KYOTO UNIVERSITY

Crowdsourcing for Big Data Analytics

Hisashi Kashima (*Kyoto University*)
Satoshi Oyama (*Hokkaido University*)
Yukino Baba (*Kyoto University*)

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

Crowdsourcing for data analytics

Part I: Crowdsourcing for data analytics (Hisashi Kashima)

Part II: Crowdsourcing for datafication (Satoshi Oyama)

Part III: Crowdsourcing for analysis (Yukino Baba)

Part IV: Future direction (Yukino Baba)

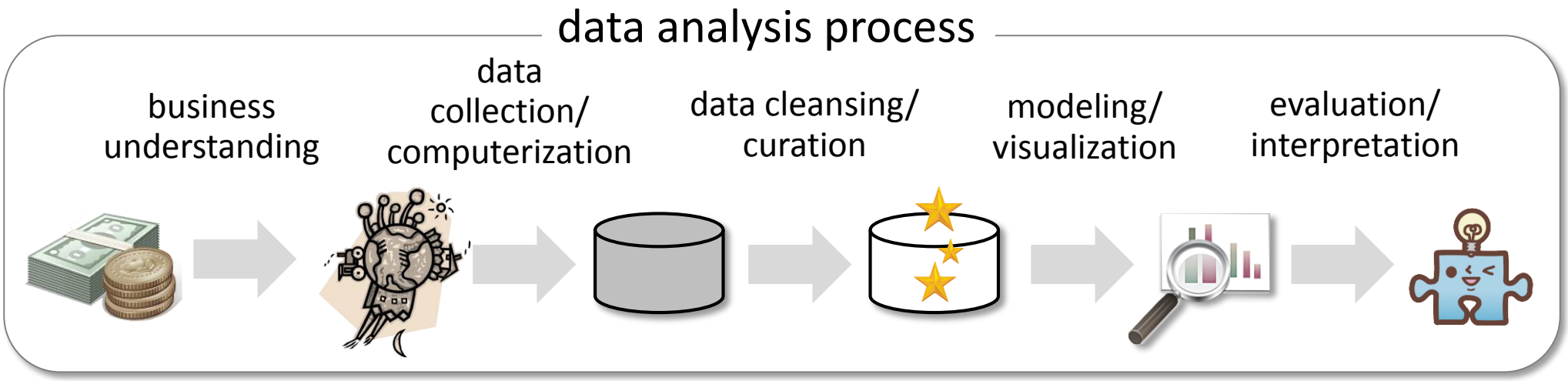
Tutorial slides are found at

http://goo.gl/Amif93

Part I: Crowdsourcing for data analytics

Big challenge in big data analytics:
Manpower bottleneck

- Automatic data analysis techniques (e.g. machine learning) are often considered as main components of data analytics
- Data analysis is heavily labor intensive
  - Manual processing dominates a large portion of data analysis process
  - 1990s-2000s: introduction of data mining techniques and data analysis process standards (e.g., CRISP-DM)



Big shortage of data scientists:
Implies labor intensity of data analysis

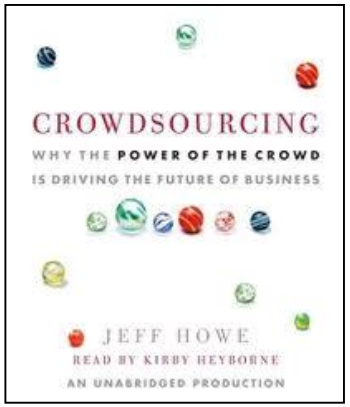
- *"By 2015, 4.4 million IT jobs globally will be created to support big data"*, but *"only one-third of the IT jobs will be filled"*
  - Peter Sondergaard (Senior VP at Gartner)
- *"Data Scientist: The Sexiest Job of the 21st Century"*
  - Thomas H. Davenport and D.J. Patil, Harvard Business Review
- These statements imply the labor intensity of data analysis

Rise of crowdsourcing:
On-demand access to massive on-line labor

- Crowdsourcing: Outsourcing human-intelligence tasks to a large group of unspecified people via Internet

  ⇔ "outsourcing" offers tasks to specified contractors

- Advantage of crowdsourcing:
  - Volume: Access to a massive amount of human intellects
  - Quality: "wisdom of crowds"



Howe, J. Crowdsourcing. 2004

Two types of crowdsourcing:
Explicit crowdsourcing and implicit crowdsourcing

1. Explicit crowdsourcing: directly asks for contributions
   - Marketplace (e.g., Amazon Mechanical Turk)
   - Volunteer (e.g., citizen science such as Galaxy Zoo)
2. Implicit crowdsourcing:
   embeds tasks into other forms in order to motivate worker participations
   - Unavoidable tasks (e.g., reCAPTCHA)
   - Games with purposes (e.g., ESP game)

Crowdsourcing marketplace:
Online labor markets for explicit crowdsourcing

- Crowdsourcing marketplaces:
  Web services for connecting requesters with workers
  - Advantages for requesters: Access to on-demand workforce
  - Advantages for workers: New work style unbound by time or place
- Emergence of online crowd-labor marketplaces
  - Mechanical Turk, oDesk, Clickworker, …

Amazon Mechanical Turk:
World-biggest microtask crowdsourcing marketplace

- amazonmechanical turk Artificial Artificial Intelligence Since established in 2005, a de-facto standard of crowdsourcing marketplace in computer science
  - 0.5 million workers from 190 countries (in 2011)
  - can access human intelligence on-demand
- Automatic creation of tasks through API

Extraction of purchase information form a receipt

6-cent reward

Extract purchased items from a shopping receipt

...w a HIT in this group

Requester: Jon Brelig **HIT Expiration Date:** Oct 31, 2013 (6 days 23 hours) **Reward:** $0.06

Time Allotted: 2 hours **HITs Available:** 24394

Description: Transcribe all of the purchased items and total from a shopping receipt

Keywords: image, receipt, categorize, transcribe, extract, data, entry, transcription, text, easy, qualification, secure, prod

Qualifications Required: None

Task types in crowdsourcing marketplaces:
Simple, structured tasks to complex, unstructured tasks

more complex

Very easy tasks:
Image labeling, entity resolution, format checking…

Output format: Yes/No

Easy tasks: Restaurant reviews, web service tests, …

Output format: Multiple-choice or short sentences

Microtasks: Popular in computer science

Tasks requiring some expertise: Logo design, report writing, …

Output format: Images or sentences

Complex tasks requiring high level of expertise:
Web/software developments, professional work, …

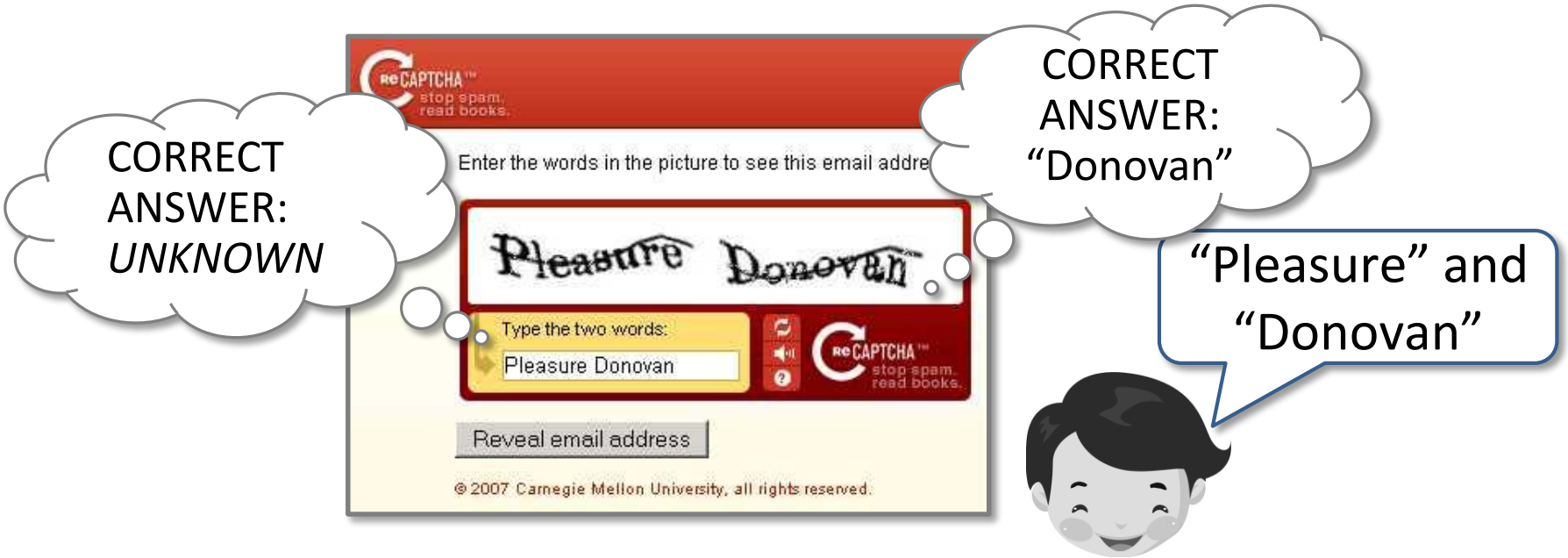
Output format: System, codes, documents

reCAPTCHA:
A ”Turing test” distinguishing humans and machines

- System shows two text images to a user and asks to read both
- Character recognition is difficult for computers but easy for humans, only human can pass the test

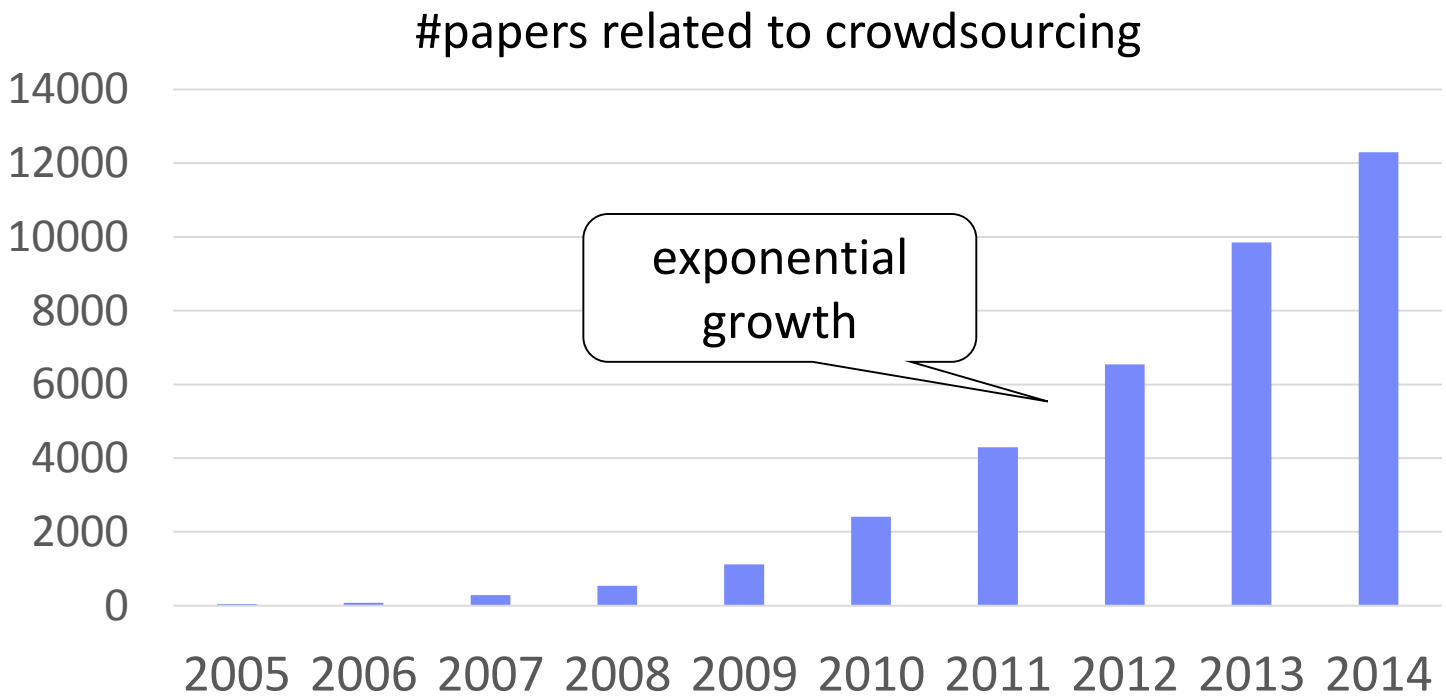
Implicit crowdsourcing of reCAPTCHA:
Use system logs for document digitization

- The system knows the correct answer of only one of the two words
- The other is a word that OCR systems fail to recognize
  - Users are forced to help document digitization



Trend in academic research:
Exponential growth of crowdsourcing research

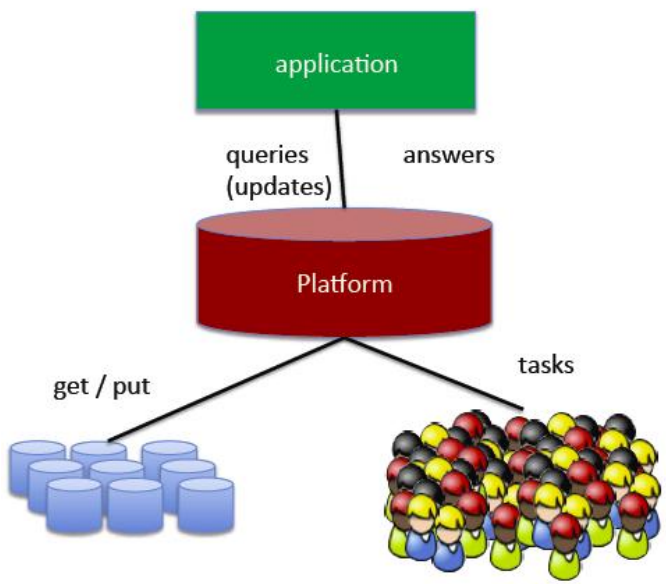
- 2005: Amazon Mechanical Turk was launched
- 2006: “crowdsourcing” and ”human computation” were advocated
- 2013: 1st conference on human computation and crowdsourcing (HCOMP)



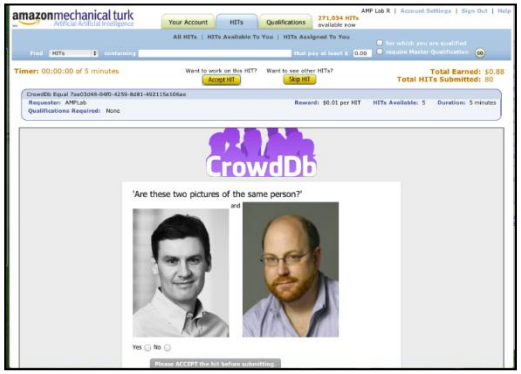
Counted using Google Scholar

Crowdsourcing in computer science:
Popular use in NLP, CV, HCI, DB, IR, ML/DM, …

- NLP: text understanding, annotation, language translation
- CV: image understanding, annotation, detection
- DB/IR: data generation/integration, search, evaluation
- ML/DM: data collection/annotation



Microtasks are automatically issued for SQL execution

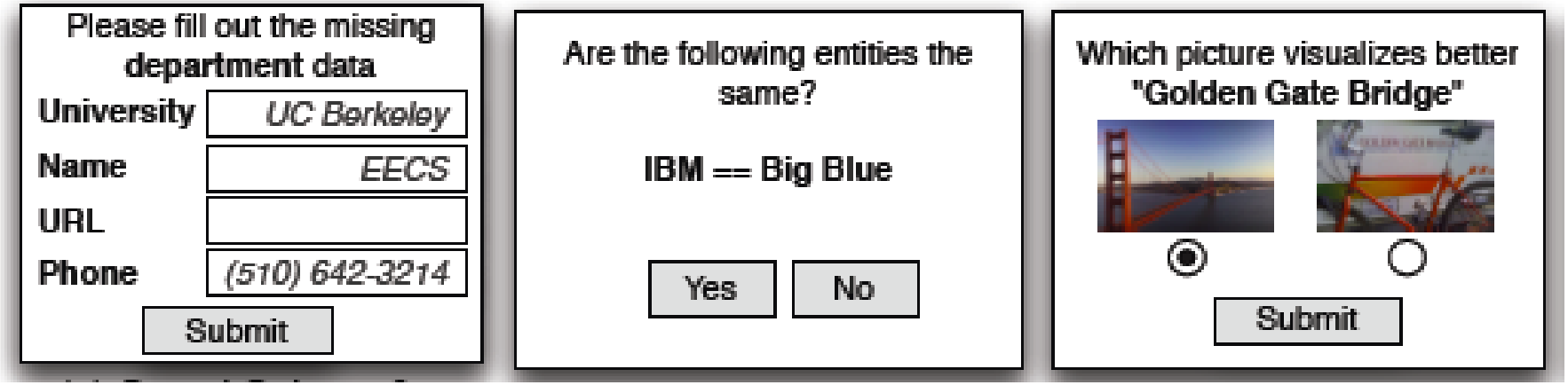


Franklin *et al*. CrowdDB: Answering Queries with Crowdsourcing. *VLDB* 2011.

Example of crowdsourcing in DB research: Data generation and comparison

- CrowdDB uses crowdsourcing for

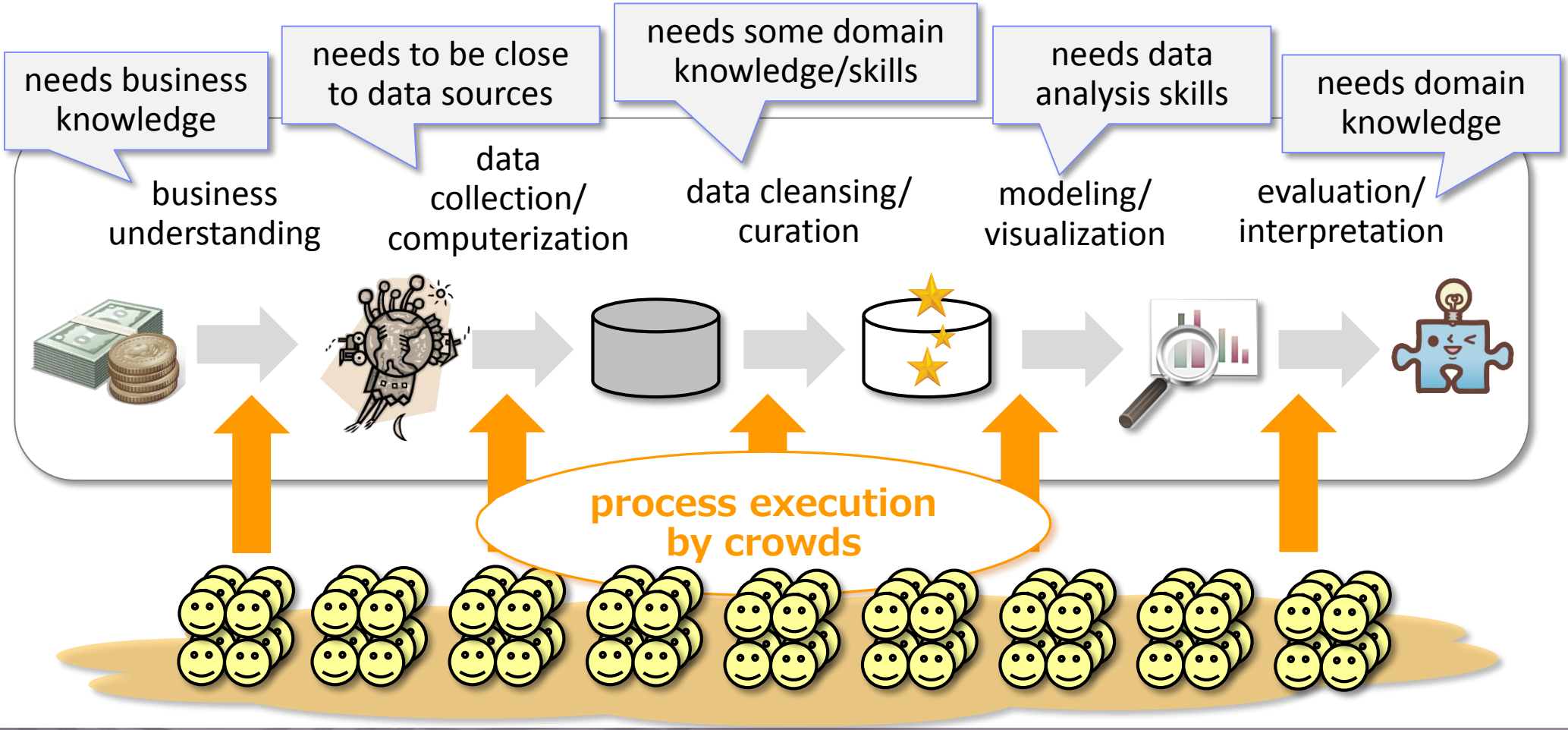
1. Data generation
2. Data comparison:
   - Identity test of two data instances
   - Sorting data instances



Franklin *et al*. CrowdDB: Answering Queries with Crowdsourcing. *VLDB* 2011.

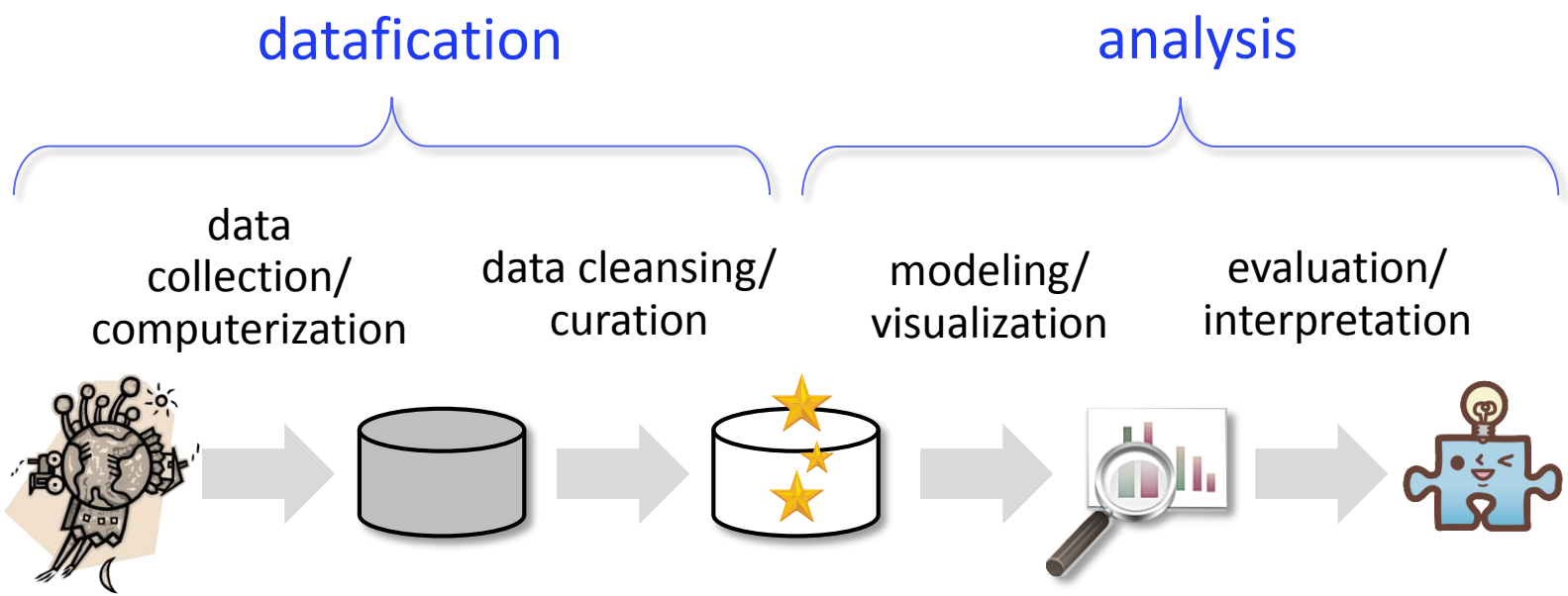
Crowdsourcing for data analytics:
Crowdsourced execution of data analysis process

- Use the power of crowds with various knowledge/skills to execute the labor-intensive data analysis process



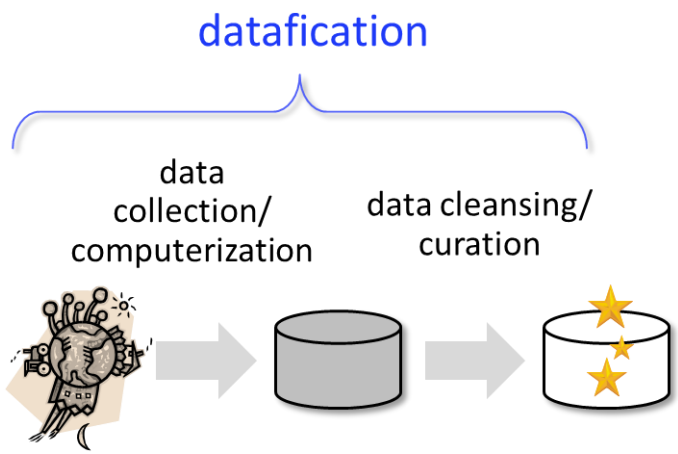
Crowdsourcing usages in data analysis:
Datafication and analysis

- Data analysis process is divided into two parts:
  1. Datafication: data collection, data cleansing, annotation, …
  2. Analysis: data modeling, visualization, evaluation, …



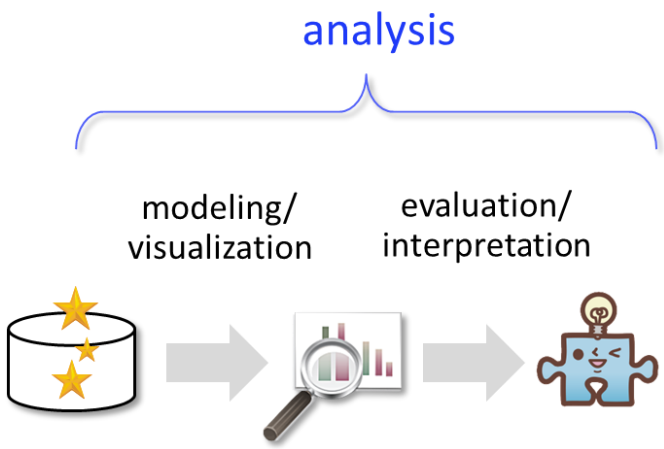
Crowdsourcing for datafication:
Microtasks for data collection, annotation, and cleansing

- *Microtasks* for data annotation and cleaning
  - Suited for simple human-intelligence tasks:
    - Image labeling, speech recognition, …
  - Small payment for a small piece of work
    - Large-scale datasets can be processed with reasonable costs



Crowdsourcing for analysis:
Competitions, peer reviewing, …

- *Competitions* for data modeling
  - Suited for creative/complex tasks:
    - Design, software development, analysis, …
  - Winner-takes-all payment
    - A few high-quality results can be obtained
- *Peer reviewing* for assessing results



Technical issues:
Incentive design and quality control

- Incentive design
  - Crowdsourcing workers are not necessarily well-motivated
  - *Gamification* and reward optimization
- Quality control
  - Crowd-workers have different skills and motivations, and sometimes they are malicious
  - Quality of crowdsourcing results is uneven
  - *Statistical quality control methods*

Following parts…:
Crowdsourced datafication and analysis