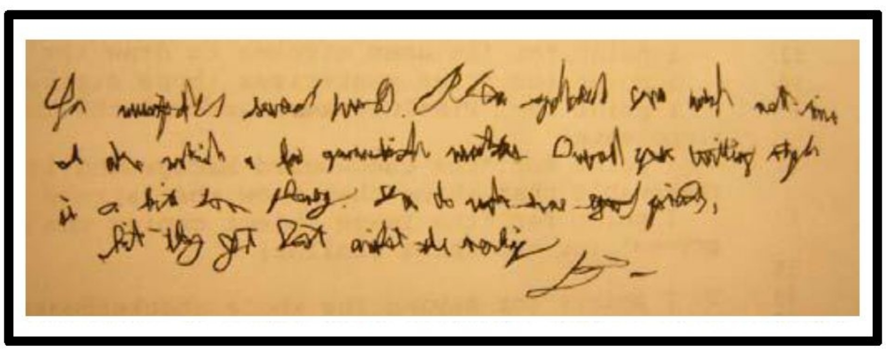
A CASIO multi-function, solar powered scientific calculator.

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.

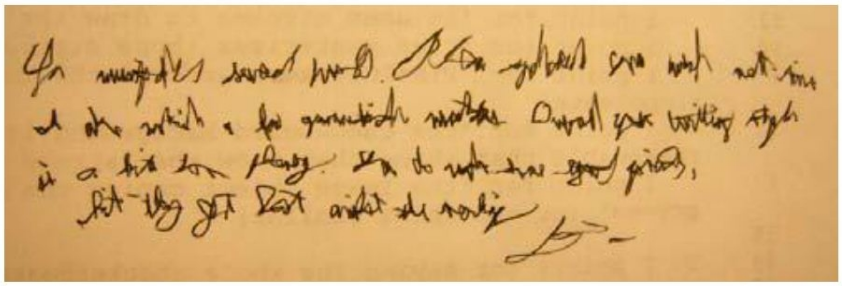
Seems to be a theme illustration for a brochure or document cover treating finance - probably personal finance.

UW CrowdLab 13 dub



[Little et al, 2010]

UW CrowdLab 14 dub

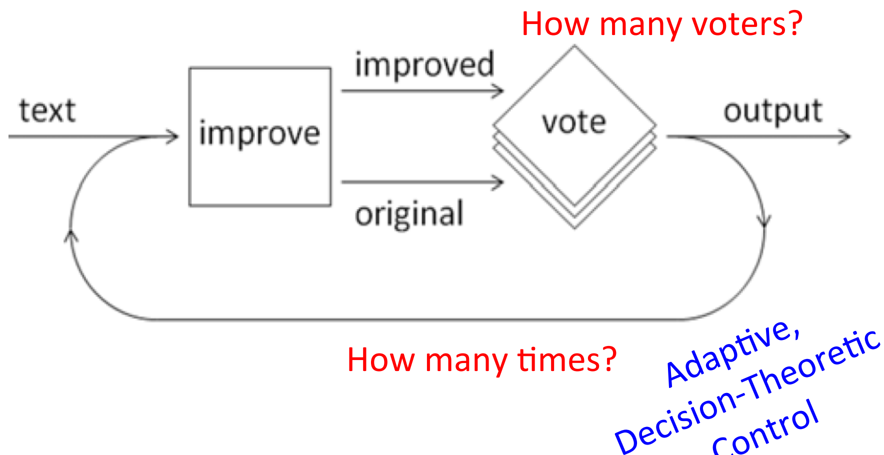


"You (misspelled) (several) (words). Please spellcheck your work next time. I also notice a few grammatical mis-takes. Overall your writing style is a bit too phoney. You do make some good (points), but they got lost amidst the (writing). (signature)"

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

[Little et al, 2010]

Workflow Control Problem



UW CrowdLab

16

Outline

✓Introduction
- Case Study: Controlling Iterative Improvement
- Case Study: Controlling Taxonomy Generation
- In Progress: Controlling ML Annotation

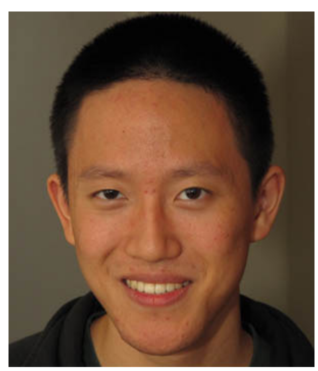
UW CrowdLab

17 dub

TurKontrol

POMDP Control of Iterative Improvement

Peng Dai

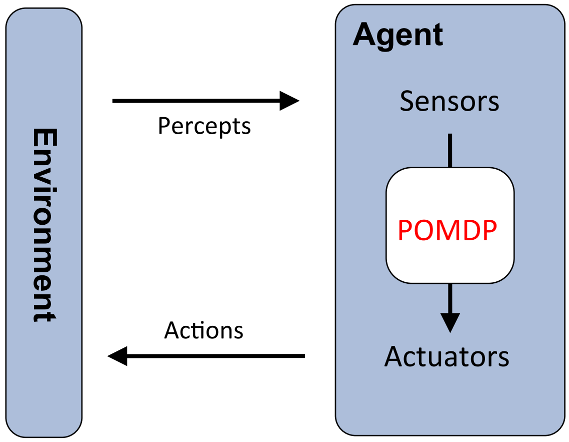


Chris Lin

Both co-advised with Mausam

dub

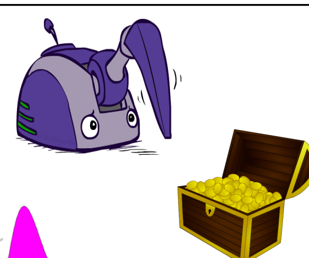
Artificial Intelligence 101



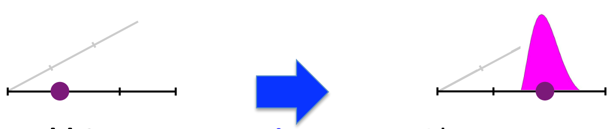
UW CrowdLab

19 dub

Markov Decision Process



Input:



World State
s = <x, y>

Actions
P(s' | s, a)
Cost c

Observe: Next State s' = <x', y'>
Reward = f(s, a, s')

Output:

Construct ***policy***, π : S→A, that chooses best action for each state
I.e., actions that ***maximize expected reward – costs*** over time

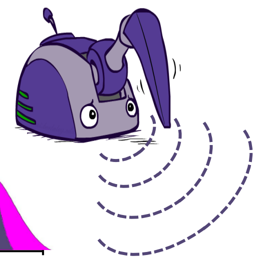
While ***learning*** action & reward probabilities
(Reinforcement learning)

UW CrowdLab

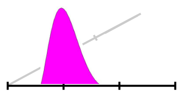
Figure from Dan Klein & Pieter Abbeel - UC Berkeley CS188: http://ai.berkeley.edu]
And Sarah Reeves (http://dear-theophilus.deviantart.com/)

20 dub

Partially-Observable
Markov Decision Process

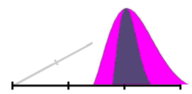


Input:



Belief State
P(s)

Actions
$P(s' \mid s, a)$
Cost c



Observe: Noisy Sensor = f(s')
Reward

Output:

Construct ***policy***, $\pi : S \rightarrow A$, that chooses best action for each state
I.e., actions that ***maximize expected reward – costs*** over time

While ***learning*** action & reward probabilities
(Reinforcement learning)

UW CrowdLab

Figure from Dan Klein & Pieter Abbeel - UC Berkeley CS188: http://ai.berkeley.edu]

dub

Solving the POMDP

Constructing the policy, π, to choose the best action

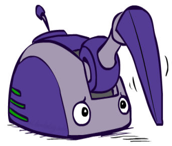
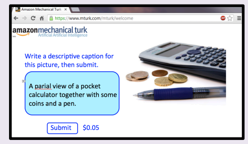
- Many algorithms
  - Point-based methods
  - UCT on discretized space
  - Lookahead search with beta distribution belief states

$$Q^*(s, a) = \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma \, \mathrm{Max}_a \, Q^*(s, a) \right]$$

- Exploration / exploitation problem
  - ε-greedy
  - UCB / Multi-armed bandit

UW CrowdLab

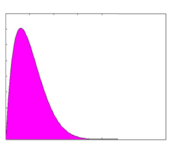
dub

From 		To 
(Hidden) **World State**	<x,y> coords	Quality Q_1, Q_2 $\in$(0,1)
Actions	Move Grasp	Improve caption task Vote best caption
Costs	Power used	$$ paid to workers
Reward		F(quality returned)

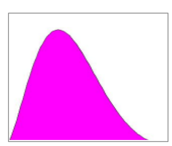
UW CrowdLab

Robot figure from Dan Klein & Pieter Abbeel - UC Berkeley CS188: http://ai.berkeley.edu.]

Belief State

P 

Quality$_{\alpha 1}$

P 

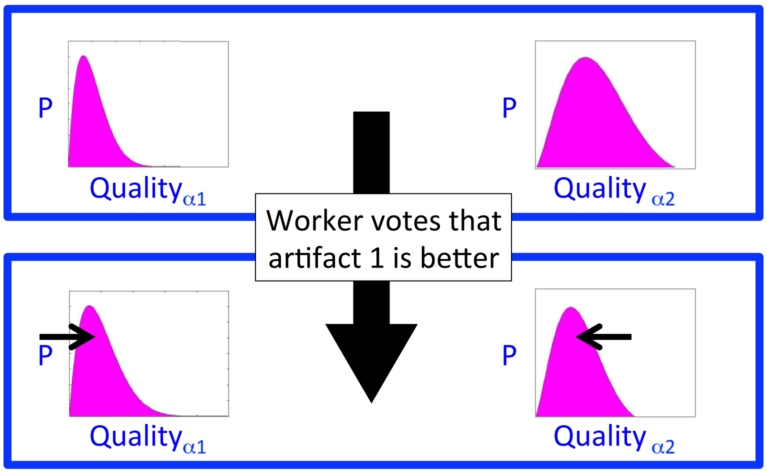
Quality $_{\alpha 2}$

UW CrowdLab

dub

Transition Model of Voting Action

Learned using Expectation Maximization

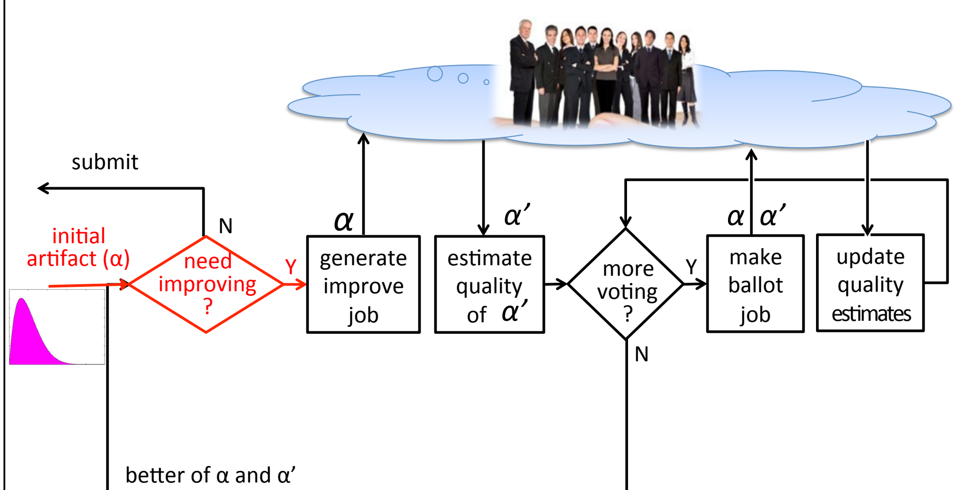


UW CrowdLab

29

dub

POMDP for Iterative Improvement

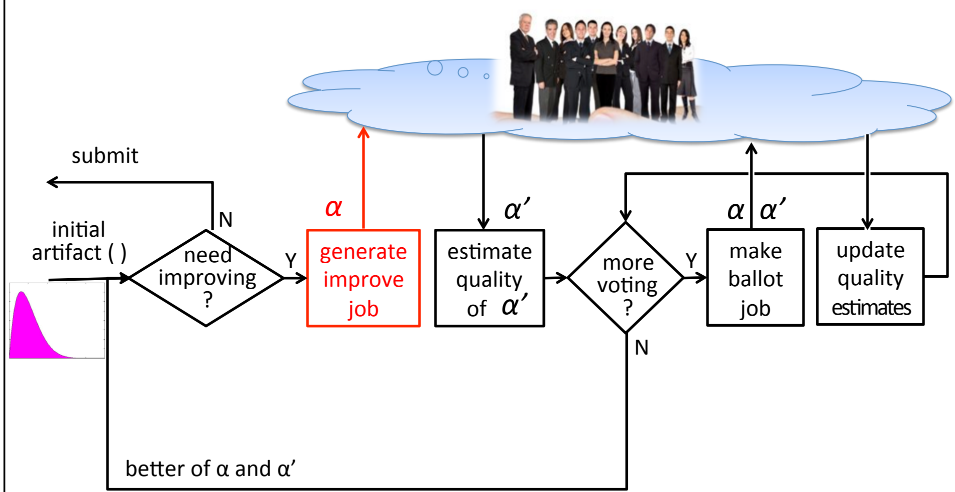


UW CrowdLab

35

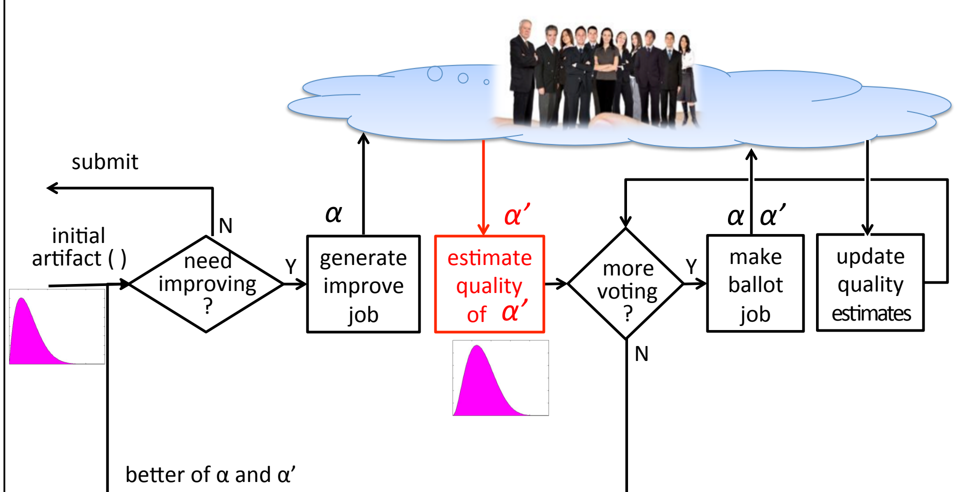
dub

POMDP for Iterative Improvement



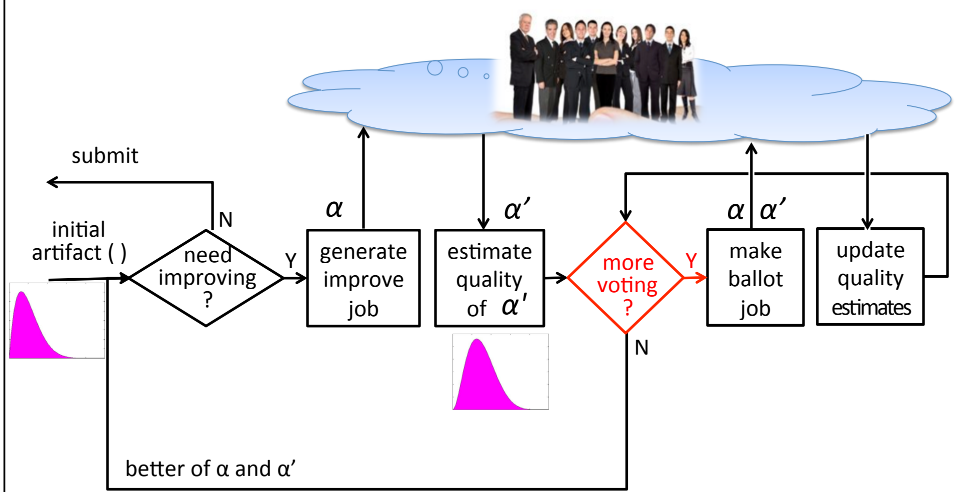
36

POMDP for Iterative Improvement



37

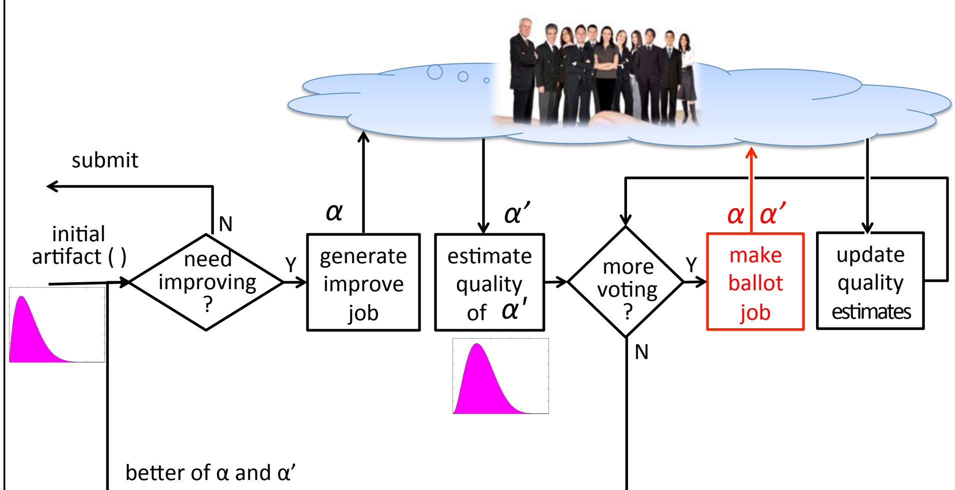
POMDP for Iterative Improvement



UW CrowdLab

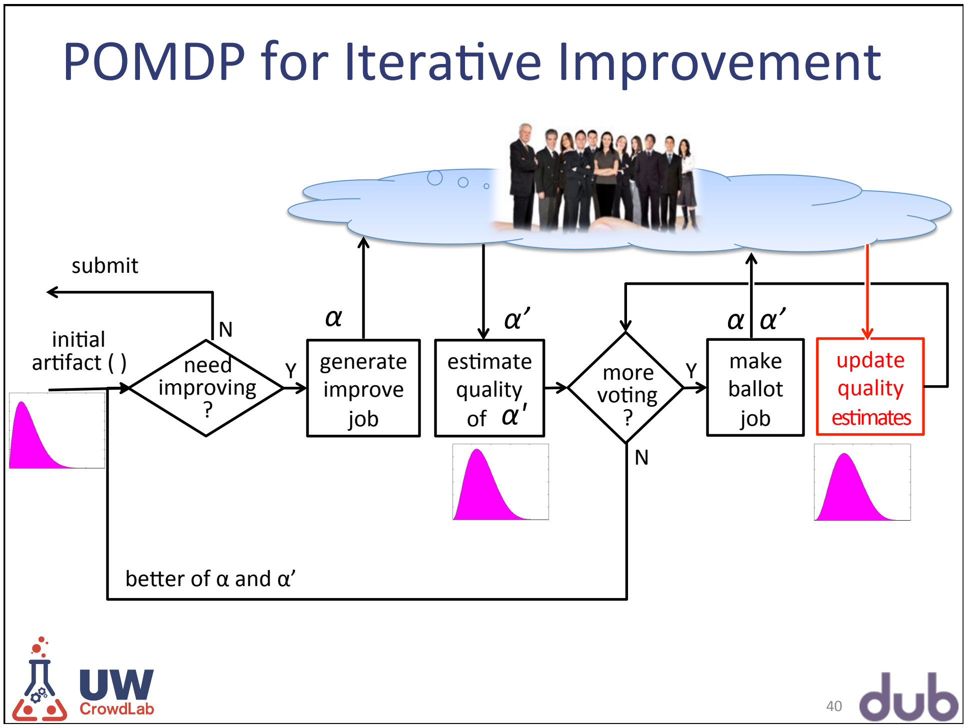
38

POMDP for Iterative Improvement

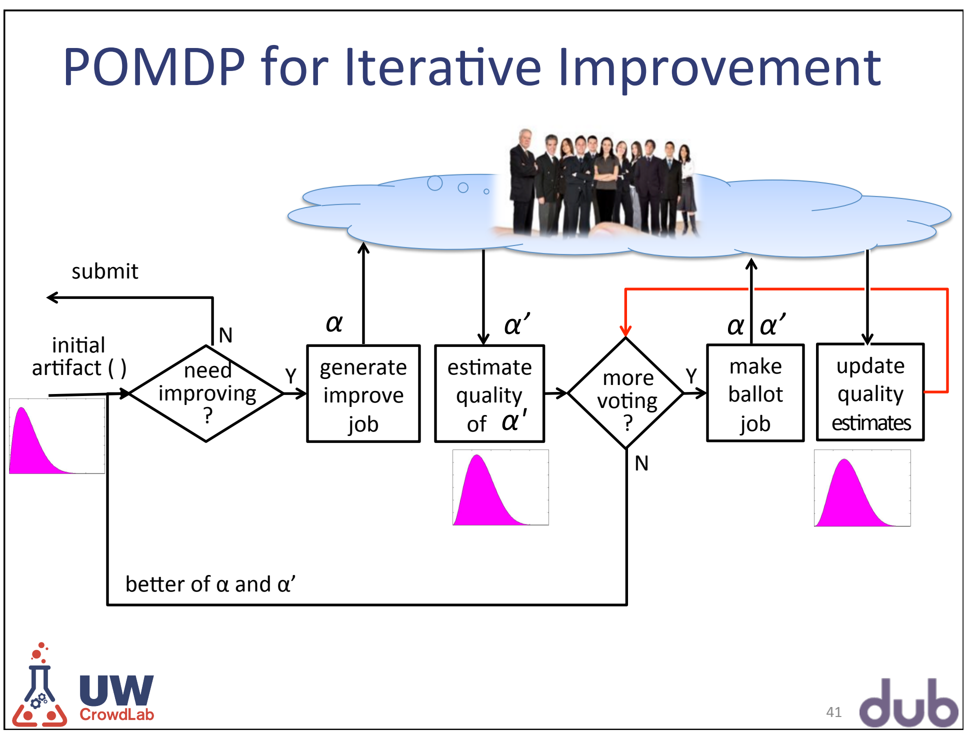


UW CrowdLab

39

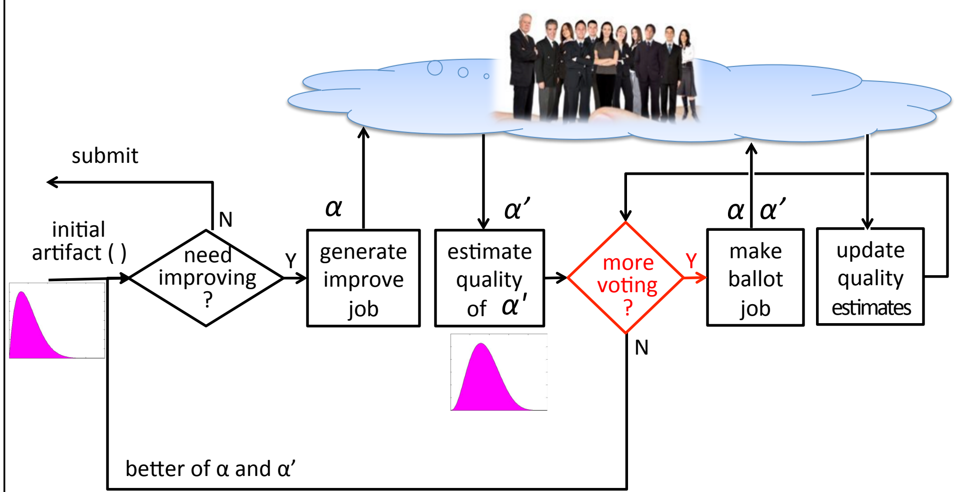


POMDP for Iterative Improvement
submit
initial artifact ( )
need improving ?
N
Y
α
generate improve job
α'
estimate quality of α'
more voting ?
Y
N
α α'
make ballot job
update quality estimates
better of α and α'
UW CrowdLab
40
dub

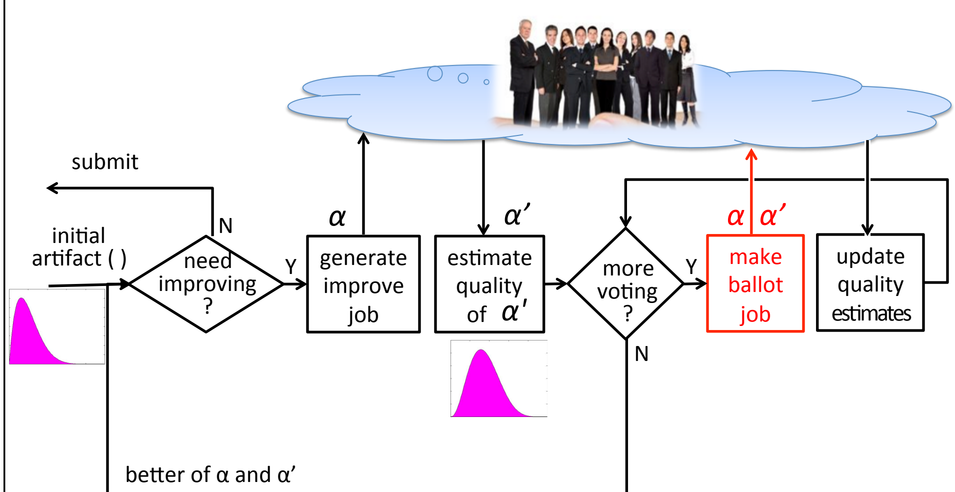


POMDP for Iterative Improvement
submit
initial artifact ( )
need improving ?
N
Y
α
generate improve job
α'
estimate quality of α'
more voting ?
Y
N
α α'
make ballot job
update quality estimates
better of α and α'
UW CrowdLab
41
dub

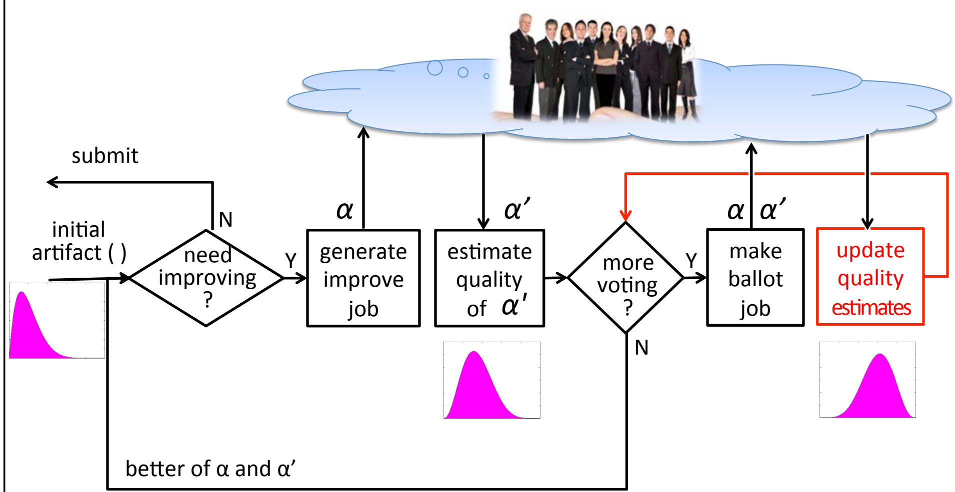
POMDP for Iterative Improvement



POMDP for Iterative Improvement



POMDP for Iterative Improvement

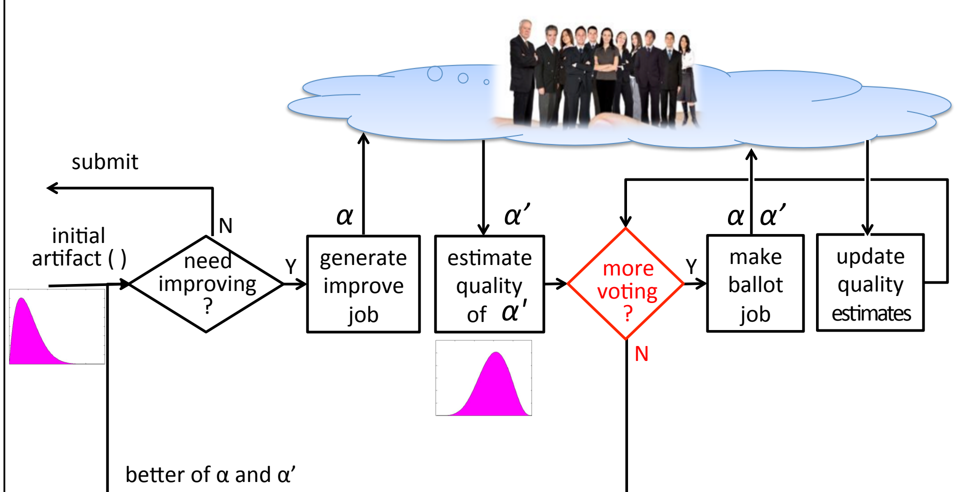


UW CrowdLab

44

dub

POMDP for Iterative Improvement

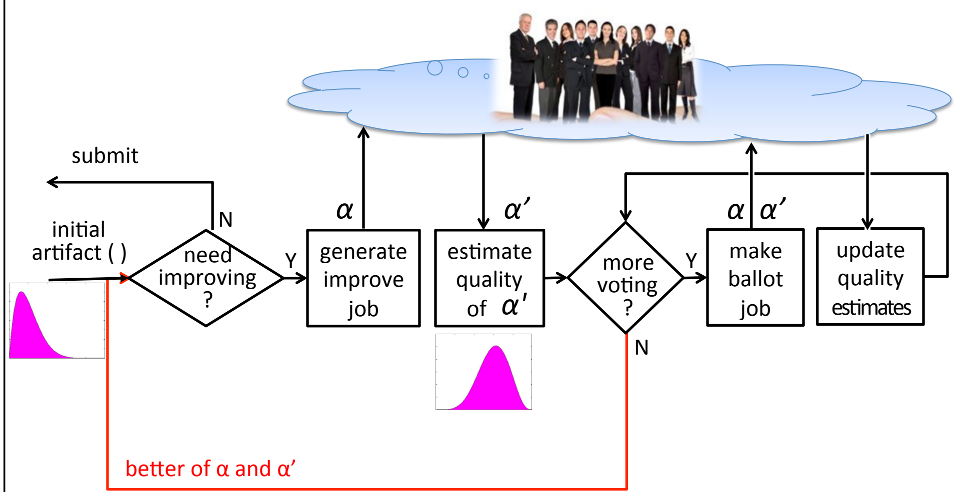


UW CrowdLab

45

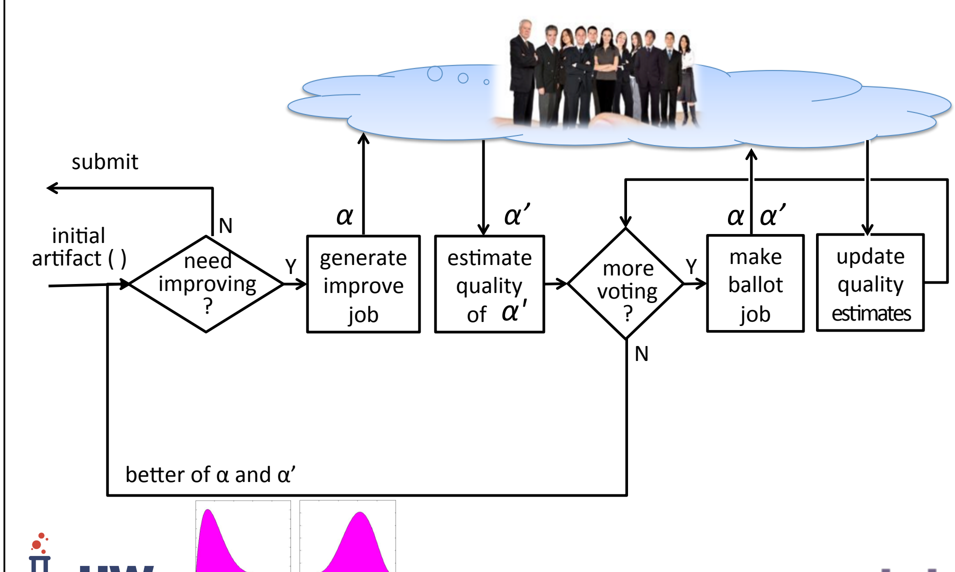
dub

POMDP for Iterative Improvement



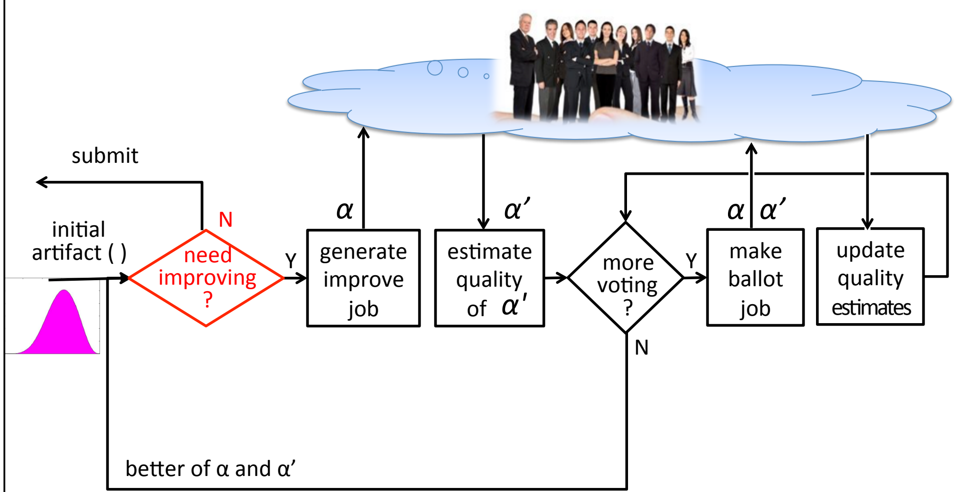
46

POMDP for Iterative Improvement



47

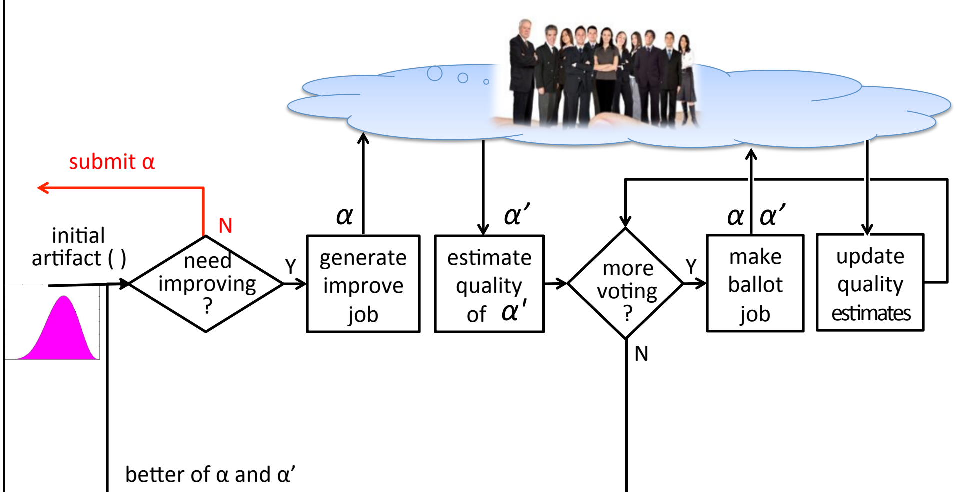
POMDP for Iterative Improvement



UW CrowdLab

48

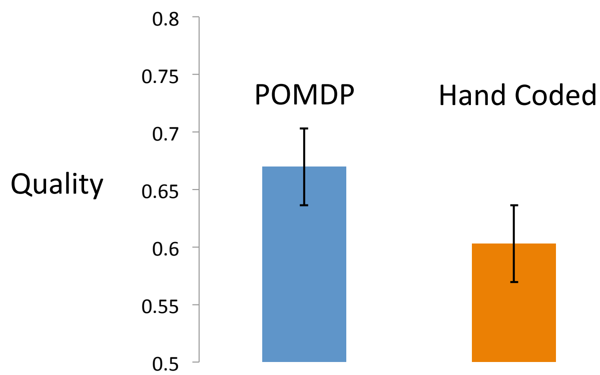
POMDP for Iterative Improvement



UW CrowdLab

49

Comparison



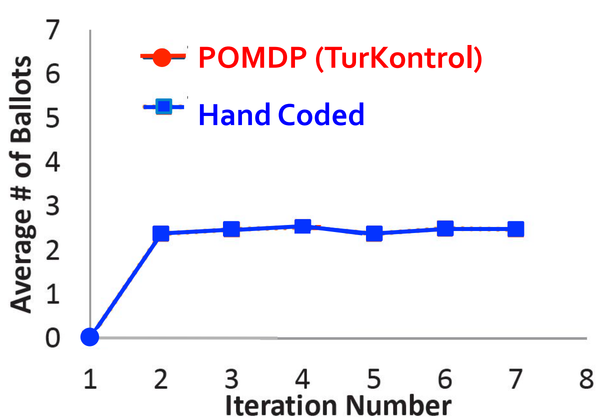
40 images, same average cost

Controlling quality: ***POMDP 30% less labor***

[Dai, Mausam & W, AAAI-11]
[Dai *et al.* AIJ 2013]

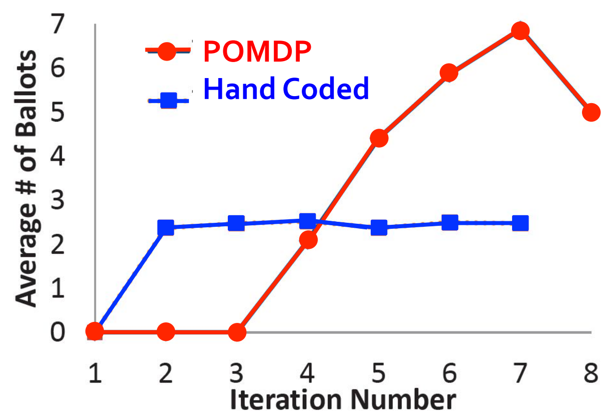
50

Allocation of Human Labor



51

Human Labor Redirected



UW CrowdLab

52

dub

Lessons So Far

- Reduced labor costs
- Improved quality
- POMDP planning
  - Update belief states about uncertain world
  - Model sensing actions
- Expectation maximization & prob inference

UW CrowdLab

53

dub

Outline

✓ Introduction
✓ Case Study: Controlling Iterative Improvement
- Case Study: Controlling Taxonomy Generation
- In Progress: Controlling ML Annotation

UW CrowdLab

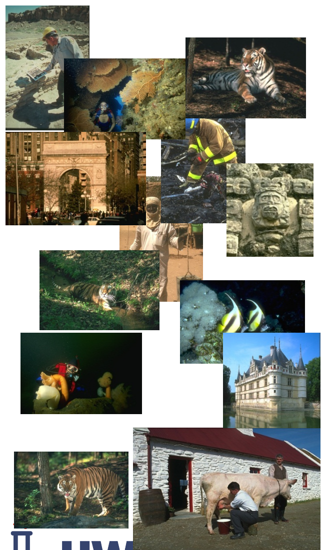
54 dub

Cascade
Crowdsourcing Taxonomy Creation

Lydia Chilton
Co-advised with James Landay

dub

Image Data Sets



- images (100)
  - field (10)
  - birds (6)
  - leopard (3)
  - tiger (3)
  - wolf (2)
- people (25)
  - worker (8)
  - hat (6)
  - world music (2
  - surfing (2)
- tree (18)
- rock formations (12)
- architecture (10)
- mountains (9)
- historical landmarks (7)
  - castle (2)
- lanscaping (5)
- winter landscapes (5)
- sculptures (5)
- polenesian culture (4)
- scuba-diving (4)
  - scuba diving (3)
- airplanes (2)
- boats (2)
- island (2)

UW CrowdLab

56

dub

Q&A Site Responses

- traveling (100)
  - air travel tips (49)
    - preparation for flying (38)
      - comfortable flying (13)
    - airport tips (26)
      - airport shortcuts (17)
    - flight (25)
      - flight comfort (11)
        - seating in airlines (5)
      - flight layovers (2)
    - in flight meals (6)
    - airport food (4)
    - best picks for airport and airline food (4)
    - where to sit on long flight (2)
    - membership discounts (2)
  - international phone usage (3)
    - international data plans (2)
  - insider tips (49)
    - making friends with locals (6)
    - airport amenities (4)

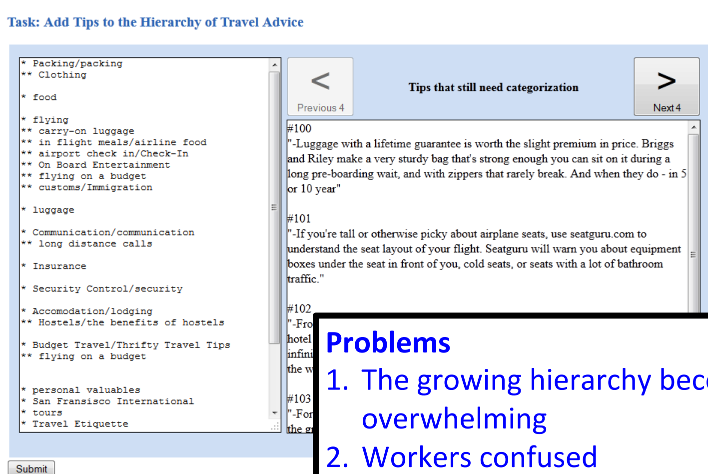
Crowdsourcing Taxonomy Generation
Is Hard !

- Good taxonomy requires a global perspective
- But workers see only a tiny fraction of data…?

UW CrowdLab

58

Iterative Improvement?



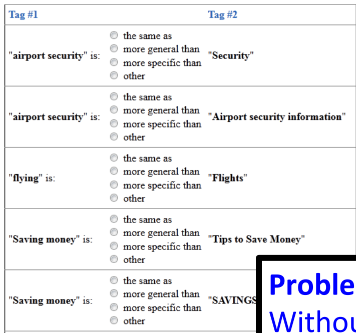
Problems
1. The growing hierarchy becomes overwhelming
2. Workers confused

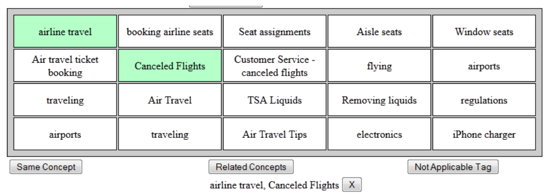
Lesson: Decompose the task into smaller steps

UW CrowdLab

59

Initial Approach 2: Category Comparison





Problem
Without context it's hard to judge relationships:
- **TSA liquids** *vs.* **removing liquids**
- **Packing** *vs.* **what to bring**

Lesson: Don't compare abstractions

UW CrowdLab

60 dub

Cascade Overview

[Chilton *et al.*, CHI-13]

Use the ***crowd*** to:

1. Generate category names
2. Select the best categories
3. Place the data into the best categories

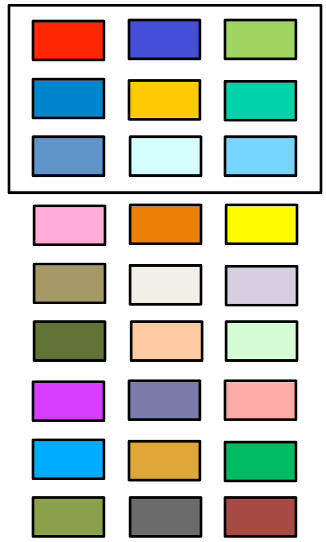
Use ***machines*** to:

4. Infer global structure of categories

UW CrowdLab

61 dub

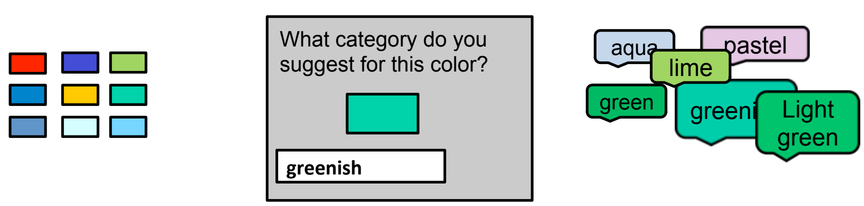
Example Input: 100 Random Colors
Step 0. Sample Data



62

Step 1. Generate Categories

For each color **Task** **Crowd responses**

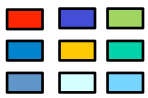


This generates an initial set of category names.

63

Step 2. Select Best Categories

For each color



Task

What is the best category for this color?

Category	Best?
Aqua	☐
Greenish	☑
Lime	☐
Pastel	☐

Crowd responses

Category	Votes
Aqua	1/5
Greenish	4/5
Lime	0/5
Pastel	0/5

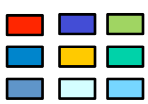
An early filter for spam and vague categories

UW CrowdLab

64 dub

Step 3. Label Data

For each color and category



Categories
Green
Greenish
Yellow
Pink

Task

What categories does this belong to?

Category	Fits	Doesn't Fit
Green	☑	☐
Greenish	☑	☐
Yellow	☐	☑
Pink	☐	☑

Crowd responses

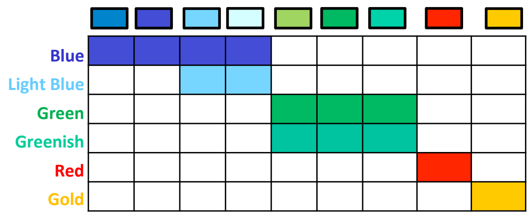
Category	Votes
Green	4/5
Greenish	5/5
Yellow	1/5
Pink	0/5

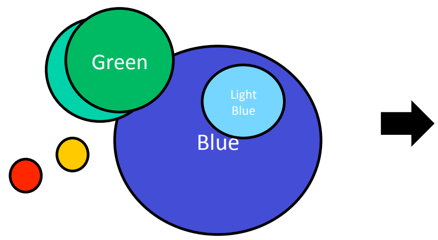
This determines category membership.

UW CrowdLab

65 dub

Step 4. Global Structure Inference





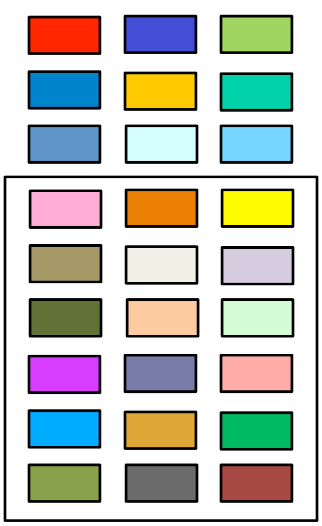
Blue:
Light Blue:
Green:
Other:

Determine parent/child relations; eliminate duplicates.

UW CrowdLab

dub

Finally, … Recurse



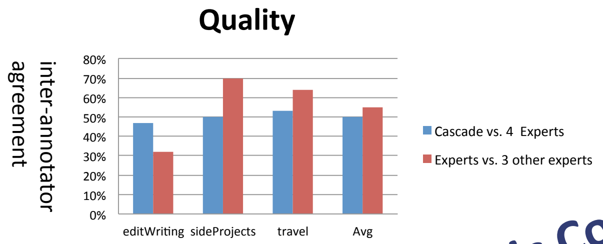
Blue:
Light Blue:
Green:
Other:

May lead to new tags & recomputing taxonomy

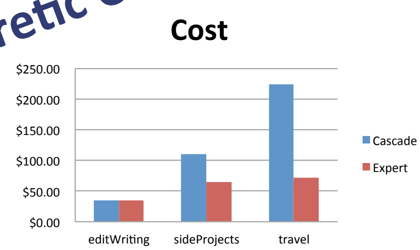
UW CrowdLab

dub

Evaluation



Decision-Theoretic Control!



UW CrowdLab

69

dub

Deluge

(Decision-Theoretic Control of Cascade)

Jonathan Bragg

Co-advised with Mausam

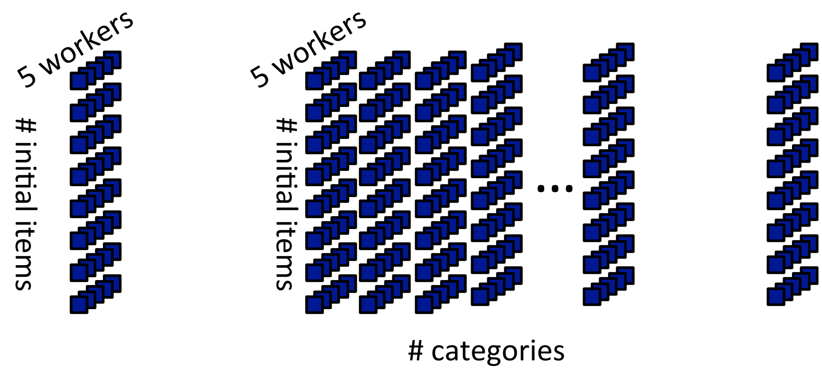
UW CrowdLab

70

dub

Why is Cascade Expensive?

Generate SelectBest Categorize

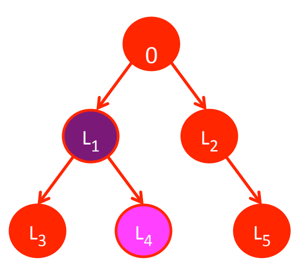


But do we really need all these questions?
What's the best order to ask them?

UW CrowdLab 72 dub

POMDP Model Agent Belief State

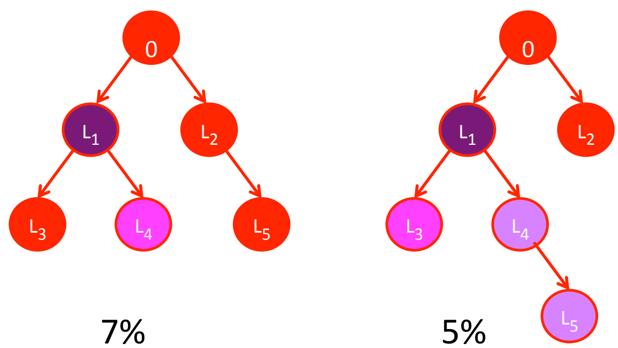
World state = taxonomy & labels applying to item



UW CrowdLab 73 dub

POMDP Model Agent Belief State

Belief state includes … distribution over taxonomies
label probabilities for item



Etc.

7% 5% …

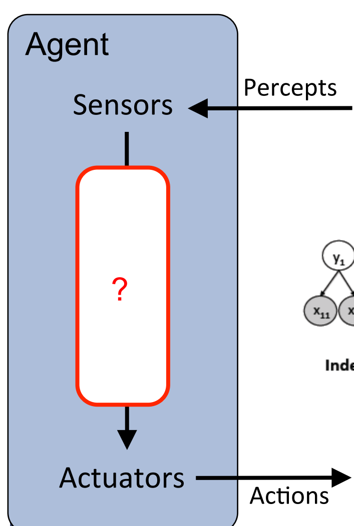
Learn & refine taxonomy during execution
Too complex for off-the-shelf POMDP solver

UW CrowdLab

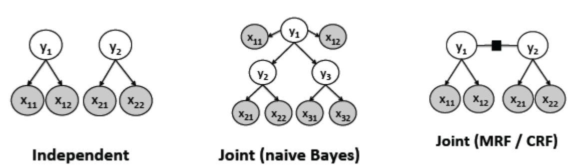
74 dub

Decision Cycle for New Item

[Bragg, Mausam & W HCOMP-13]



Probabilistic inference to update
- Posterior probabilities
- Co-occurrence model for labels
- Worker accuracy

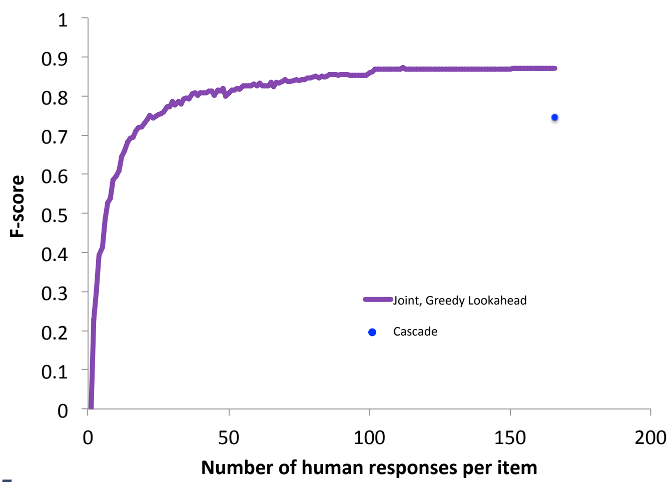


Ask about label with max VOI

UW CrowdLab

76 dub

Performance of Decision-Theoretic Model

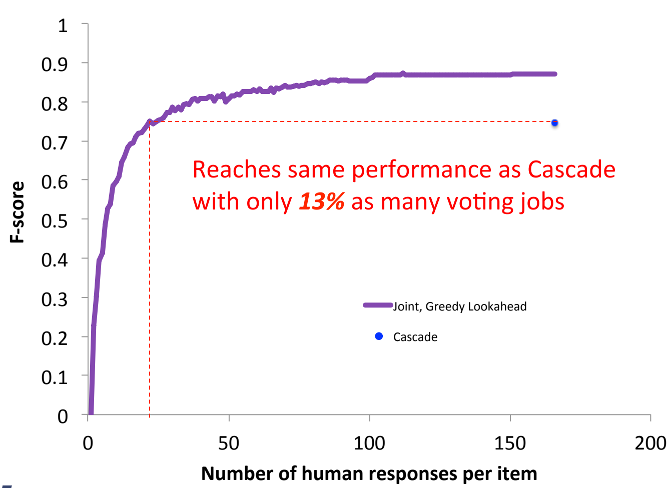


UW CrowdLab

77 dub

Performance of Decision-Theoretic Model

Now crowd is cheaper than experts!



UW CrowdLab

78 dub

Lessons So Far

- Decision-theoretic planning
  - Probabilistic inference
  - Expectation maximization
- Reduced labor & improved quality
  - Iterative Improvement
  - Taxonomy Generation
  - ***???***

UW CrowdLab

79 dub

Outline

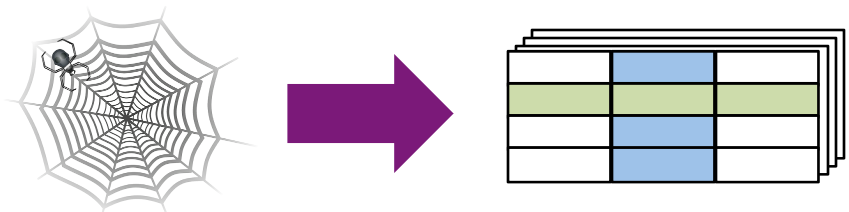
✓ Introduction
✓ Case Study: Controlling Iterative Improvement
✓ Case Study: Controlling Taxonomy Generation
- In Progress: Controlling ML Annotation

UW CrowdLab

80 dub

Information Omnivore Project

- Large Scale Information Extraction



- Train via 2 kinds of Weak Supervision
  - **Align Corpus to Background Knowledge Base**
    [Wu & W CIKM-07; … Koch *et al.* EMNLP-14]
  - **Identify & Extract Events from Newswire**
    [Zhang & W EMNLP-13; Zhang, Soderland & W TACL-15]

NewsSpike

UW CrowdLab

81

dub

Information Omnivore Project

- Augment with Crowdsourced Annotations
  - Eg: "Calling himself Guccifer, Marcel-Lehel Lazar rampaged through the email accounts of rich and powerful Americans…"
    AliasOf(p, p)
  - **For improved machine learning performance**

- Train via **Semi**-Distant Supervision
  - **Align Corpus to Background Knowledge Base**
    [Wu & W CIKM-07; … Koch *et al.* EMNLP-14]
  - **Identify & Extract Events from Newswire**
    [Zhang & W EMNLP-13; Zhang, Soderland & W TACL-14]

NewsSpike

UW CrowdLab

82

dub

Observation

- Vast proportion of micro-task crowdsourcing… is used to create training data for ML classifiers
  - Chris Caliston-Burch (UPenn) $250,000 on MTurk
  - LDC: 44 FT employees ***just*** creating NLP training data
  - Google, MSFT – internal CS: each larger than MTurk
- Common approach
  - Get two humans to annotate
  - If they agree, … done
  - Else recruit a third to arbitrate

"2/3 Relabeling"

j/k Relabeling

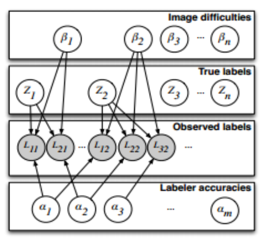
UW CrowdLab

83 dub

$$U(p(\beta_r|y_{i',l'})) = \left\|E_{p(\beta_r)}(\beta_r) - E_{p(\beta_r|y_{i',l'})}(\beta_r)\right\|_2 \quad (12)$$

$$\approx \left\|E\left(\frac{1}{S-1}\sum_{s=2}^{S} Z_r^{s-1\top}\left[(\gamma|\gamma^{s-1}, Z^{s-1}) - (\gamma_{(i',l')}|\gamma^{s-1}, Z^{s-1})\right]\right)\right\|_2. \quad (13)$$

$$\begin{aligned} Q(\boldsymbol{\alpha},\boldsymbol{\beta}) &= E\left[\ln p(\mathbf{l},\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\beta})\right] \\ &= E\left[\ln \prod_j \left(p(z_j)\prod_i p(l_{ij}|z_j,\alpha_i,\beta_j)\right)\right] \\ &\quad \text{since } l_{ij} \text{ are cond. indep. given } \mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta} \\ &= \sum_j E\left[\ln p(z_j)\right] + \sum_{ij} E\left[\ln p(l_{ij}|z_j,\alpha_i,\beta_j)\right] \end{aligned}$$



$$p(z|L,\theta) = \int p(z,q|L,\theta)dq = \prod_{j\in[M]}\int_0^1 p(q_j|\theta)q_j^{c_j}(1-q_j)^{\gamma_j-c_j}dq_j \stackrel{def}{=} \prod_{j\in[M]}\psi_j(z_{\mathcal{N}_j}), \quad (4)$$

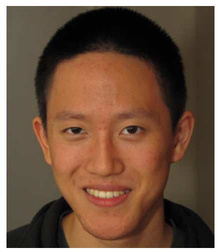
[Dawid *et al* 79, Whitehill *et al* 09, Welinder *et al* 10, Raykar *et al* 10, Karger *et al* 11, Kajino *et al* 12, Baba *et al* 13, Liu *et al* 12, etc, etc…]

UW CrowdLab

84 dub

[Lin, Mausam & W HCOMP-14]

How should one best spend a fixed annotation budget… ***when training an ML classifier?***



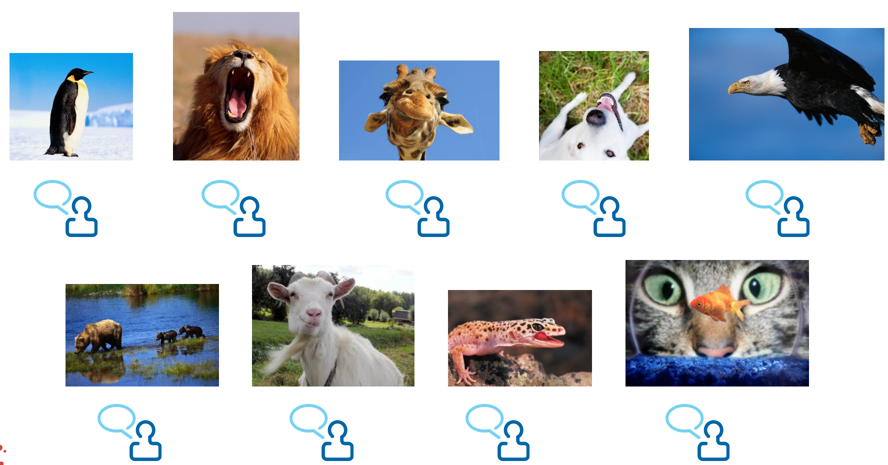
UW CrowdLab

85

dub

Unilabel?

9 examples with labels that are 75% accurate?

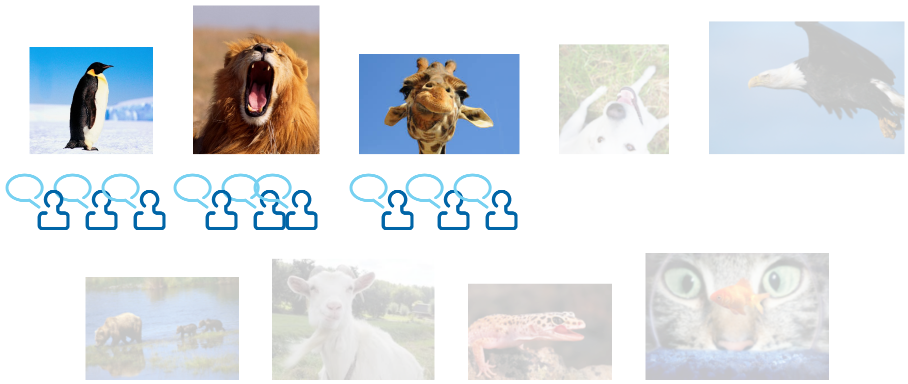


UW CrowdLab

dub

2/3 Relabel?

3 examples each with 3 labels – consensus 84% accurate?

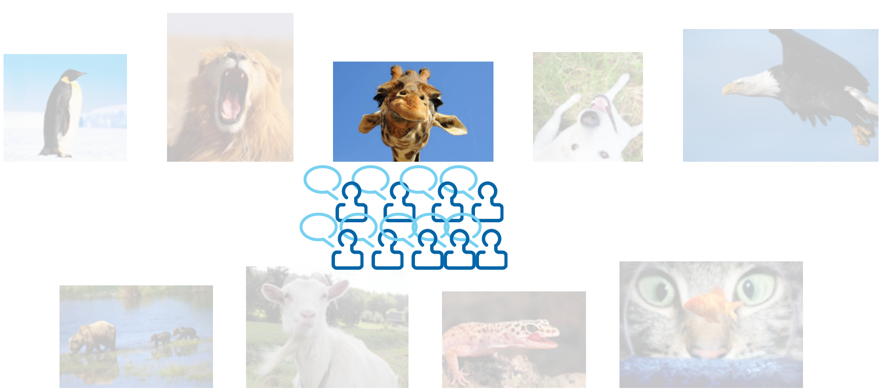


UW CrowdLab

dub

Or Even?

1 example with 9 labels – consensus 98% accurate?



UW CrowdLab

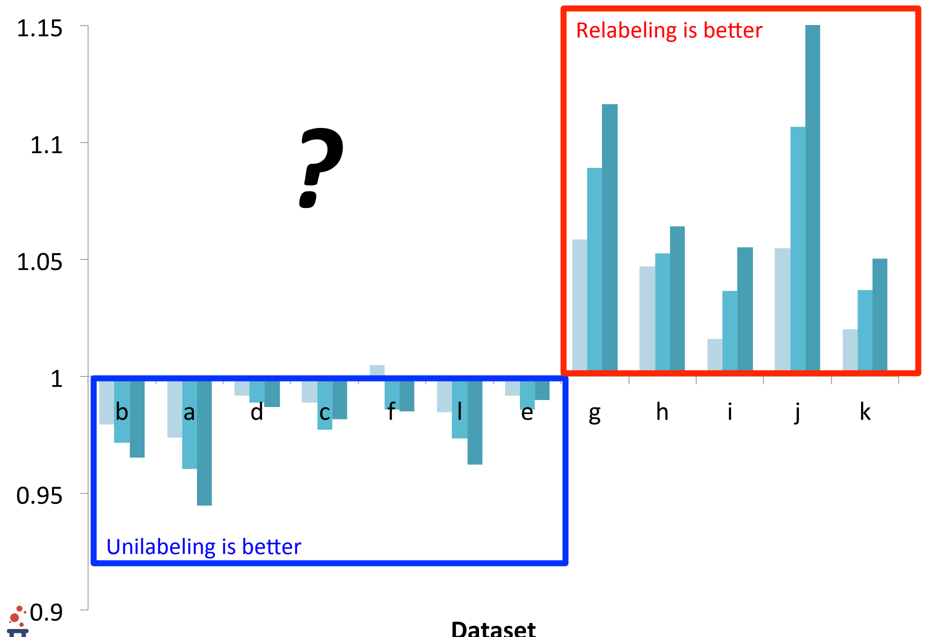
dub

Existing Data Sets?

Dataset	# Features	# Examples
(a) Breast Cancer	9	699
(b) Bank Note Authentication	4	1372
(c) Seismic Bumps	18	2584
(d) EEG Eye State	14	14980
(e) Sonar	60	208
(f) Breast Cancer Diagnostic	30	569
(g) Hill-Valley	100	606
(h) Hill-Valley with Noise	100	606
(i) Internet Ads	1558	2359
(j) Gisette	5000	6000
(k) Farm Ads	54877	4143
(l) Spambase	57	4601

UW CrowdLab

dub



UW CrowdLab

dub

Factors that Affect Relabeling Efficacy

Inductive Bias of Classifier

“Strong” → limited expressiveness

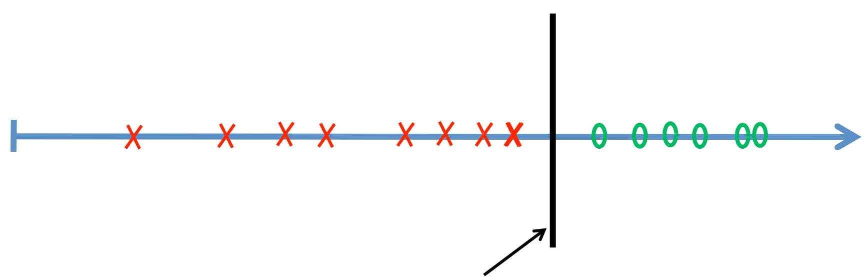
“Weak” → can learn many different concepts

Worker Accuracy

Budget

UW CrowdLab

If Data was Clean

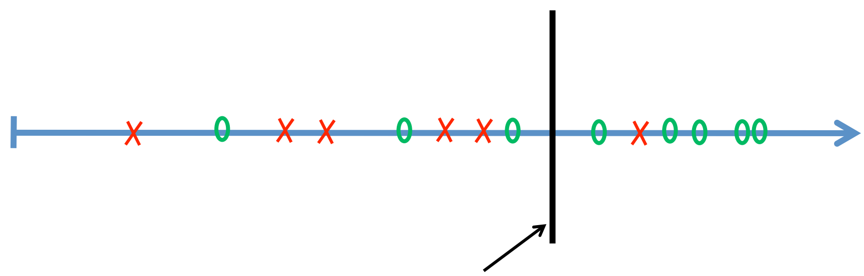


True Concept: 65 and older -> “Senior Citizen”

UW CrowdLab

92

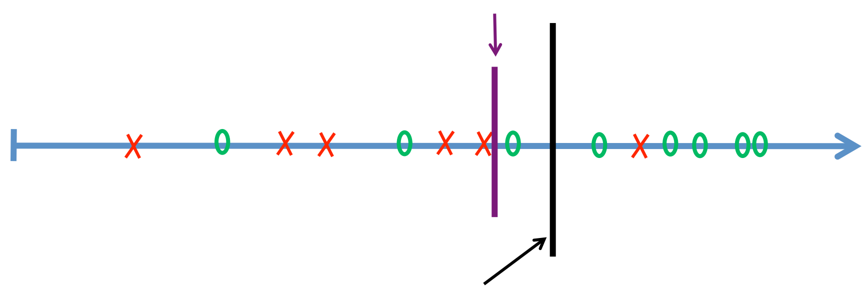
With Noisy Annotation



True Concept: 65 and older -> “Senior Citizen”

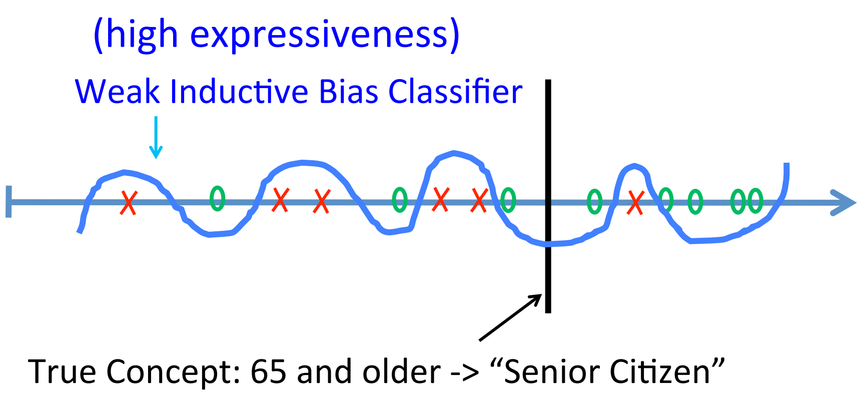
(low expressiveness)

Strong Inductive Bias Classifier



True Concept: 65 and older -> “Senior Citizen”

Overfitting to Noise



UW CrowdLab

95

dub

Conjecture

- Relabeling more important for classifiers with weak inductive bias

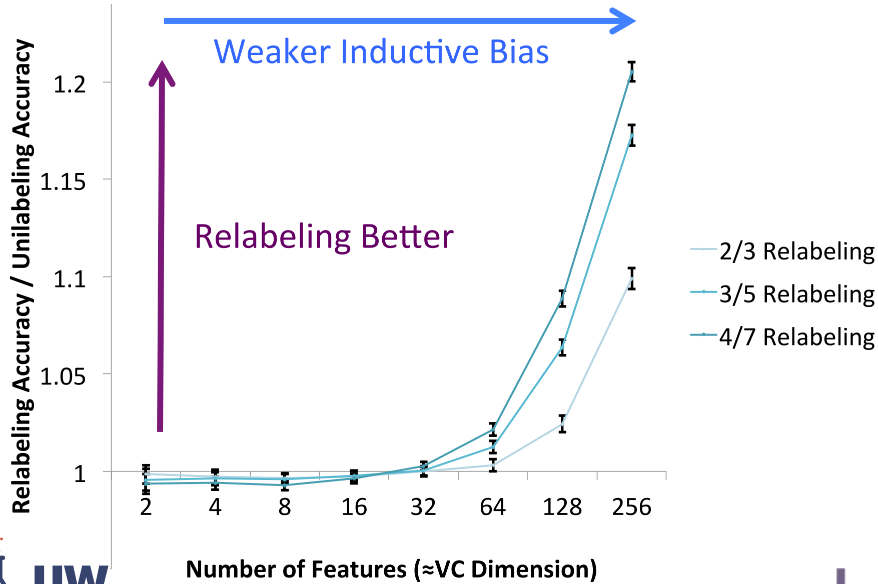
  (e.g., in domains with myriad features)

UW CrowdLab

96

dub

Experiments on Synthetic Data

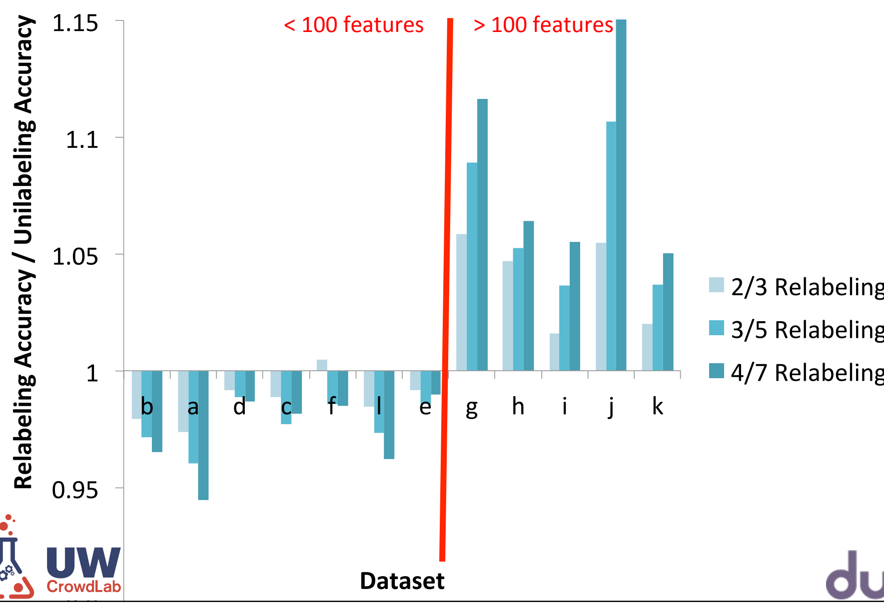


UW CrowdLab

98

dub

Revisiting the Real Data



UW CrowdLab

dub

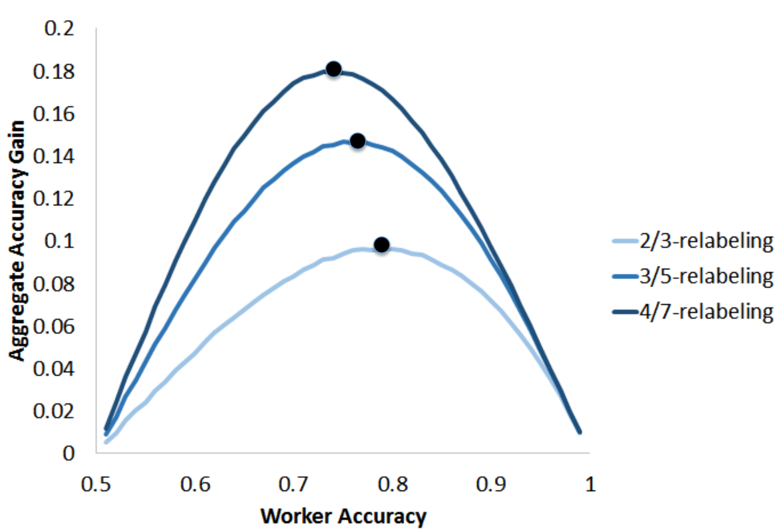
Factors that Affect Relabeling Efficacy

- Inductive Bias of Classifier
- Worker Accuracy

UW CrowdLab

100

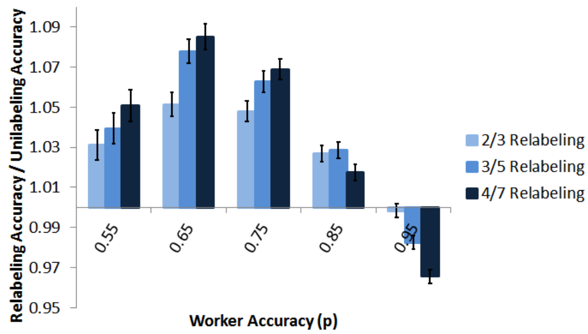
Accuracy of **Training Data**



UW CrowdLab

Factors that Affect Relabeling Efficacy

- Inductive Bias of Classifier
- Worker Accuracy



Results on simulated Gaussian data, fixed dimensionality = 50

UW CrowdLab

dub

Factors that Affect Relabeling Efficacy

- Inductive Bias of Classifier
- Worker Accuracy
- Budget

Future Work

- Relax Assumptions
- Complete Decision-Theoretic Control

UW CrowdLab

103

dub

Outline

✓ Introduction
✓ Case Study: Controlling Iterative Improvement
✓ Case Study: Controlling Taxonomy Generation
- Future Challenges

UW CrowdLab

104 dub

Other Challenges

- Usually assume ***workers choose*** job to perform

amazonmechanical turk
beta
Artificial Artificial Intelligence

- What if employer can ***assign*** jobs to best workers?
  – Google internal crowdsourcing
    - Street-view/maps, knowledge graph, search relevance
    - Task routing (expert / novice) in citizen science

UW CrowdLab

105 dub

Matching Jobs to Workers

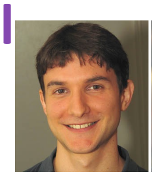
- Set of jobs, each with difficulty
- Set of workers, each with
  - Skill
  - Capacity (bound on # jobs)
  - Independent errors (conditioned on difficulty)
- Minimize overall error *wrt* fixed budget

- Knapsack?
  - "Pack" jobs with workers

UW CrowdLab

106 dub

Unknown Difficulty v **Skill**


Jonathan Bragg


Andrey Kolobov

- If skill levels are known…
  - Assigning unknown problem is like MAB "arm"
  - Once find hard problem (workers disagree), add expert
- If difficulty is known…
  - Assigning unknown worker is like MAB "arm"
- Exploration / Exploitation Tradeoff
  - [Bragg, Kolobov, Mausam & W HCOMP-14]

UW CrowdLab

107 dub

Additional Challenges

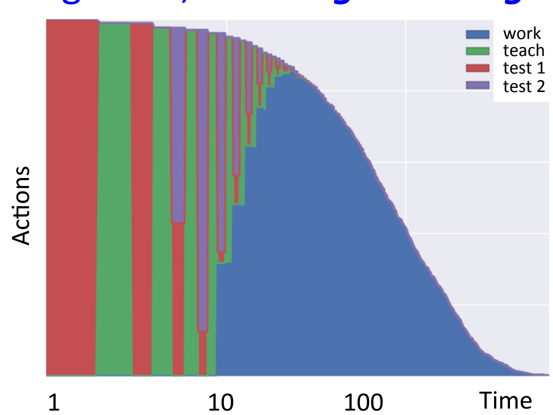
- Balancing worker desires w/ central needs
  - Frenzy [Chilton et al. CHI-14]

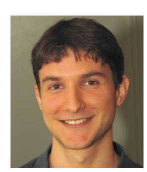
UW CrowdLab

108

Additional Challenges

- Balancing worker desires w/ central needs
- Optimizing for time
- Balancing ***work***, ***teaching*** & ***testing***





Jonathan Bragg

UW CrowdLab

109

Additional Challenges

- Balancing worker desires w/ central needs
- Optimizing for time
- Interleaving work, education & testing
- Workers improving job instructions
- Aggregation when majority is wrong
  - Bayesian truth serum
  - MicroTalk – focused argumentation

UW CrowdLab

dub

Related Work

- DT Crowdsourcing / Active Learning with Noise
  - GalaxyZoo – [Kamar & Horvitz 2012]
  - BBMC – [Wauthier & Jordan 2011]
  - ITS – Poppovic & Brunskill
  - [Sheng et al. 2008, Donmez et al. 2009;
  - Etc.

- Crowdsourcing Global Structure
  - Mobi – [Zhang *et al.* 2012]
  - Context Trees - [Verroios & Bernstein 2014]
- Information Omnivore
  - Never-Ending Language Learning – [Carlson et al 2012]
  - [Angeli & Manning 2014, Pershina *et al.* 2014]

UW CrowdLab

dub

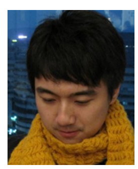
Conclusion

- Crowdsourcing is huge & growing rapidly
  - Specialized communities, citizen science & labor mkts
- Decision theoretic planning – large potential
  - Reduce required labor by 30-85%
  - Sequential decision making is crucial
  - Must model uncertainty & noisy sensors
- Many open questions…

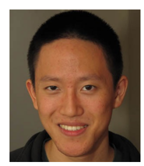

Jonathan Bragg


Lydia Chilton

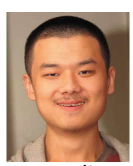

Peng Dai

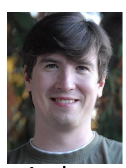

Shih-Wen Huang


James Landay


Chris Lin

Thanks


Angli Liu


Andrey Kolobov


Mausam


Stephen Soderland

116

Extra Slides

UW
CrowdLab

117

dub