UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

eXascale Infolab

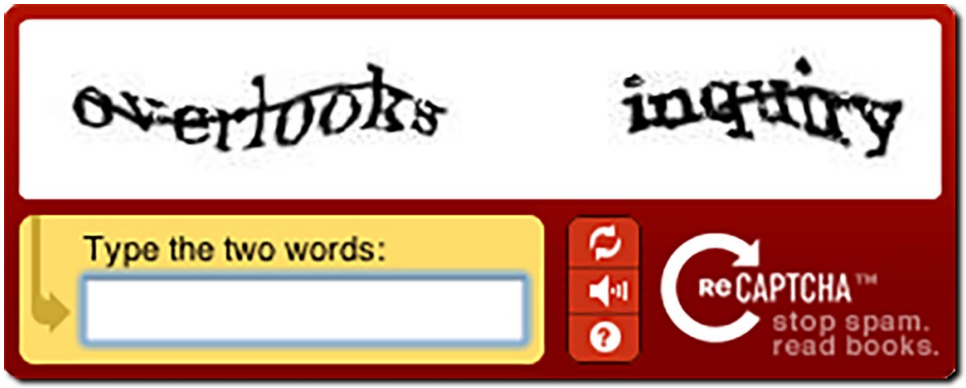
Crowdsourcing and Human Computation

Jie Yang
Senior Researcher
eXascale Infolab, University of Fribourg, Switzerland
Email: jie@exascale.info
Homepage: http://yangjiera.github.io

Outline

- **Introduction to human computation**
  - CAPTCHA and reCAPTCHA
  - Definition as a scientific discipline
- **Core questions and algorithms**
  - Human computation algorithms
  - Effectiveness: output aggregation
  - Effectiveness: task routing
  - Efficiency
- **The future**
  - Combining human and machine intelligence
- **Concluding remarks**

CAPTCHAs



Completely **A**utomated **P**ublic **T**uring Test To Tell **C**omputers and **H**umans **A**part

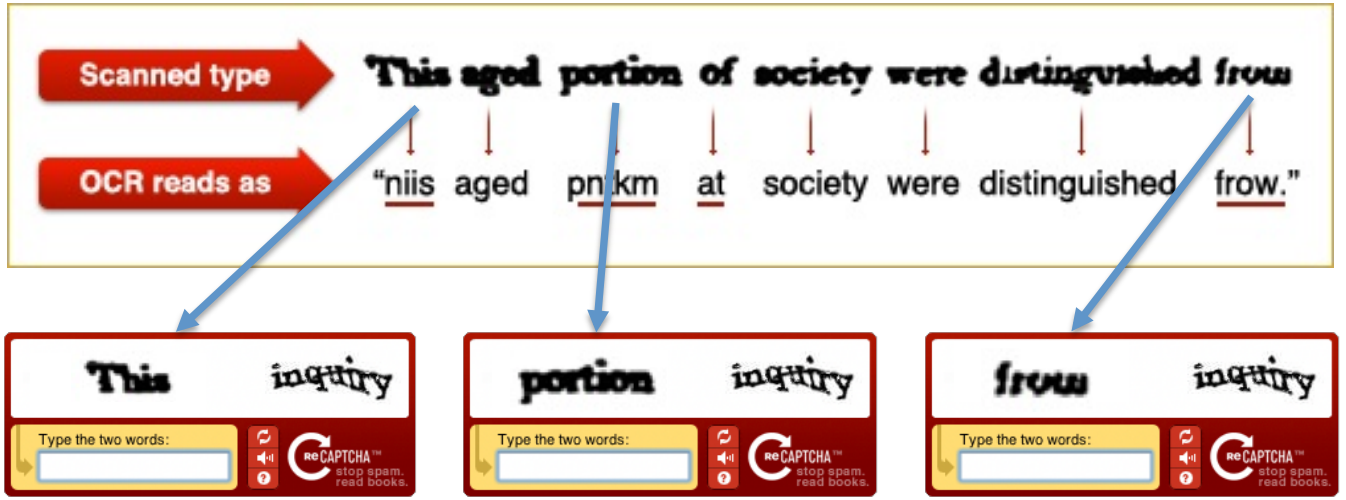
Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Eurocrypt, 2003

100 million CAPTCHAs every day!

How can this insane amount of work be exploited?

ReCHAPTCHA

- Digitalize news articles and books



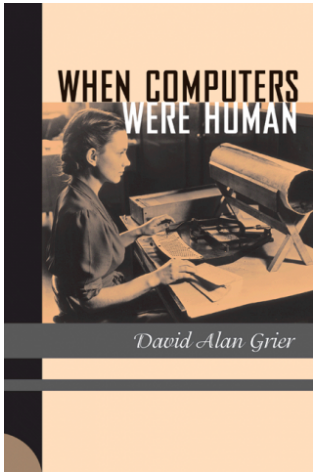
- As of 2011, more than 13 millions articles of The New York Times dating from 1851 have been digitized

Human Computation

- **Computation: the process of mapping input to output**
  - Multiplication: two numbers ➠ product
  - Sorting: set of object ➠ set of object sorted
  - Medical diagnosis: x-ray, lab tests ➠ diagnosis
  - Object recognition: image ➠ tag
  - Translation: source sentence ➠ target sentence
  - …
- **Human computation: computation performed by humans**

Human Computers

- **The term “computer” was used to refer to humans who did computation**



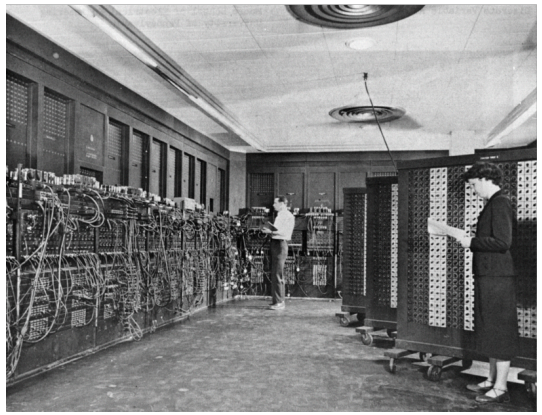
When Computers Were Human: a 250-year epoch
[David A. Grier 2005]

NASA Computers at work in 1949

More: https://www.youtube.com/watch?v=YwqltwvPnkw

Human vs. Electronic Computers

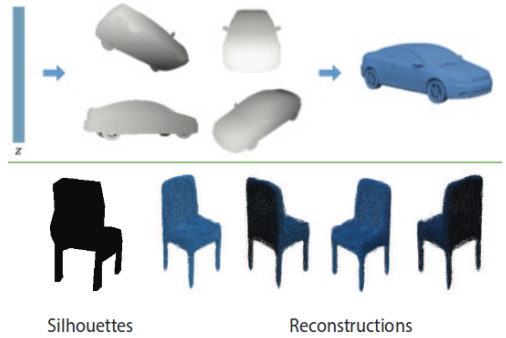
- Electronic
  - Fast
  - Deterministic
  - Arithmetic



- Human
  - Slow
  - Inconsistent/noisy
  - But, **still better at many tasks**

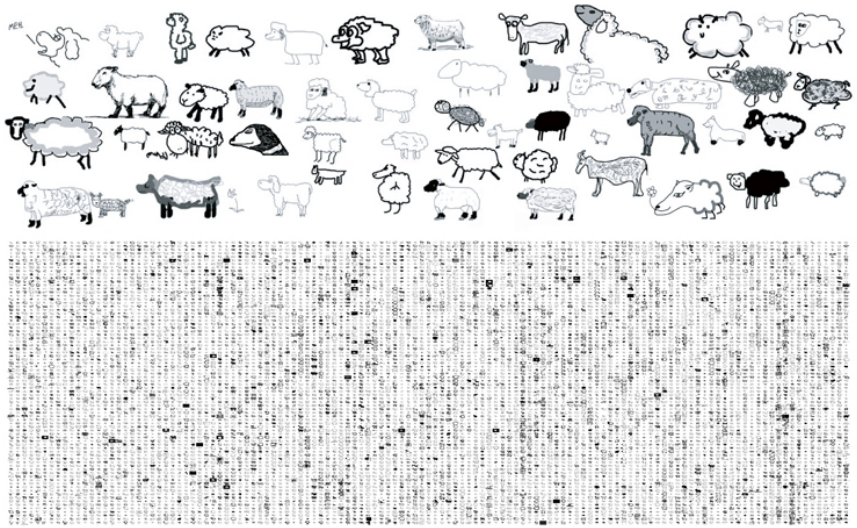
The Human Advantage

- **Humans excel at many tasks that computers cannot efficiently and/or effectively solve**
  - Perception
    - Perception/comprehension: reconstructing information that wasn't captured at capture-time (as in a photo or surface scan)
    - Constructing/inferring information that was never recorded using knowledge humans naturally possess
    - Sketch
    - Recognizing emotions
    - Labeling images
  - Preference/aesthetic judgments
    - Evaluate goodness ("beauty") for sorting or optimization
  - Common sense



The Human Advantage (cont.)

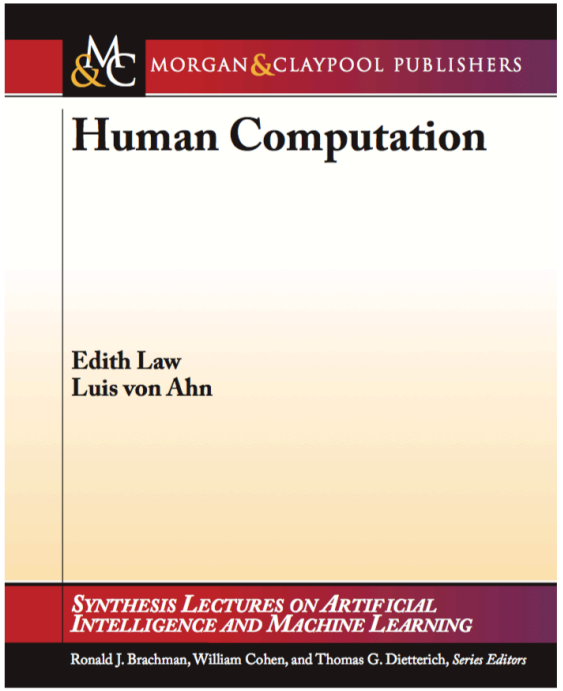
- Creativity
  - search: finding images that go well together
  - art projects like The Sheep Market [Koblin 2006]
  - [Little 2009/10] for expanding text/jokes/shirt design
  - [Yu and Nickerson 2011] for sketching chair designs
  - Etc.



Human Computation as a Scientific Discipline

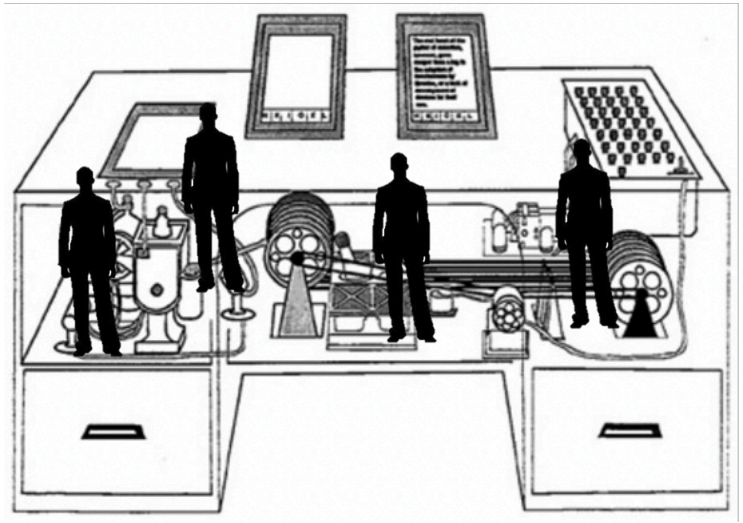
- "A paradigm for utilizing *human* processing power to solve problems that computers *cannot* yet solve"

  "We treat human *brains as processors* in a distributed system, each performing a small part of a massive computation." [VonAhn2005]

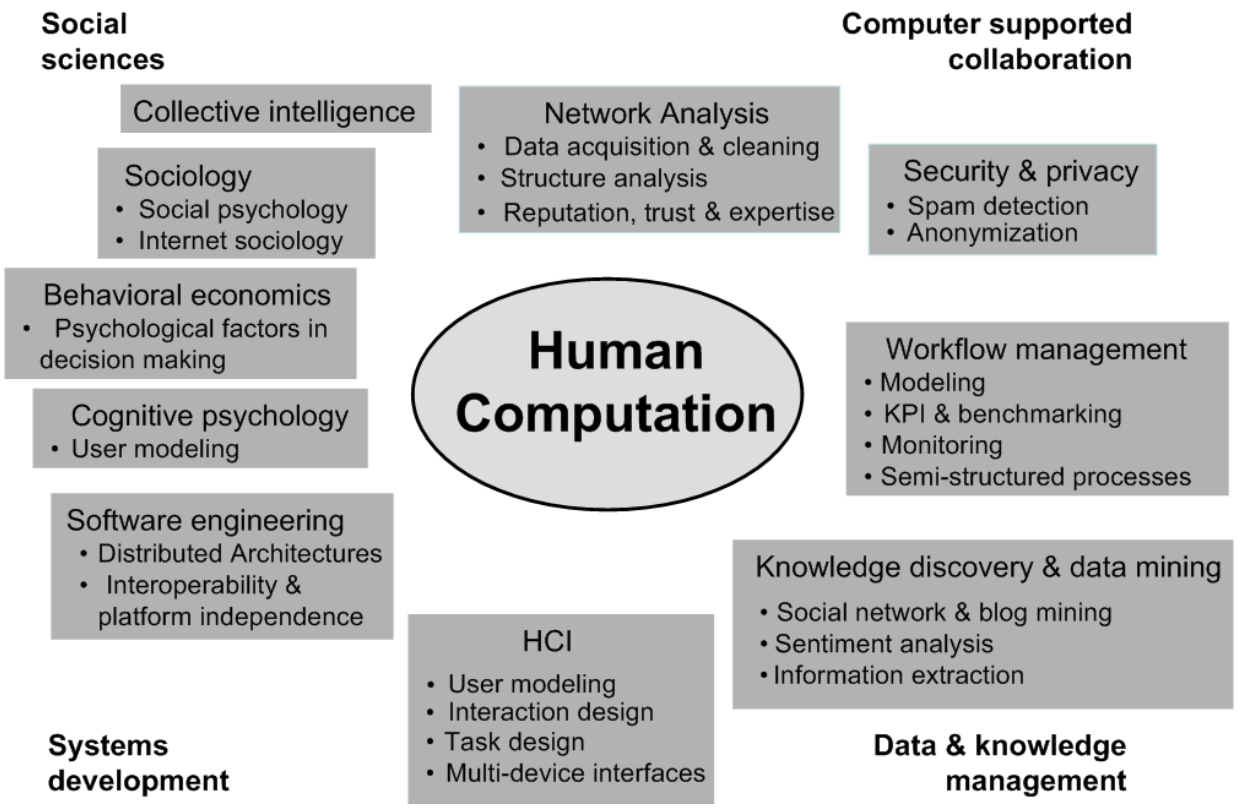


Distinguishing Features

- **“Human” In The Loop**
  - Not bacteria, not ants, not fish
- **Conscious Effort**
  - Humans are actively computing something, not merely carrier of sensors and computational devices
- **Explicit Control**
  - The outcome of the computation is determined by an algorithm, and not the natural dynamics of the crowd



A Growing, Multidisciplinary Field

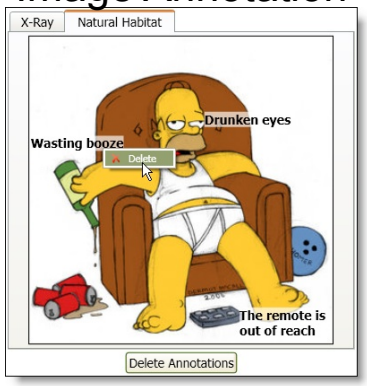


Relationship with AI

- **Computer Scientists in the artificial intelligence field have been trying to emulate human abilities (e.g., speech recognition, vision, natural language processing)**

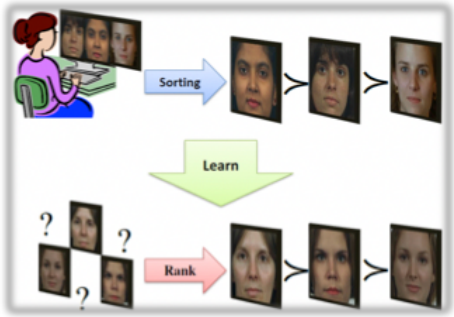
- **Training AI algorithms can be seen as intelligence transfer from humans to machines (human-in-the-loop training/debugging pipelines)**

Image Annotation



Translation

Ordering

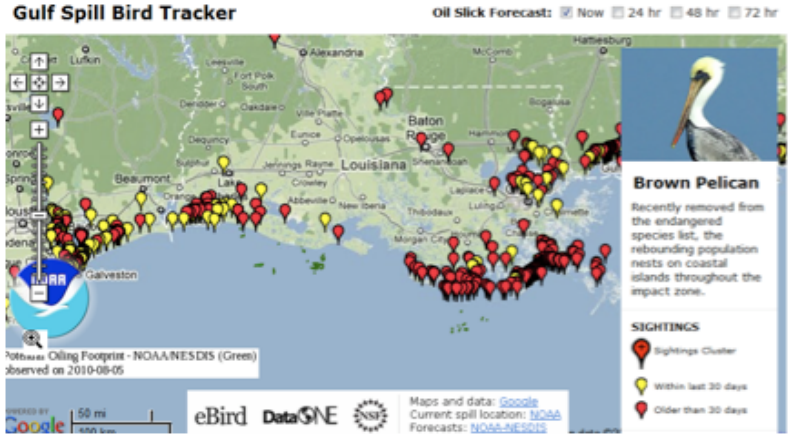


- **Even with AI, humans are still better at machines in many tasks (e.g., most perception tasks, not to mention those that require creativity)**

Human Computers + The Web

GWAP: Game with a purpose

200K participants, 170M observations



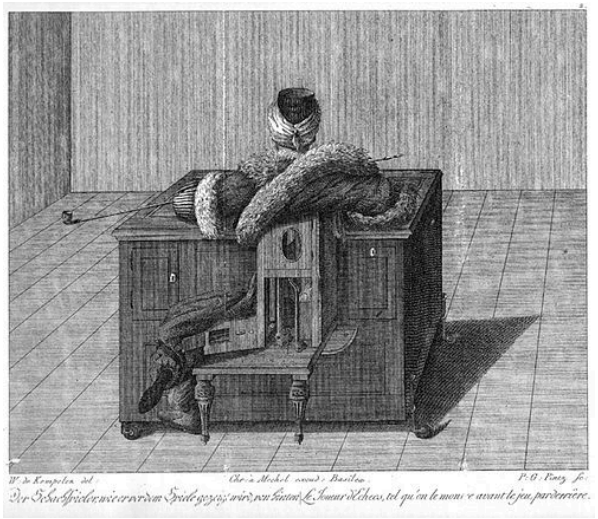
1M volunteers, 100 publications

Crowdsourcing

- **“Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.” (Jeff Howe)**
  - Human computation replaces computers with humans
  - Crowdsourcing replaces traditional human workers with members of the public
  - Crowdsourcing facilitates human computation (but they are not equivalent)

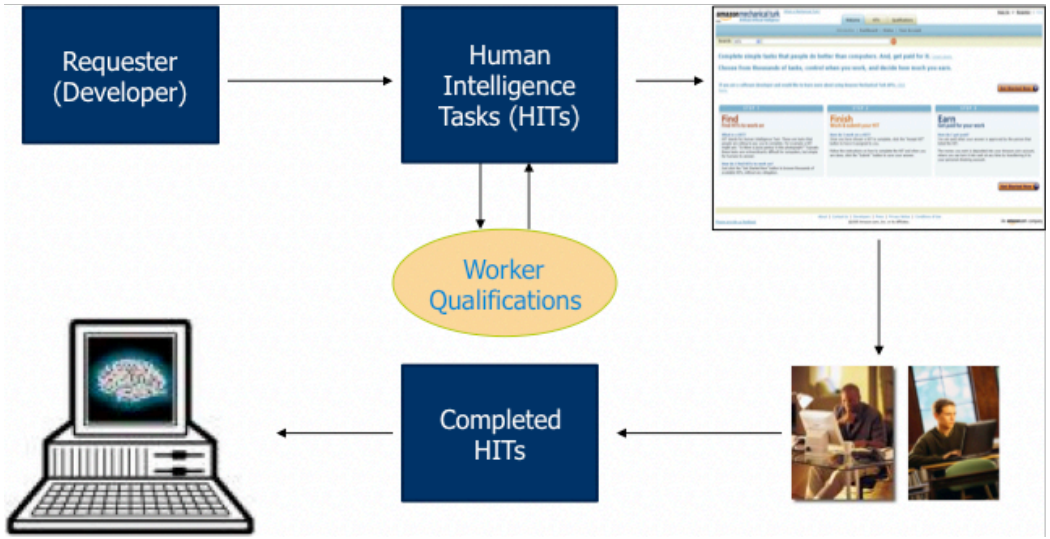
    E.g., Citizen journalism, sensing, ...

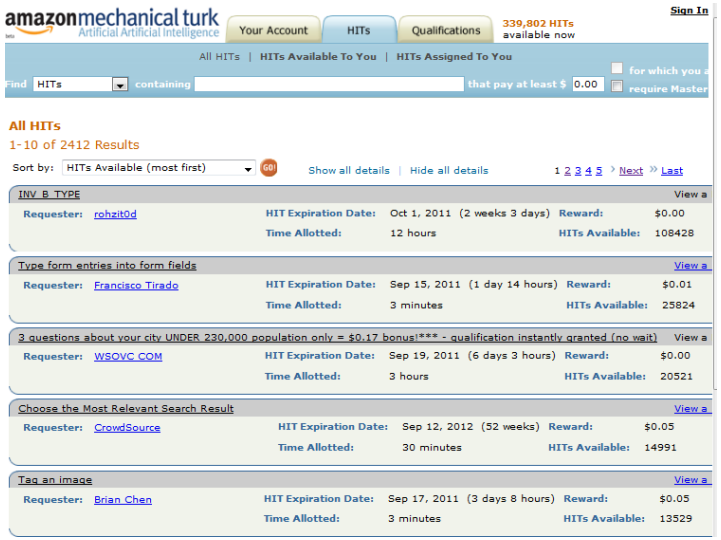
Amazon Mechanical Turk



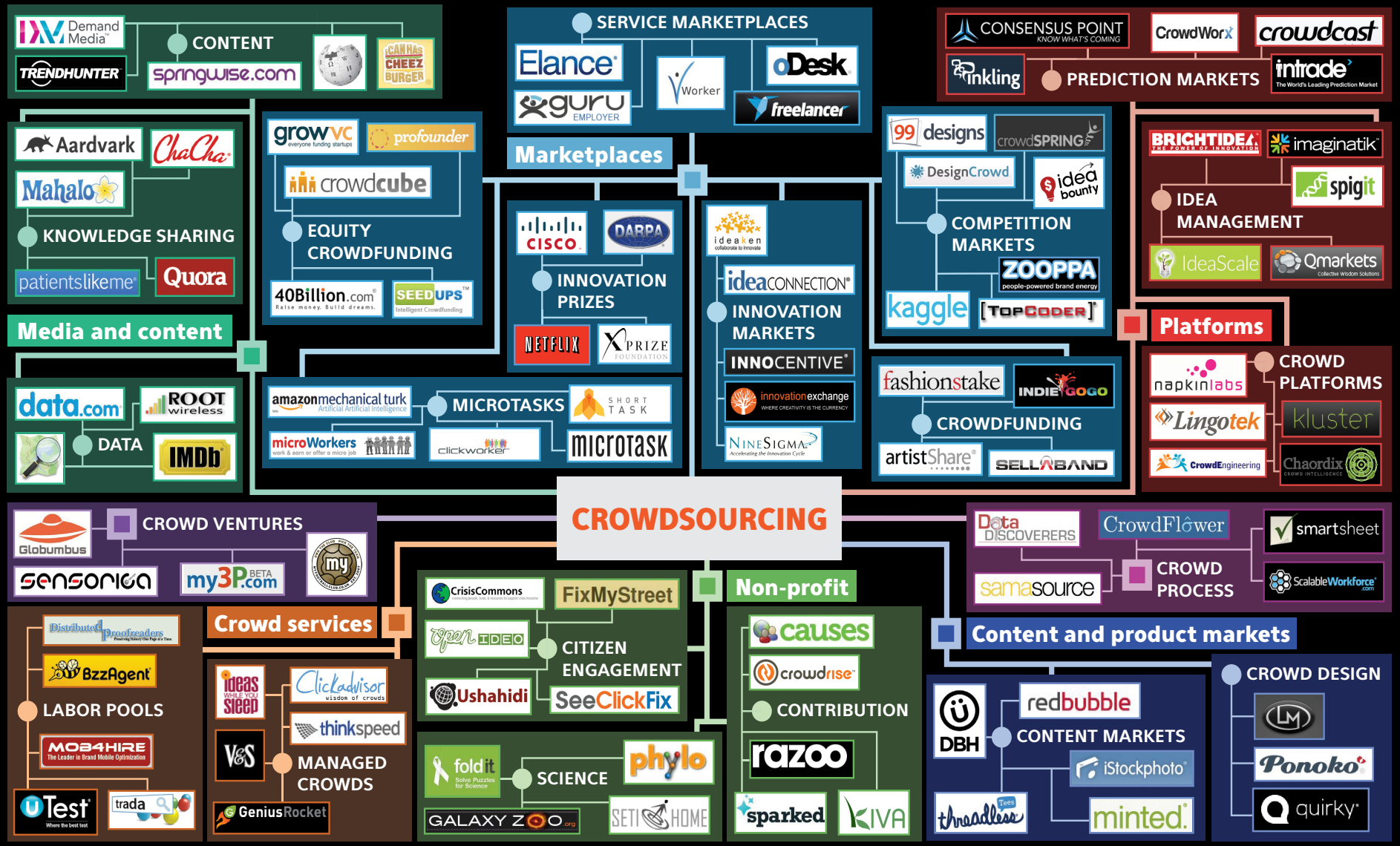
the Turk from Karl Gottlieb von Windisch's 1784 book

- Artificial Artificial Intelligence
- Micro-task crowdsourcing
- Provides a UI and Web Services API to allow developers to easily integrate human intelligence directly into their processing





Crowdsourcing landscape Beta v2



Excerpted from

Getting Results From Crowds

by Ross Dawson and Steve Bynghall

For definitions, analysis, free book chapters, and other crowdsourcing resources go to:

www.resultsfromcrowds.com

Note: examples only, see website for full list of crowdsourcing services

Advanced Human Technologies

ROSSDAWSON Social Media Analytics

Core Research Questions

- **Given a computational problem, design a solution using human computers and automated computers**
  - "How hard is the problem? Is it efficiently solvable?"
  - Is the human computation algorithm correct and efficient?
  - How do we aggregate the outputs of many human computers?
  - How to make the tradeoff between human versus machine?
  - "To whom do we route each task, and how?"
  - How to design tasks, motivate participation and incentivize truthful outputs?

What are Human Computation Algorithms

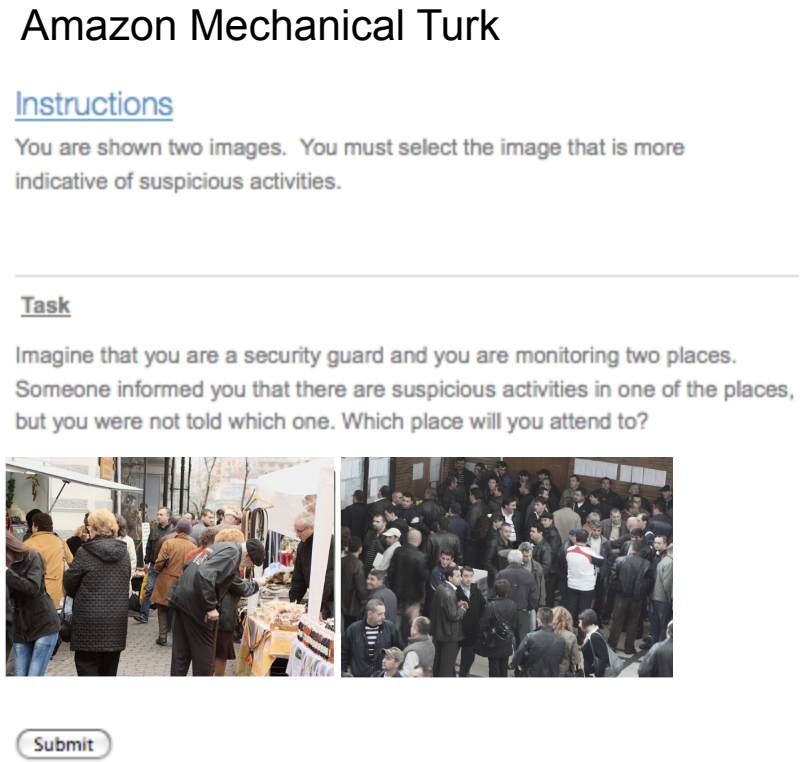
- **Human driven operations**

```
function quicksort(A)
    initialize empty lists L and G
    if (length(A) ≤ 1)
        return A
    pivot = A.remove(find_pivot(A));
    for x in A
        if compare(x, pivot)
            L.add(x)
        else
            G.add(x)
    return concatenate(quicksort(L), pivot, quicksort(G))

function pivot(A)
    return randomIndex(A);

function compare(x, pivot)
    return human_compare(x, pivot)
```

Amazon Mechanical Turk



TurKit (Little et al., 2010)

What are Human Computation Algorithms

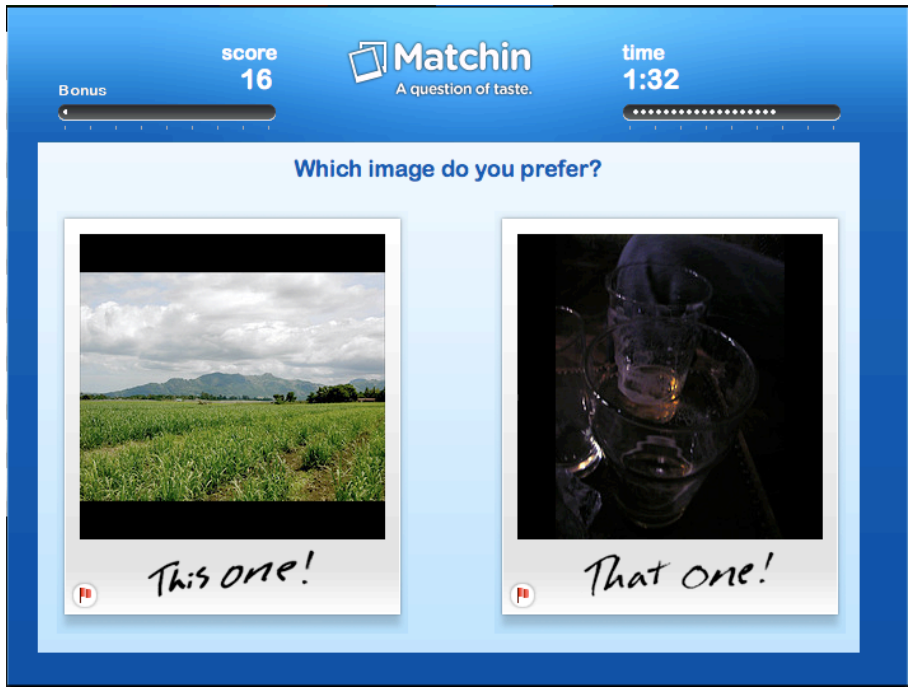
- **Human driven operations**

```
function quicksort(A)
    initialize empty lists L and G
    if (length(A) ≤ 1)
        return A
    pivot = A.remove(find_pivot(A));
    for x in A
        if compare(x, pivot)
            L.add(x)
        else
            G.add(x)
    return concatenate(quicksort(L), pivot, quicksort(G))

function pivot(A)
    return randomIndex(A);

function compare(x, pivot)
    return human_compare(x, pivot)
```

Game with a purpose



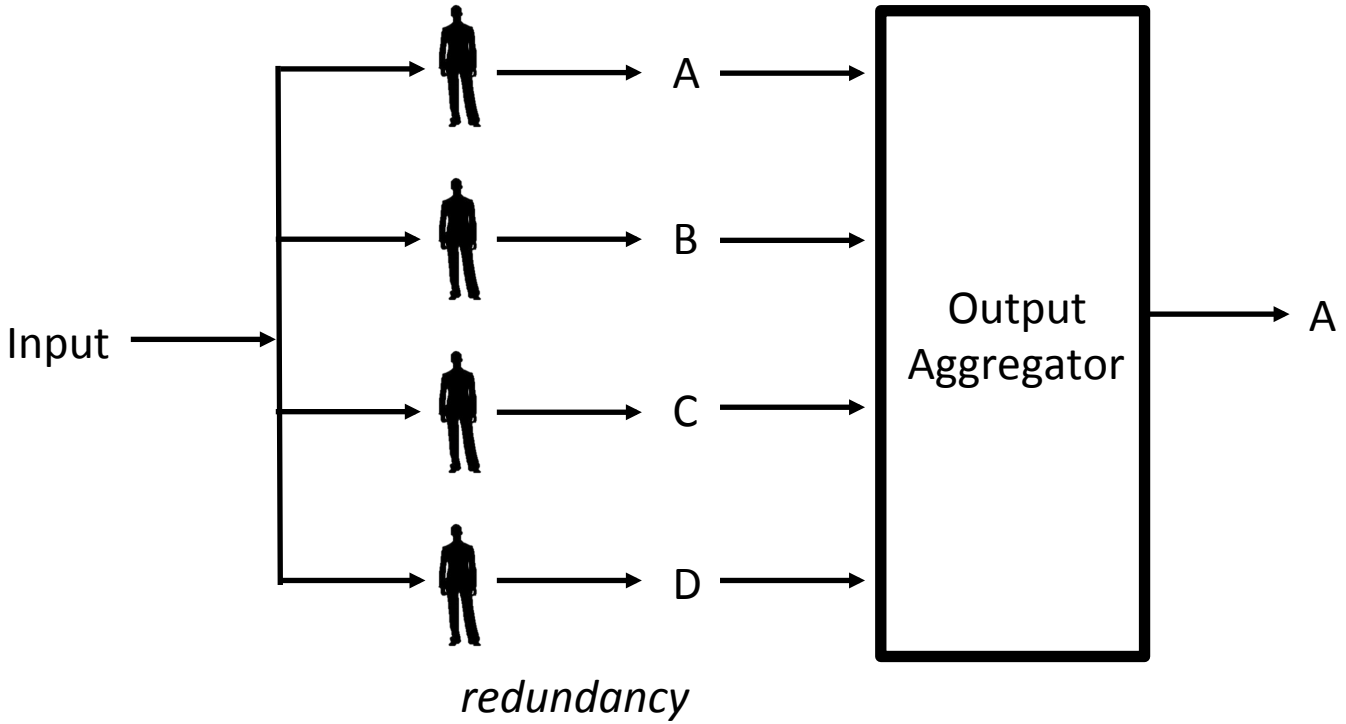
Correctness of a Human Computation Algorithm: Output Aggregation

- **Outputs generated by human computers can be noisy**
- **Basic motivation: The "truth" exists, and through *redundancy* we can find it ("wisdom of the crowd")**
- **Truth can either be objective or subjective/cultural**
  - Objective: a definitive answer exists beyond human judgments, but hard to reach
  - Subjective/cultural: shared beliefs of a group of people, often involving perceptual judgments

Objective
- cancer or not
- number of volcanos on Venus
- location or time of a photo

Subjective
- is this music calm?
- is this image pornographic?
- is this disease contagious?

Output Aggregation in a Nutshell



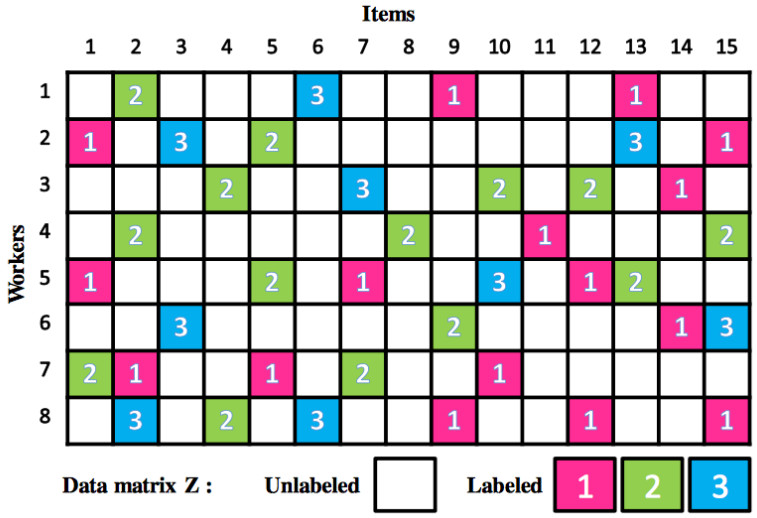
- **Outputs can be aggregated by humans or automatically**

The Simplest Way: Majority Voting

- **Assumption**
  - Workers independently generate output, depending on the truth
- **No assumption of any hidden factors that may influence the annotation process**
  - Prior about which categories are more or less likely to be the true classification
  - Worker properties
    - Expertise (e.g., bird identification)
    - Bias (e.g., mother vs college students)
    - Physical condition (e.g., fatigue)
  - Task properties
    - Quality (e.g., blurry pictures)
    - Difficulty (e.g., transcription of non-native speech)

Problem Formulation

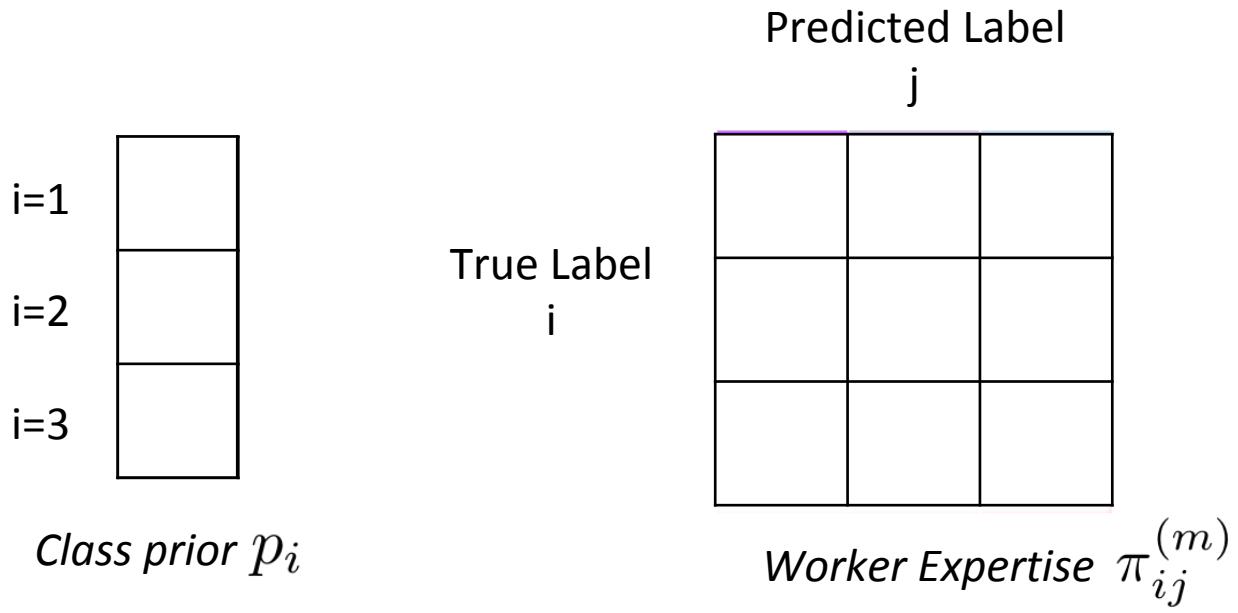
- **Suppose we have**
  - M workers, N items
  - and a label matrix L in which $L_{mn} \in C$ indicates the label of worker m for item n , where C is the set of classes



- **Our goal is to infer the *true label* for each item**

The Dawid-Skene Model: Assumptions

- **The Dawid-Skene model assumes**
  - Class prior p_i: different classes are differently likely to occur
  - Each worker has their own expertise represented as a *confusion matrix* $\pi_{ij}^{(m)}$ indicating the probability of worker m classifying an item to class j given the true label is i

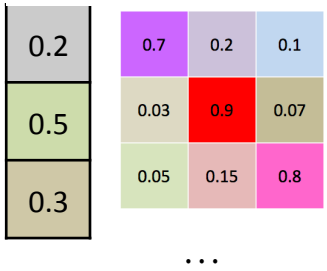


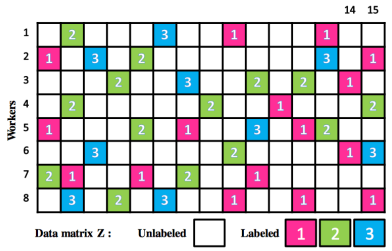
- **True labels, class prior and worker expertise are all unknown!**

The Dawid-Skene Model: Inference Goal

- **The Dawid-Skene model aims at inferring**
  - Parameters: class prior p_i , worker expertise $\pi_{ij}^{(m)}$
  - True labels T_{ni} for each item n
    - T_{ni}=1 if item n has true label i , otherwise 0
- **Maximum likelihood estimation (MLE)**
  - Find the parameters and true labels that maximize the likelihood of the observed label matrix

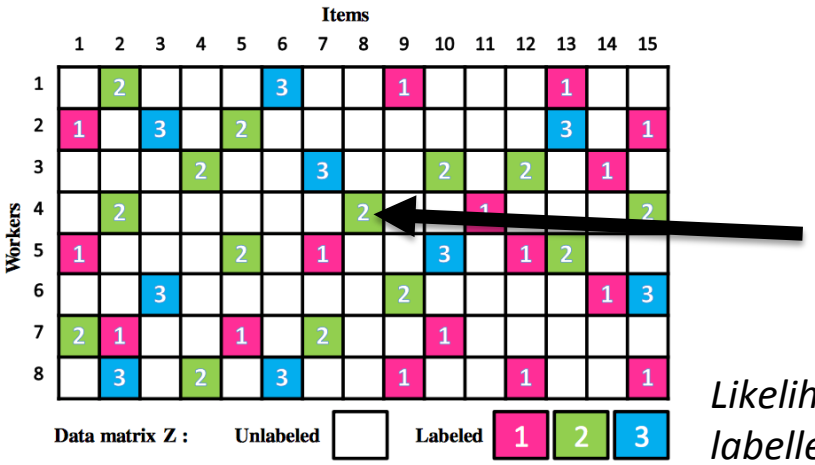
$$p_i, \pi_{ij}^{(m)}, T_{ni} = \text{argmax}\ \mathcal{L}(L)$$





The Dawid-Skene Model: Likelihood (Single Label)

- **View the process of worker label as a generative process**
  - For each item, true label is drawn from the prior class distribution
  - Worker label is generated according to the true label and worker expertise

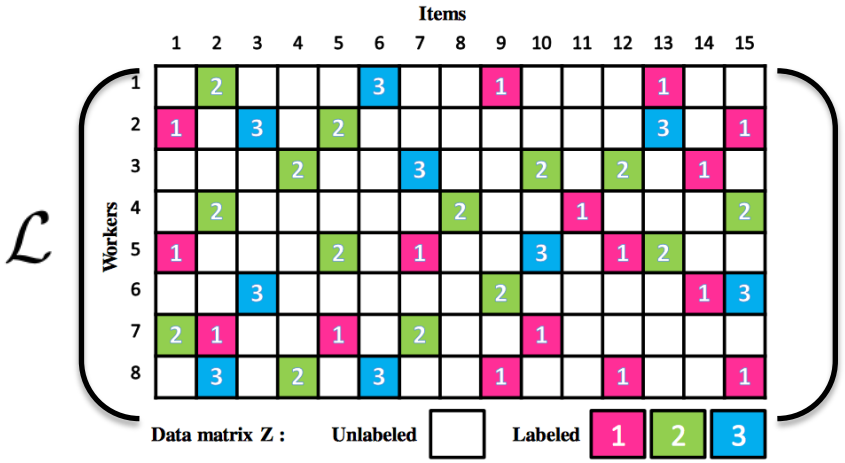


$$L_{4,8} = \sum_{i}^{3} p_i \times \pi_{i,2}^{(4)}$$

Likelihood that item 8 is labelled as 2 by worker 4

The Dawid-Skene Model: Likelihood (Overall)

- **(Continued)**
  - For all items *n*, workers *m* and classes (*C*):



$$\mathcal{L} = \prod_{n}^{N} \sum_{i}^{C} p_i \prod_{m}^{M} \pi_{i,(j=\text{observed})}^{(m)}$$

$$\text{or} \quad \prod_{n}^{N} \sum_{i}^{C} p_i \prod_{m}^{M} \prod_{j}^{C} (\pi_{ij}^{(m)})^{t_{nj}^{(m)}}$$

if item n receives multiple labels from m; $t_{nj}^{(m)}$ is the number of times worker m assigns label j to item n.

The Expectation Maximization Algorithm

- **MLE for the following problem is intractable**
  - Exponential complexity

$$\operatorname{argmax} \prod_{n}^{N} \sum_{i}^{C} p_i \prod_{m}^{M} \prod_{j}^{C} (\pi_{ij}^{(m)})^{t_{nj}^{(m)}}$$

- **Expectation Maximization - a general solution for this kind of problem**
  1. E-step: estimates the true labels of each item by weighing the votes of the workers according to our current estimates of their expertise (as given by the confusion matrix)
  2. M-step: estimates the class prior and confusion matrices based on the current beliefs about the true labels of each item
  3. Repeat 1 and 2 until convergence (**guaranteed**)

The E- and M-step, and initialization

- **E-step:** $p(T_{ni} = 1) \propto p_i \prod_m^M \prod_j^C (\pi_{ij}^{(m)})^{t_{nj}^{(m)}}$
- **M-step:** $\hat{p}_i = \frac{1}{N} \sum_n^N T_{ni}$

$$\hat{\pi}_{ij}^{(m)} = \frac{\sum_n^N T_{ni} t_{nj}^{(m)}}{\sum_j^C \sum_n^N T_{ni} t_{nj}^{(m)}}$$

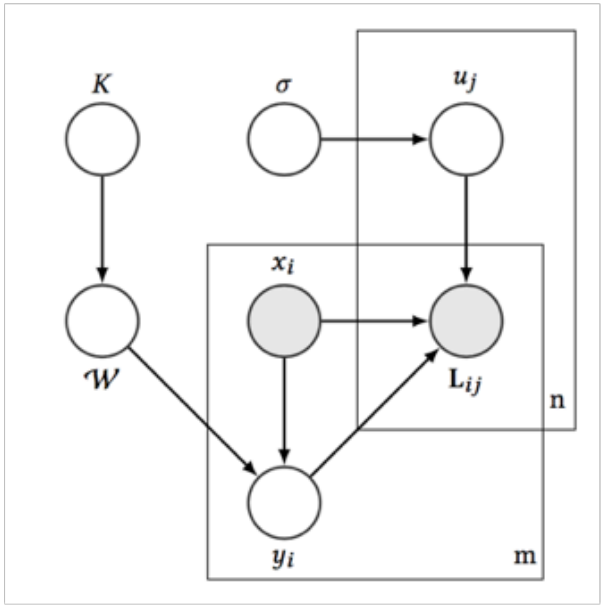
- **Initialization:**
  - Randomly initialize the true label
  - Or, using majority voting to initialize the true label (faster convergence)
  - Once initialized, go into the M-step, then repeat E- and M-steps

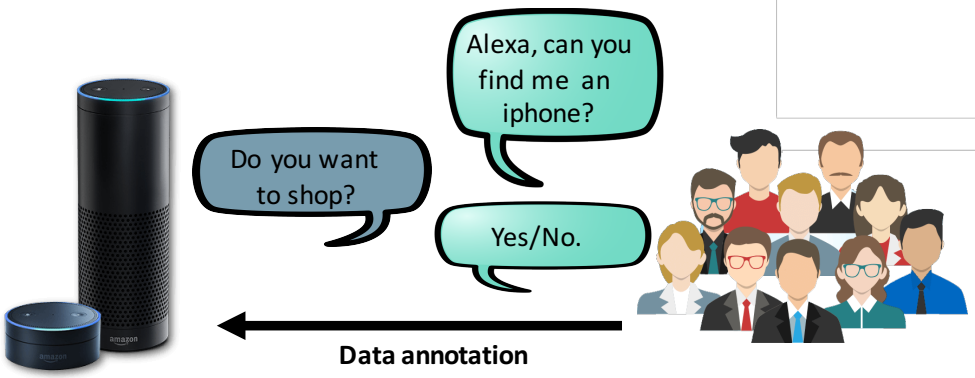
Remarks on Performance

- **Majority voting works well in many cases**
  - In particular, when redundancy > 20
- **EM-based methods are superior with less redundant worker annotations, but not too little …**
- **Dawid-Skene is more robust than many other methods**
  - Regularization methods to reduce overfitting:
    - Class-Conditional Dawid-Skene model (b): class-conditioned error
    - Homogenous Dawid-Skene model (c): same error across all classes

Extensions

- Expertise learning with content
- Combining with machine learning: Learning from Crowds
- An example in Amazon Alexa: deep learning from customers





Other Aggregation Task: Rank Aggregation

- **Goal: individual rankings ➡ full ranking**

Paired comparison

Rating (scale 1-4)

Ordering

- How should we aggregate a set of partial rankings to generate a complete ranking that actually *reflects the opinion of the crowd*?
  - Conflicts due to errors and biases
- Related to many problems in other domains
  - Aggregating search results from multiple search engines
  - Recommend a product that is generally liked by users (Amazon Choice)
  - Vote aggregation for collective decision or social welfare

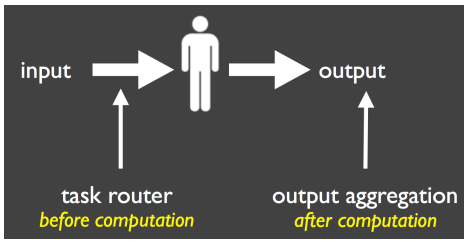
Task Routing: find the right workers

- **Knowledge-intensive tasks**
  - Annotate flower/bird types from digitalized artworks
  - Determining cancer from X-ray images
  - French-Engilish translation
- **Subjective/cultural tasks**
  - Assess attractiveness of an image
  - Assess offensiveness of a tweet
- **Context-dependent tasks**
  - Local air quality measurement
  - Real-time report of traffic jams

- **Task routing: early intervention to improve work quality**

Task Routing: finding the right workers

- **Task routing: early intervention to improve work quality**



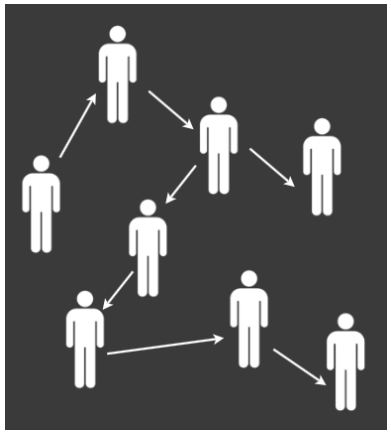
- **The most popular task routing method is WHTBT**
  - "Whoever Happens To Be There"
  - Used in many paid crowdsourcing platforms

- **Why is task routing difficult**
  - Difficult to infer worker reliability (anonymous, disposable)
  - Difficult to model knowledge demand

- **Push vs. Pull (D.E. Difallah, WWW2013)**
  - Push: workers are passive receivers of task; the system takes complete control over who is assigned which task (*Expert Finding*).
  - Pull: workers are active seekers of tasks; the system supports a set of interfaces that enable workers to look for tasks to assign themselves (e.g., Amazon Mechanical Turk interface).

Peer Routing

- **Workers routing tasks to one another**
  - Workers can either
    - accept a task
    - reject a task
    - or recommend another worker who may have more expertise to handle the task
  - Related to *expertise location and recommendation*
  - E.g., DARPA Red Balloon Challenge, fashion influencer finder



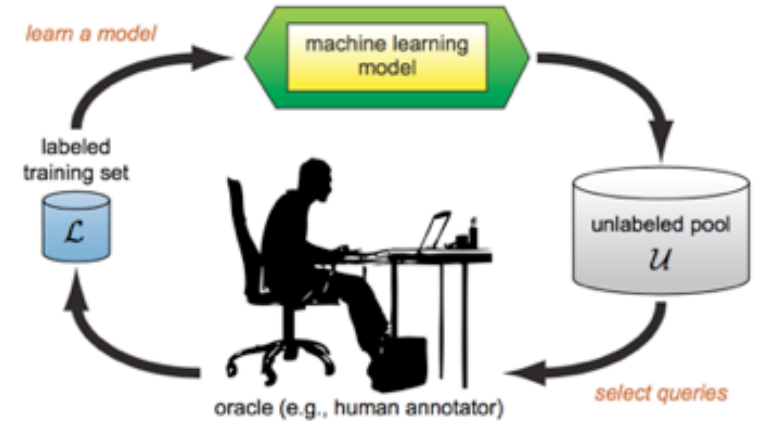
Efficiency

- **Time efficiency**
  - The need for real-time crowdsourcing

- **Does increasing the reward of a micro-task increase efficiency or effectiveness?**
  - Mason & Watts 2010

Efficiency (Cont.)

- Active Learning
  - "The learner can select the data from which it learns the best." (Settles, 2011)



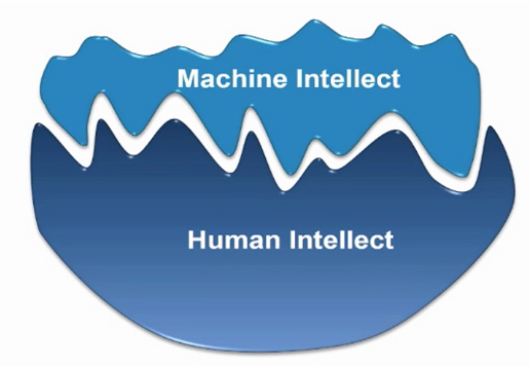
(Burr Settles 2010)

  - Which input should we process? What questions should we ask?

Human-Machine Hybrid Computing

- **An example: Metastatic Breast Cancer Identification [wang et al. 2016]**
  - Wining solution: 7.5% error
  - By pathologist: 3.4% error
  - Combining the two: **0.5**% error

- **Envision**
  - Models of people and tasks
  - Models of complementarity
  - Coordination of initiative



Concluding Remarks

- **Human Computation & Crowdsourcing: a rising discipline with huge scientific and societal impacts**
- **One of the central problems in human computation is truth inference from noisy output**
- **Great potential in future online labor markets and human-machine hybrid computing**