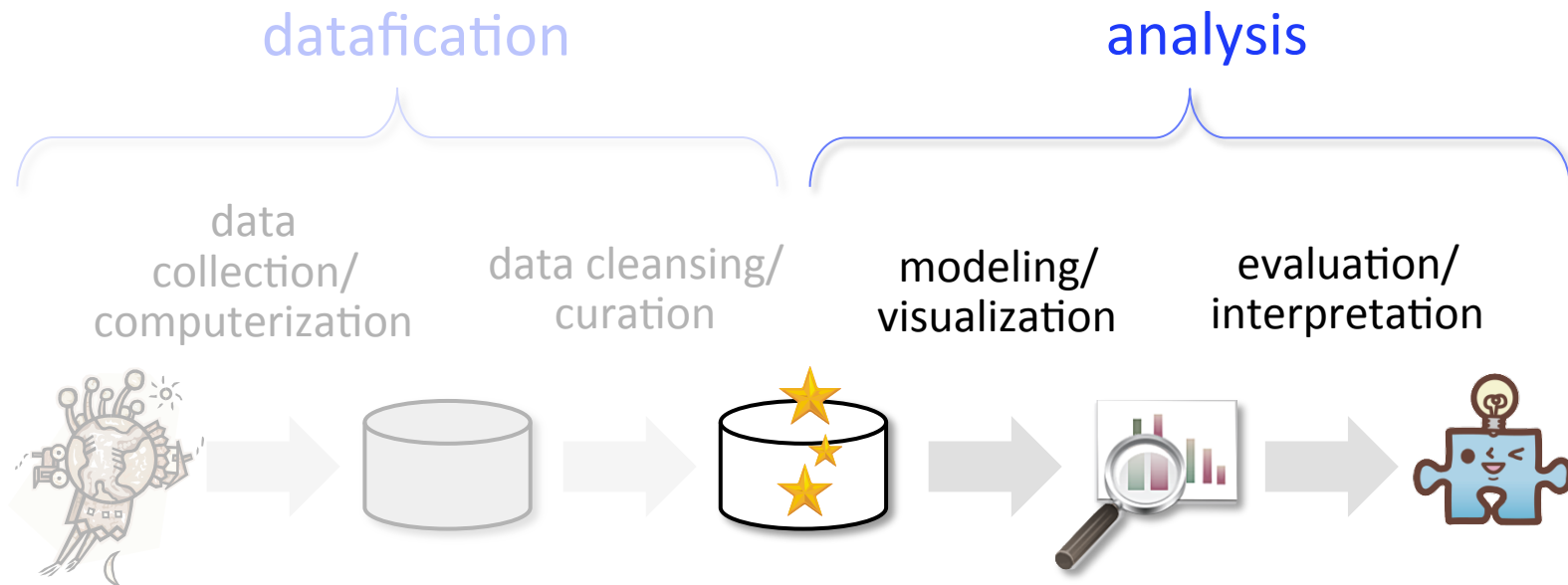


# Part III: Crowdsourcing for analysis



# Crowdsourcing for analysis:

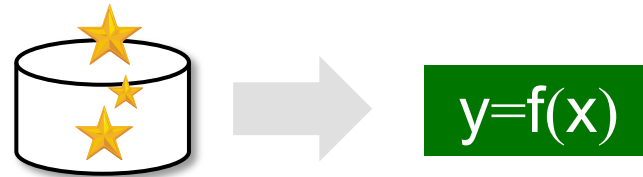
## Predictive and descriptive analysis

---

### ■ Two types of analysis

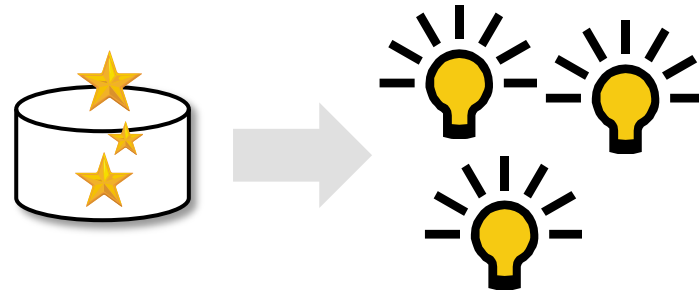
#### — *Predictive analysis*

- GOAL: Building models to make predictions about future
- OUTPUT: Predictive models



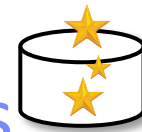
#### — *Descriptive analysis*

- GOAL: Summarizing and making sense of data
- OUTPUT: Findings



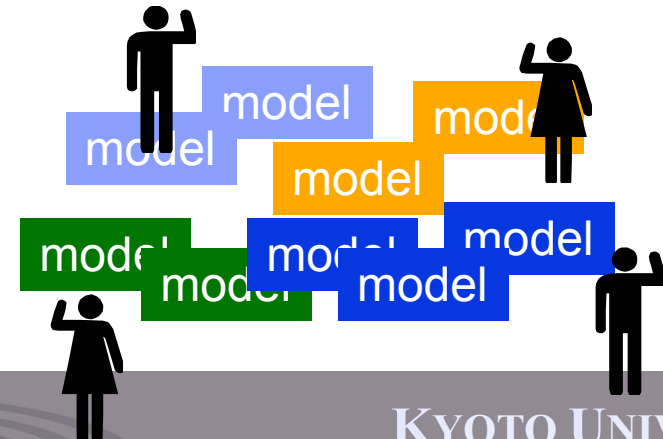
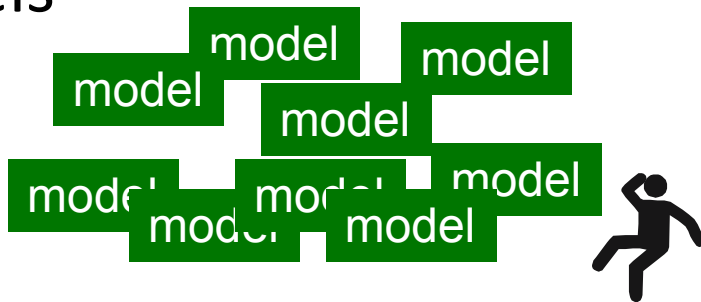
# Crowdsourcing for predictive analysis

Modeling requires crowds of data scientists



$$y=f(x)$$

- Predictive modeling is actually labor-intensive
  - No free lunch theorem: there is no universal good model
  - Performance improvement is often brought by careful
- Broad exploration of models is hard for a *single data scientist*
- *Crowds of data scientists* can examine a large number of models



# Predictive modeling competition:

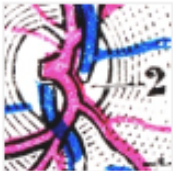

## Platform for leveraging crowds of data scientists

- Predictive modeling competition

- E.g, **kaggle**

- Over 150 competitions have been hosted
- Over 140K data scientists have been registered
- KDD Cup have been hosted from 2012

- Participants build models and compete for monetary rewards

Competition Name		Reward		
	<b>Diabetic Retinopathy Detection</b> Identify signs of diabetic retinopathy in eye images	\$100,000		
	<b>West Nile Virus Prediction</b> Predict West Nile virus in mosquitos across the city of Chicago	\$40,000	561	32 days

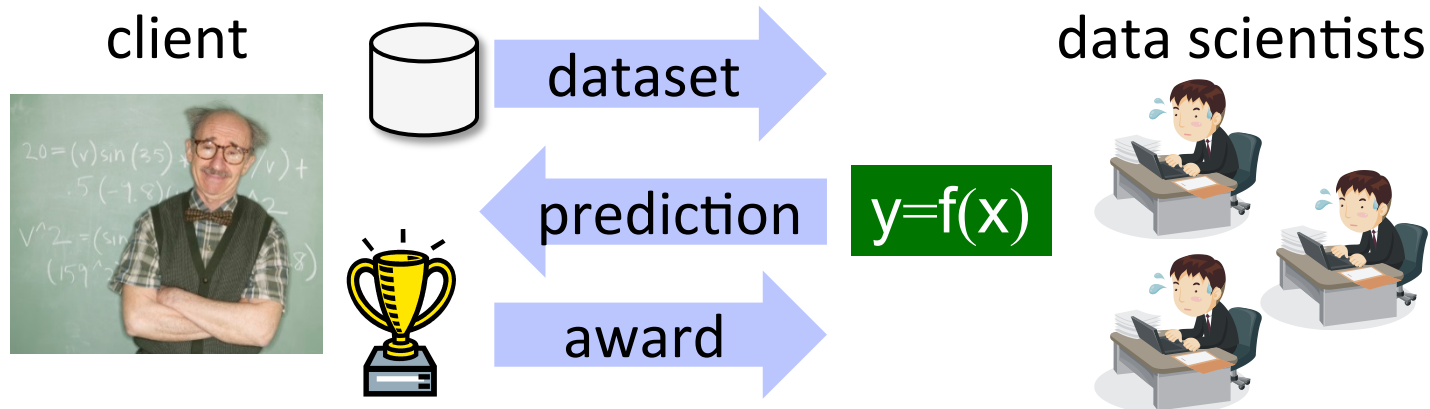
Winner gets the reward

# Predictive modeling competition protocol:

## Participants explore model spaces to find better one

### ■ Competition protocol:

1. Training and test dataset are published
  2. Participants submit predictions for test dataset
  3. Participants get feedback via a live leader board  
..... (repeat step 2 and 3)
  4. Winner is determined by results on test data
- several weeks to months



# Wikipedia link prediction challenge: Modeling competition experiment

- We ran a competition on CrowdSolving (for one month)
- Dataset: link prediction among articles of Wikipedia
  - Goal: given several link information, predict the other (hidden) links
    - ~ 23,000 articles with links and page categories
    - warm start and cold start



Jimmy Wales



Wikimedia foundation



San Francisco



The screenshot shows the CrowdSolving competition page. The header includes the CrowdSolving logo and navigation links. The main content area is titled 'コンペ詳細' (Competition Details) and includes sections for 'スポンサー' (Sponsor: infocom), '趣意' (Purpose), and '懸賞金総額' (Total Prize Money: 100,000円). A callout bubble points to the prize money section, stating 'Total amount of prize: \$1,000'. The prize breakdown is as follows:

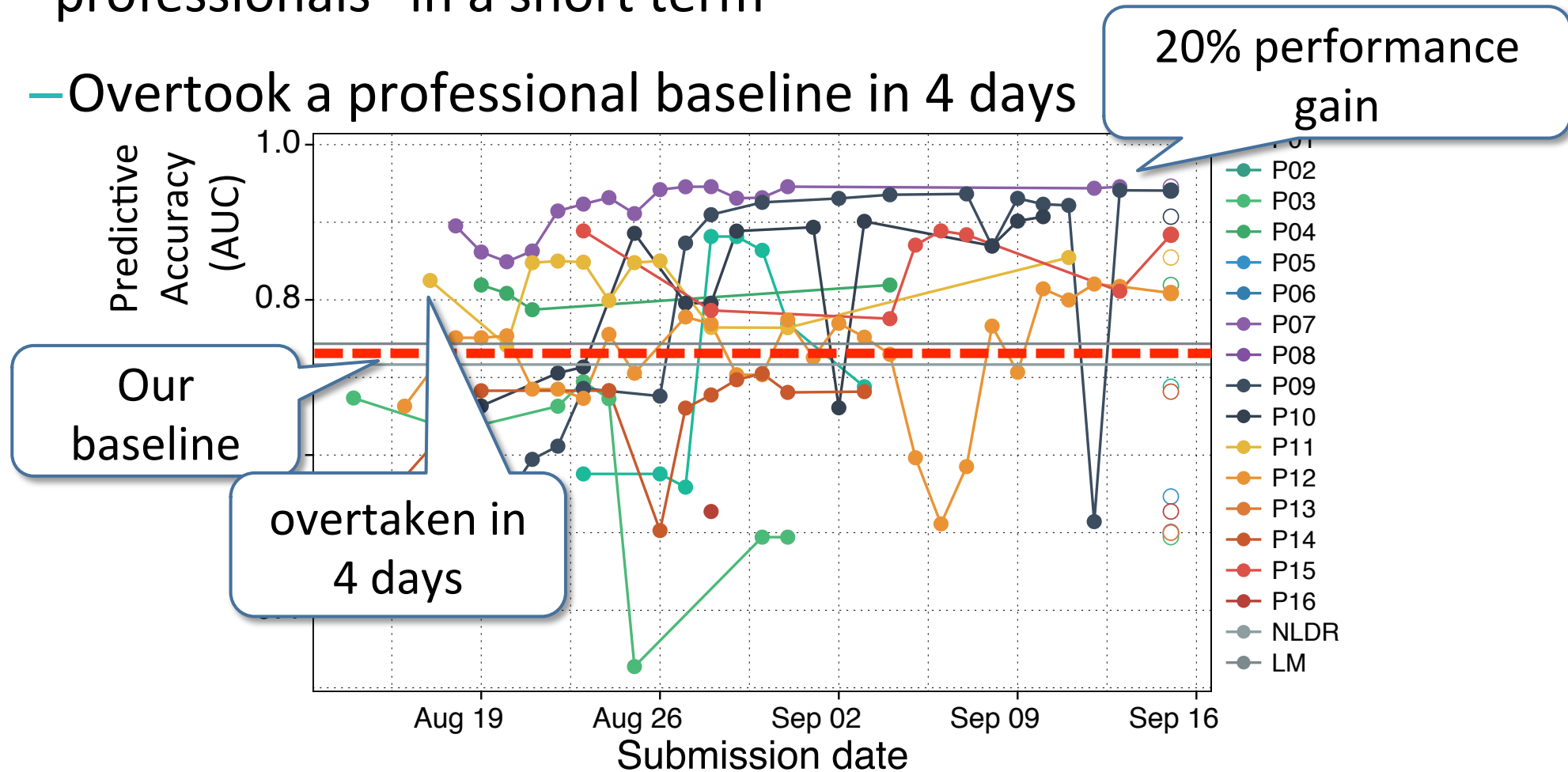
順位	懸賞金
1位	50,000円
2位	20,000円
3位	15,000円
4位	10,000円
5位	5,000円

# Competition results:

## Power of crowd data scientists defeats “professionals”

- Crowd of data scientists achieved far better results than “professionals” in a short term

—Overtook a professional baseline in 4 days



# Methods used by professionals and crowds:

## Crowd winners used simple heuristic techniques

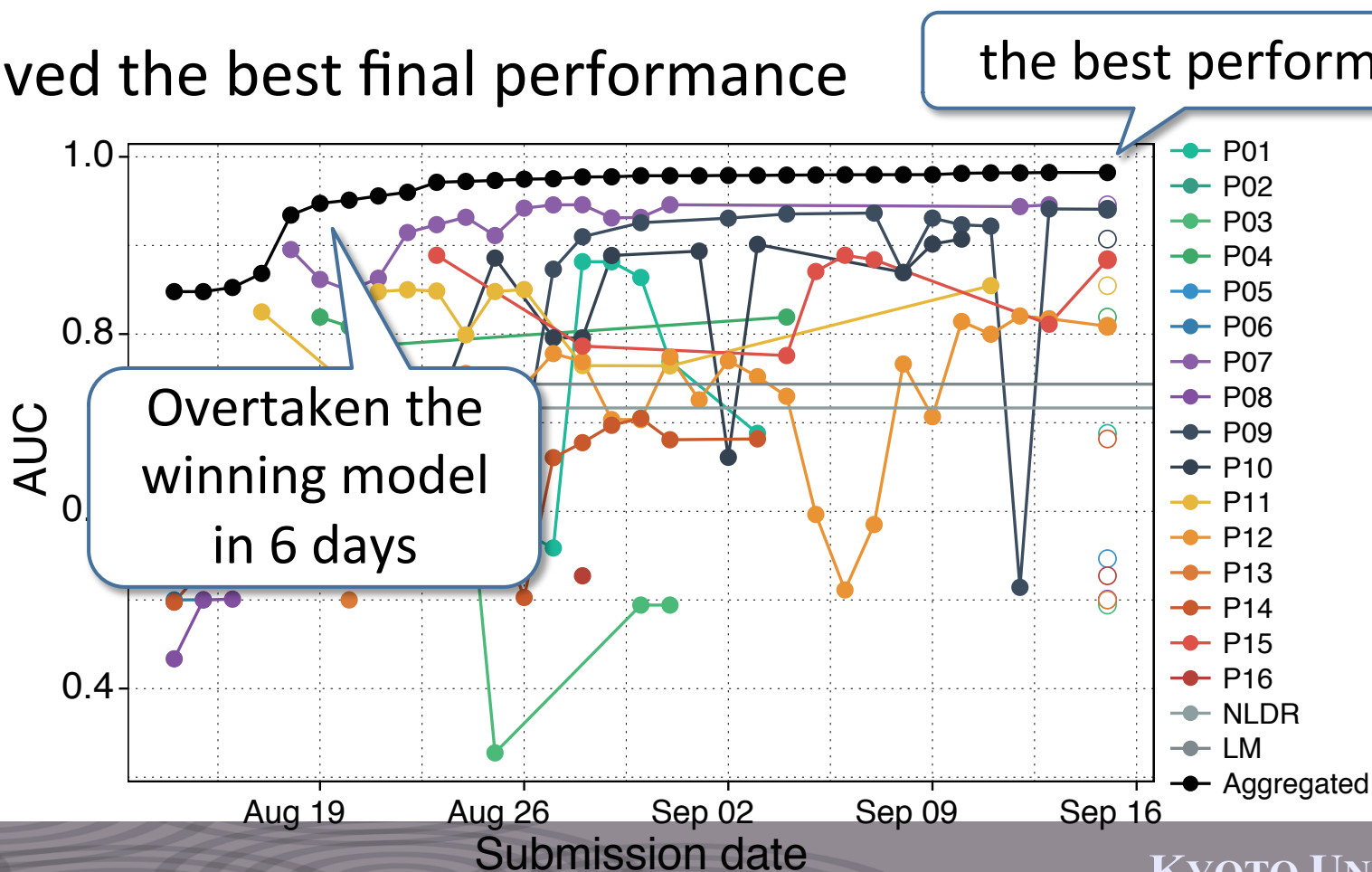
---

- Professionals used a state-of-the-art link prediction method
  - Proposed in AAAI'12
- Crowd winners used simple heuristic techniques
  - 1<sup>st</sup> place:  
Random forest + heuristic features
  - 2<sup>nd</sup> place:  
carefully designed link occurrence measure
- Broad exploration of various models is a key to win



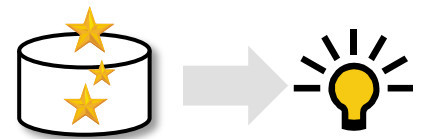
# Human-machine hybrid prediction: Machine learning further boosts performance

- Combination of submitted models with machine learning
  - Overtook the best individual performance in 6 days
  - Achieved the best final performance



# Crowdsourcing for descriptive analysis

## Crowds view data from various aspects

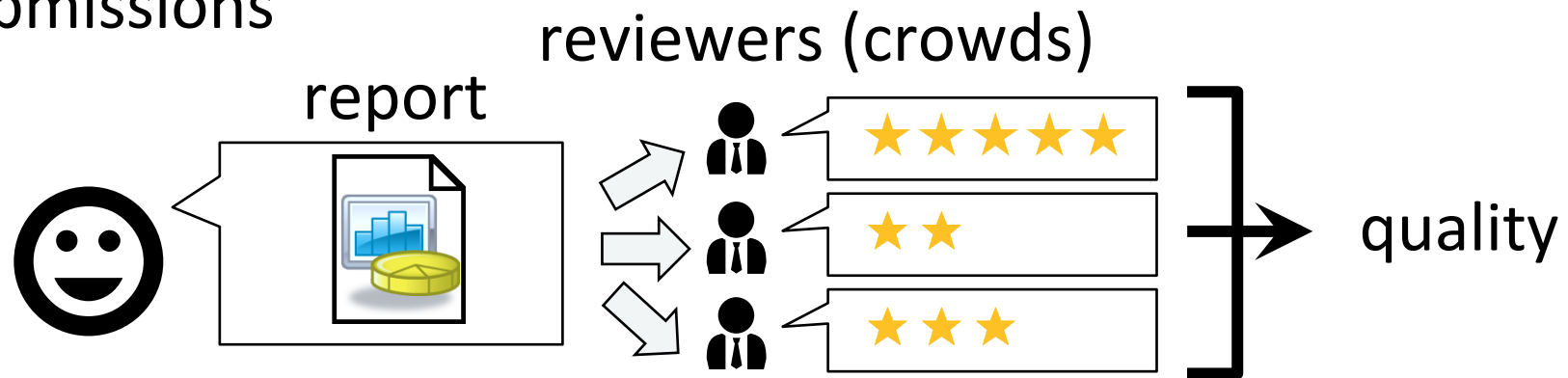


- Not all analysis can be easily formalized as prediction
- Crowds are expected to provide findings from various aspects
- CrowdSolving hosted a competition for descriptive analysis
  - Given: access log and purchase history in an e-book store
  - Task: find typical patterns in consumer behavior, and provide insights for business decision making



# Quality control for descriptive analysis: Peer-reviewing

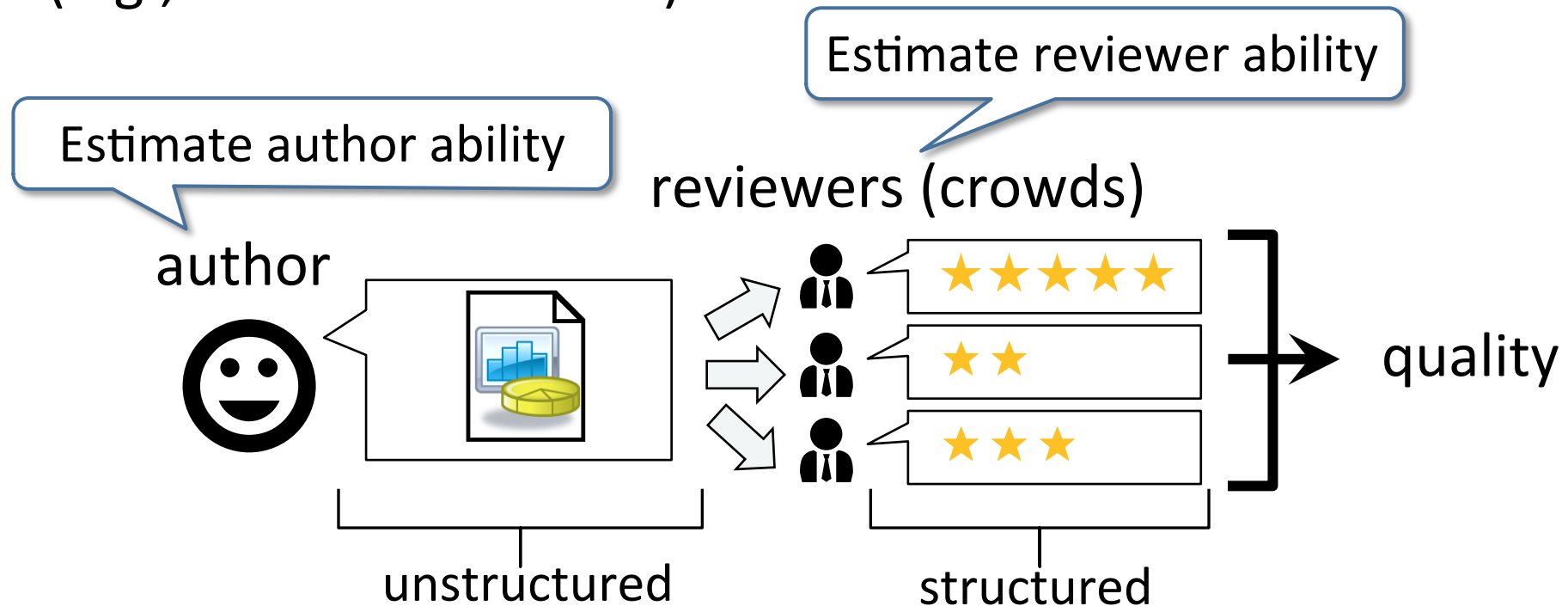
- Predictive analysis:  
easy to evaluate models by prediction performance
- Descriptive analysis:  
hard to evaluate findings in an automatic way
- Peer-reviewing is a possible approach
  - Used in MOOCs to evaluate a large number of student submissions



# Quality control for peer-reviewing:

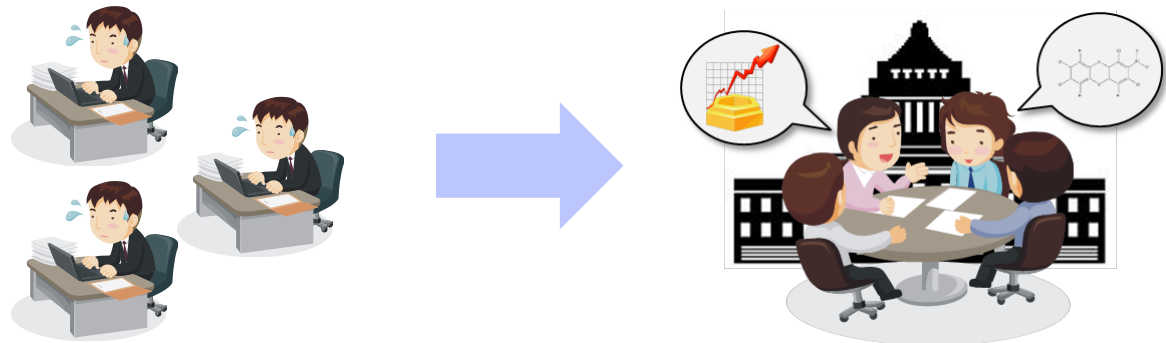
## Introducing redundancy and estimating abilities

- No guarantee that crowds accurately evaluate submissions
- Ratings are given in *structured format*
  - ➔ quality control methods in datafication can be applied (e.g., latent class model)



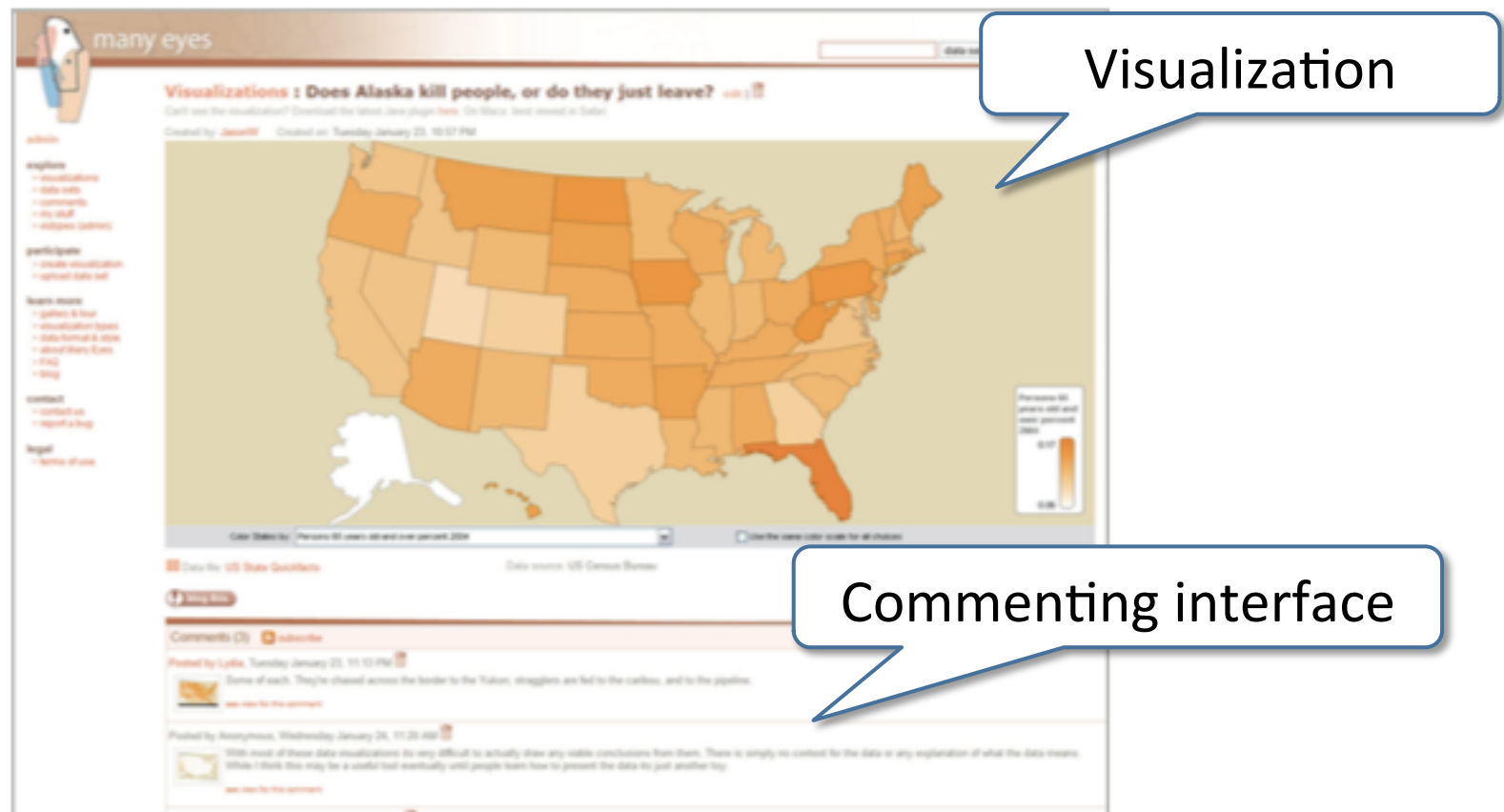
# Collaborative predictive analysis: Mechanism design for collaboration

- Competition scheme makes participants work individually
  - Knowledge and expertise are not shared
- Collaboration should be encouraged for efficient modeling
  - Mechanism design for collaboration [Abernethy+, 2011]
    - Submitted models are shared
    - Participants receive payment according to performance improvement



# Collaborative descriptive analysis: Platforms for social data analysis (1)

- **manyeyes** [Viégas+, 2007]
  - Platform for data visualization and discussion



# Collaborative descriptive analysis: Platforms for social data analysis (2)

- Platform with systematic collaboration scheme [Willett+, 2012]
  - Several subtasks are generated to involve the crowds
    - chart-reading, annotation, etc.



Visualization

**b** 1. What **occupation(s)** are shown in this chart?

**c** 2. What **gender(s)** are shown in this chart?

**c** 3. Click and drag on the chart to highlight strong **peaks and/or valleys** (if any exist).

**d** 4. Explain **why** the strong **peak or valley** highlighted in the chart might have occurred.

**e** 5. Provide the URL of the source of the data.

Chart-reading subtask:  
Asking to enter chart labels

Annotation subtask:  
Asking to highlight strong peaks

## Part IV: Future direction



## Future direction:

### For reliable and advanced crowd analytics system

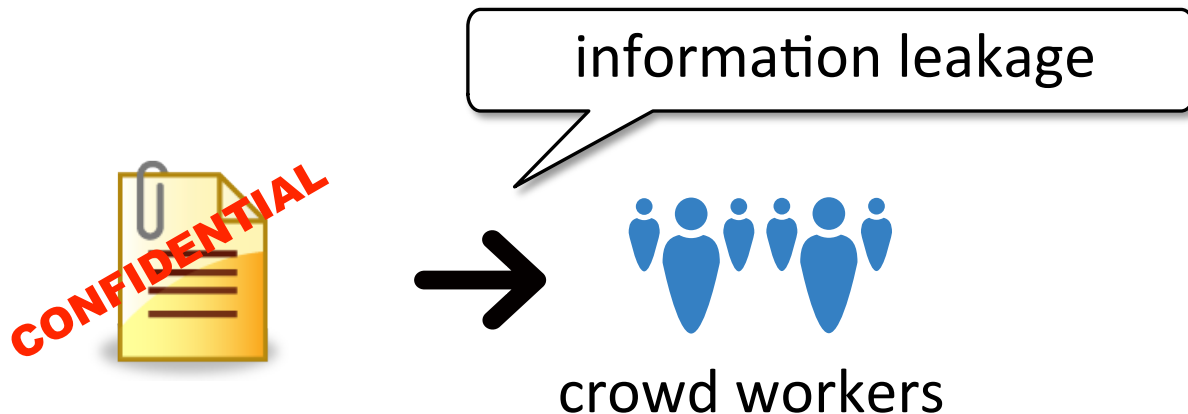
---

- Reliable crowdsourcing system
  - Addressing security and privacy issues
- Closer corporation between machine and human
  - Human computation
  - Workflow control
- Data science education

# Security and privacy issues:

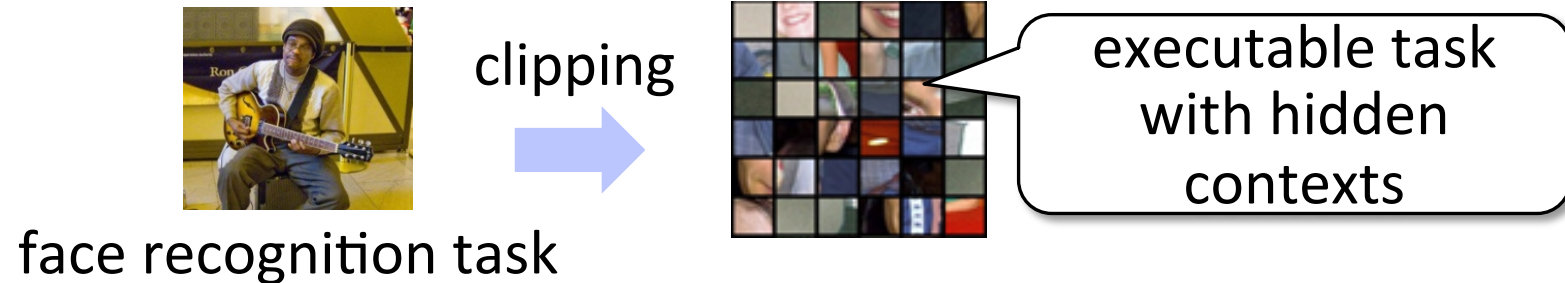
## Concerns about disclosing sensitive information

- Task instances may contain sensitive or privacy information
  - E.g., Face images in image classification tasks, recording of business meeting in audio transcription tasks
- Can we apply the concept of Privacy preserving data mining (PPDM) to guard the security of task instances?
  - Show task instances to workers while sensitive information is preserved

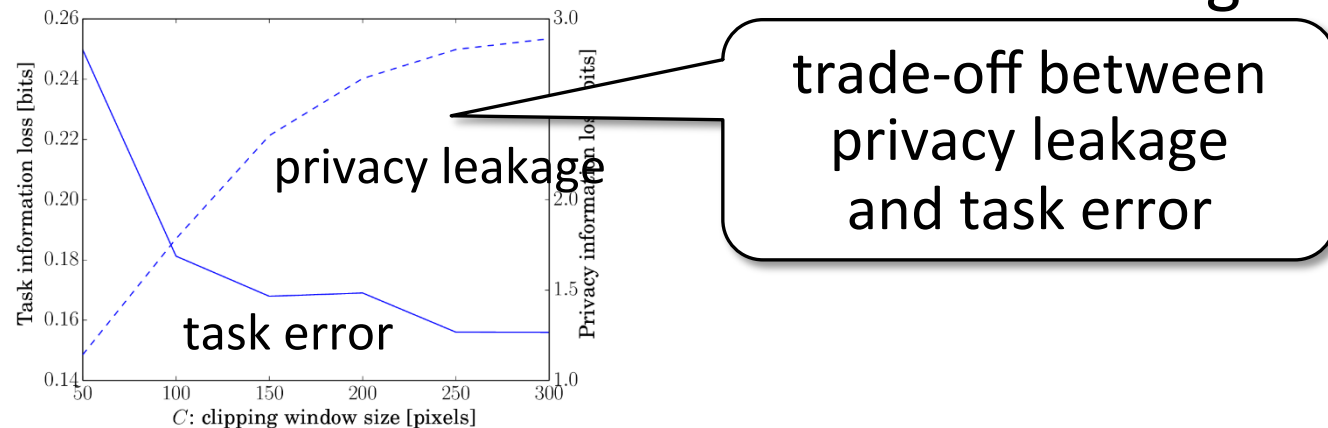


# Privacy-preserving crowdsourcing: Task conversion approach to hide task contexts

- Task conversion: Convert task instances into different forms to hide contexts in task instances
  - Example: Task clipping in face recognition tasks



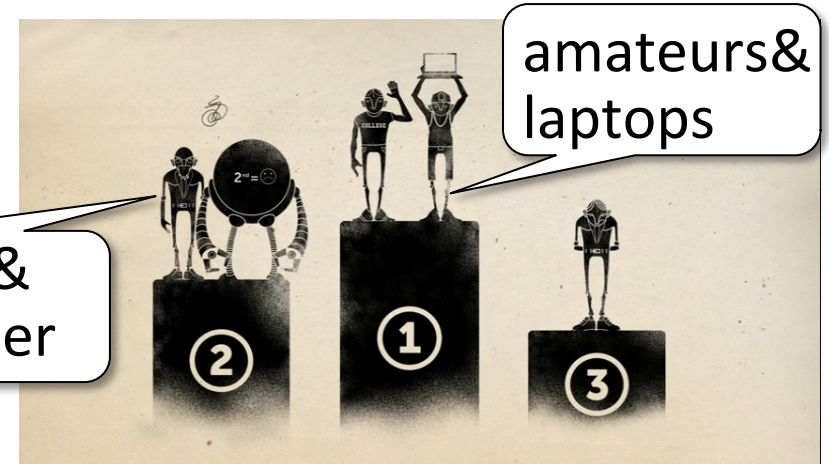
- Balance between task error and information leakage



# Human computation:

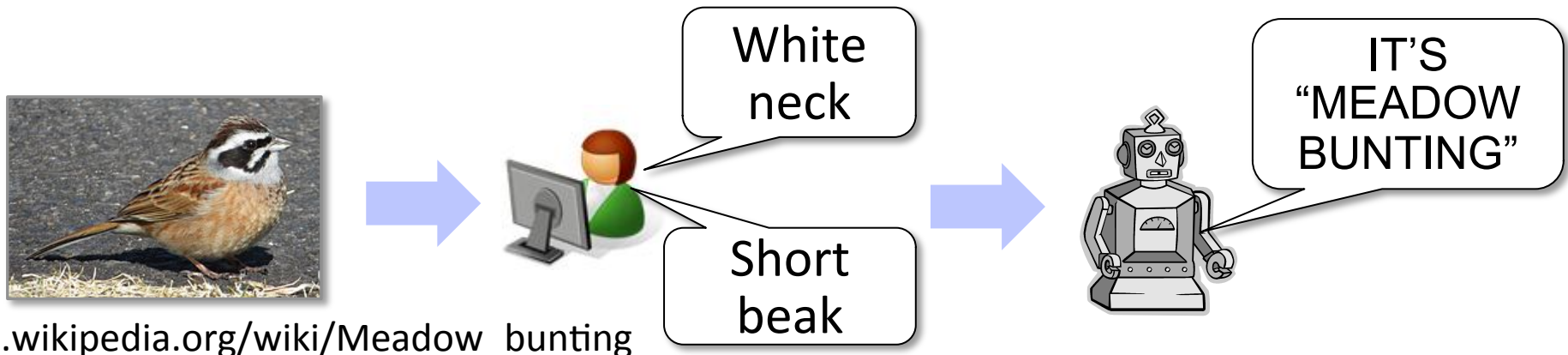
## Combine machine power and human intellect

- Human computation
  - Combine computation power and human intellect
  - Tackle problems that neither can solve alone
- Freestyle chess tournament in 2005
  - “Amateurs&laptops” beat “grandmaster&supercomputer”
  - Good coordination of human and machine is important



# Human computation for data analytics (1): Classification with crowd-generated features

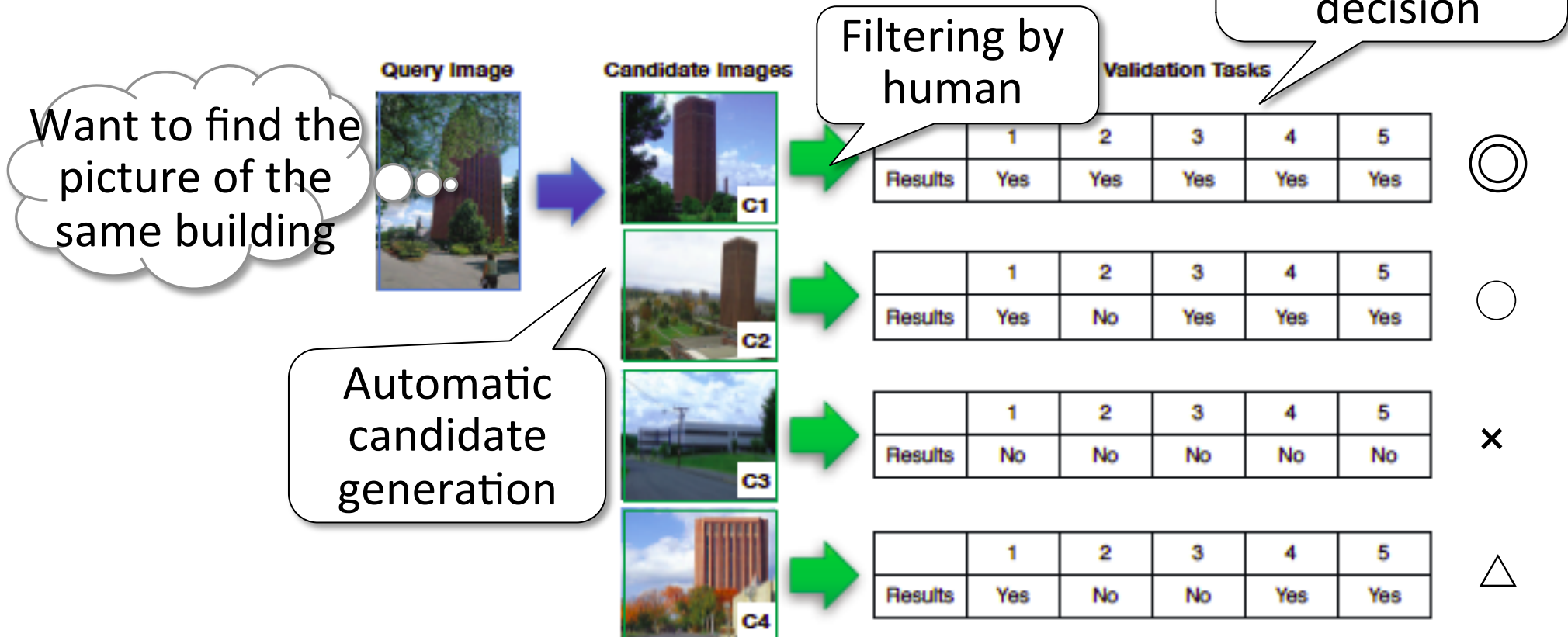
- Problem that neither machines nor humans can solve alone
  - e.g., visual recognition of wild birds
    - Difficult for both non-experts and computers
- Good coordination of human and machine solve the problem
  - Non-experts can extract several features (e.g., color of neck)
  - Machines classify images using the extracted features



# Human computation for data analytics (2): Filtering and re-ranking of samples by crowds

- Humans accurately discover the desired images but the scalability is poor

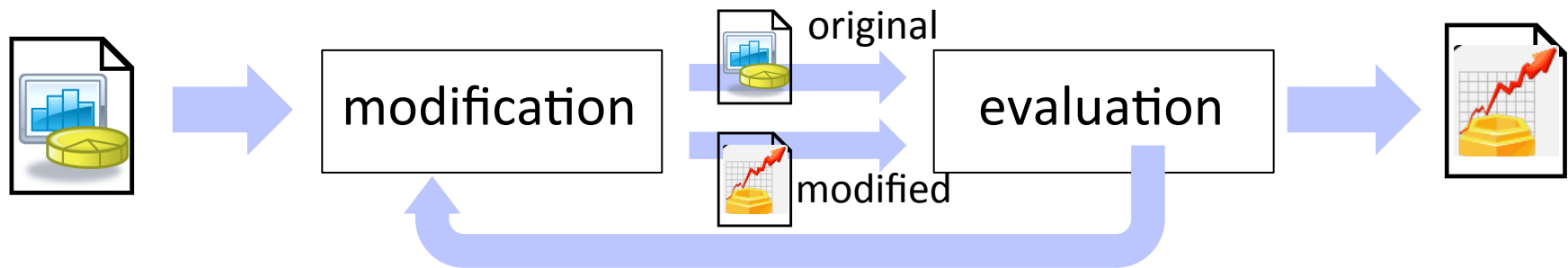
➔ Combination of human and machine



# Workflow control

## Modeling and control by Markov decision processes

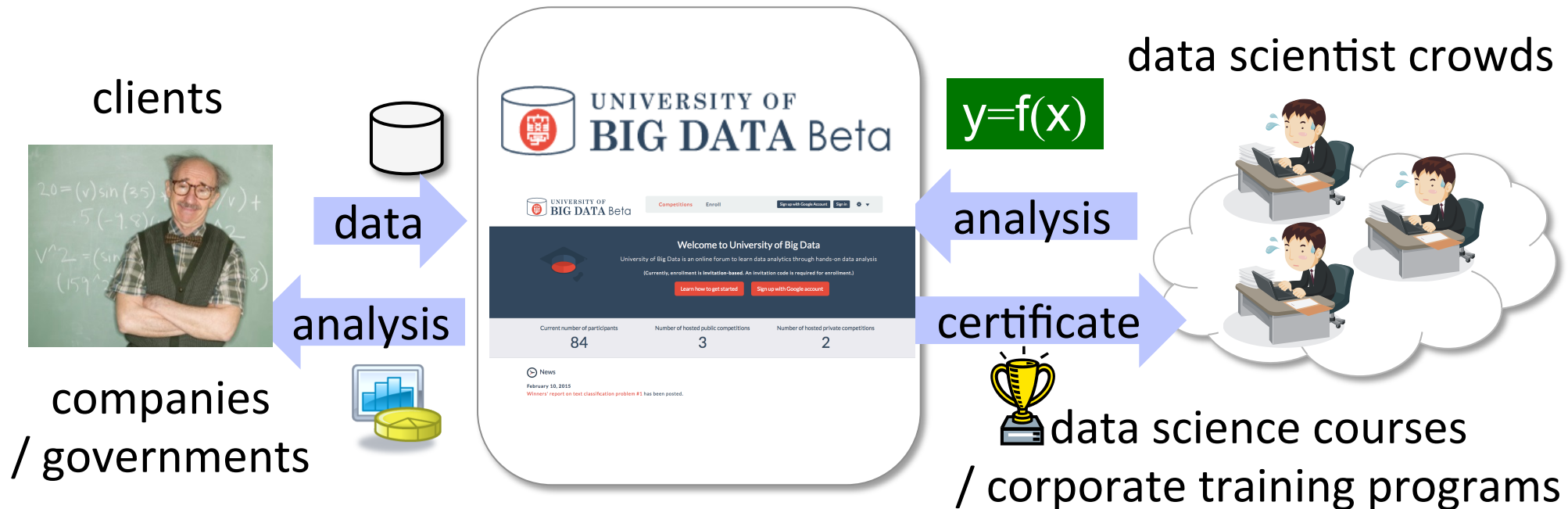
- Complex workflow in crowdsourcing
  - E.g., Report submission -> modification <-> evaluation
    - Is modification required? Does modification improve the quality? Is the quality evaluation reliable?
- Workflow control
  - TurKontrol: automated control by MDPs [Dai+, 2011]
  - Turkomatic: manual control by crowds [Kulkarni+, 2011]





# Data science education: “Educational” data analysis competition platform

- Data science education through competition-style data analysis  
Not *Classwork*, Not with *monetary rewards* (e.g. Kaggle)
- Eco-system with data providers and learners





# Summary

---

Part I: Crowdsourcing for data analytics (Hisashi Kashima)

Crowds for executing the labor-intensive data analysis process

Part II: Crowdsourcing for datafication (Satoshi Oyama)

Gamification and introducing redundancy

Part III: Crowdsourcing for analysis (Yukino Baba)

Predictive modeling competitions and peer-reviewing

Part IV: Future direction (Yukino Baba)

Security/privacy, human computation, and data science education