**enetCollect Annual Meeting**

# Quality control in crowdsourcing

## A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions

POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA

**Florian Daniel**

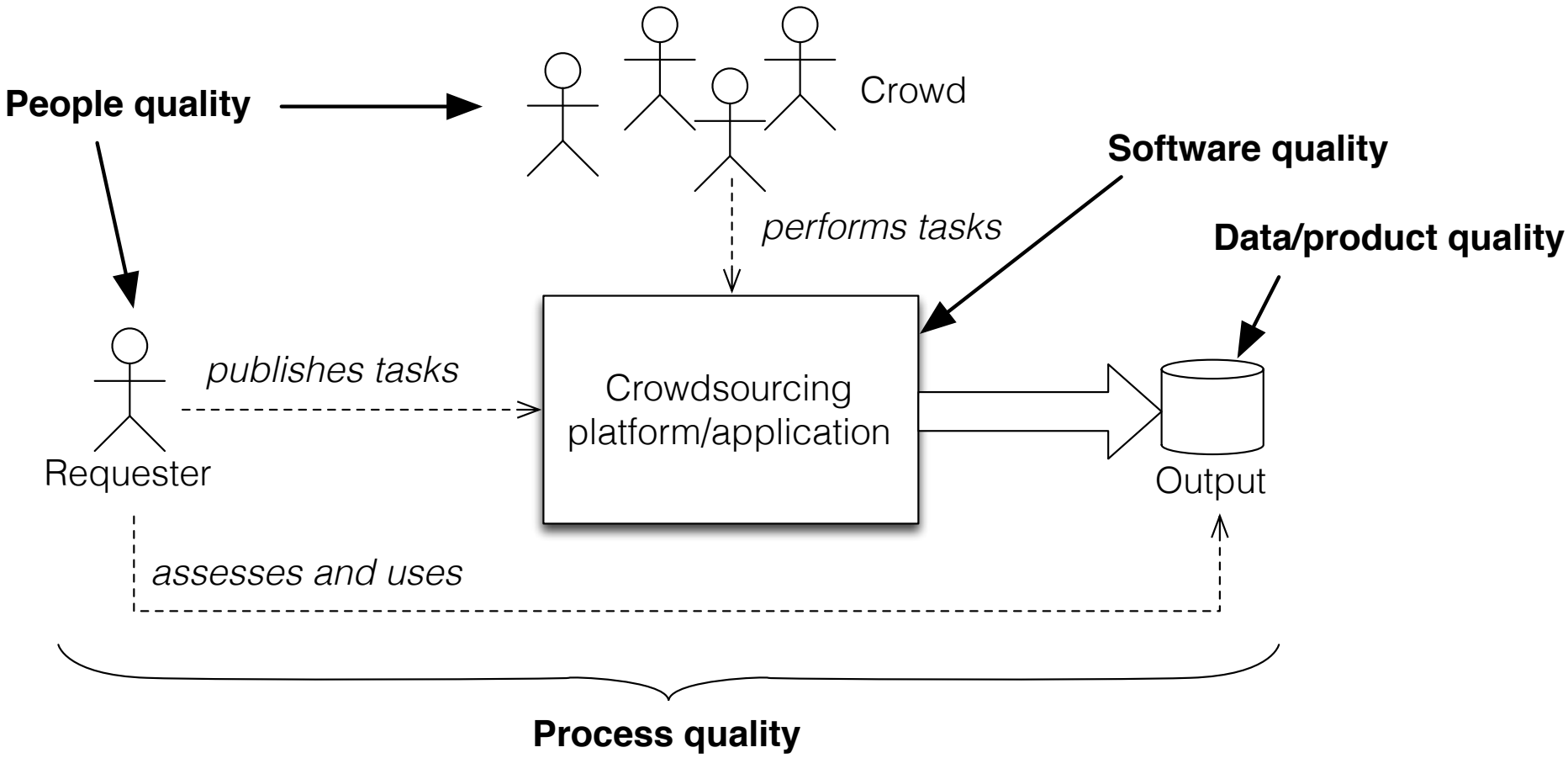
florian.daniel@polimi.it

My **goal** today = mini intro to quality control in crowdsourcing

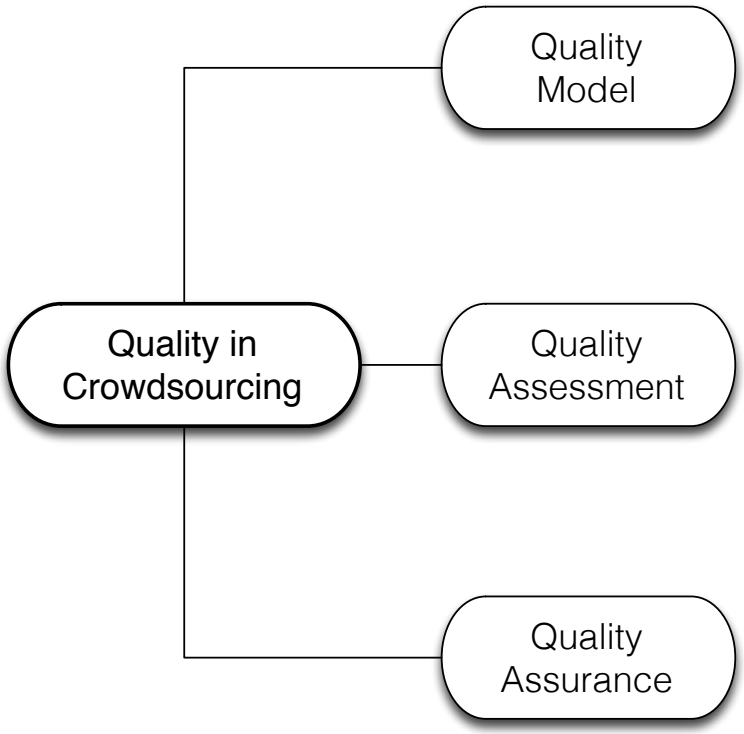
**Quality** of a crowdsourced task = the extent to which the output meets or exceeds the requester's expectations

Talk based on: F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, M. Allahbakhsh. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques and Assurance Actions. *ACM Computing Surveys* 51(1), Article No. 7, April 2018.

# **Types** of quality in crowdsourcing



## High-level **taxonomy**



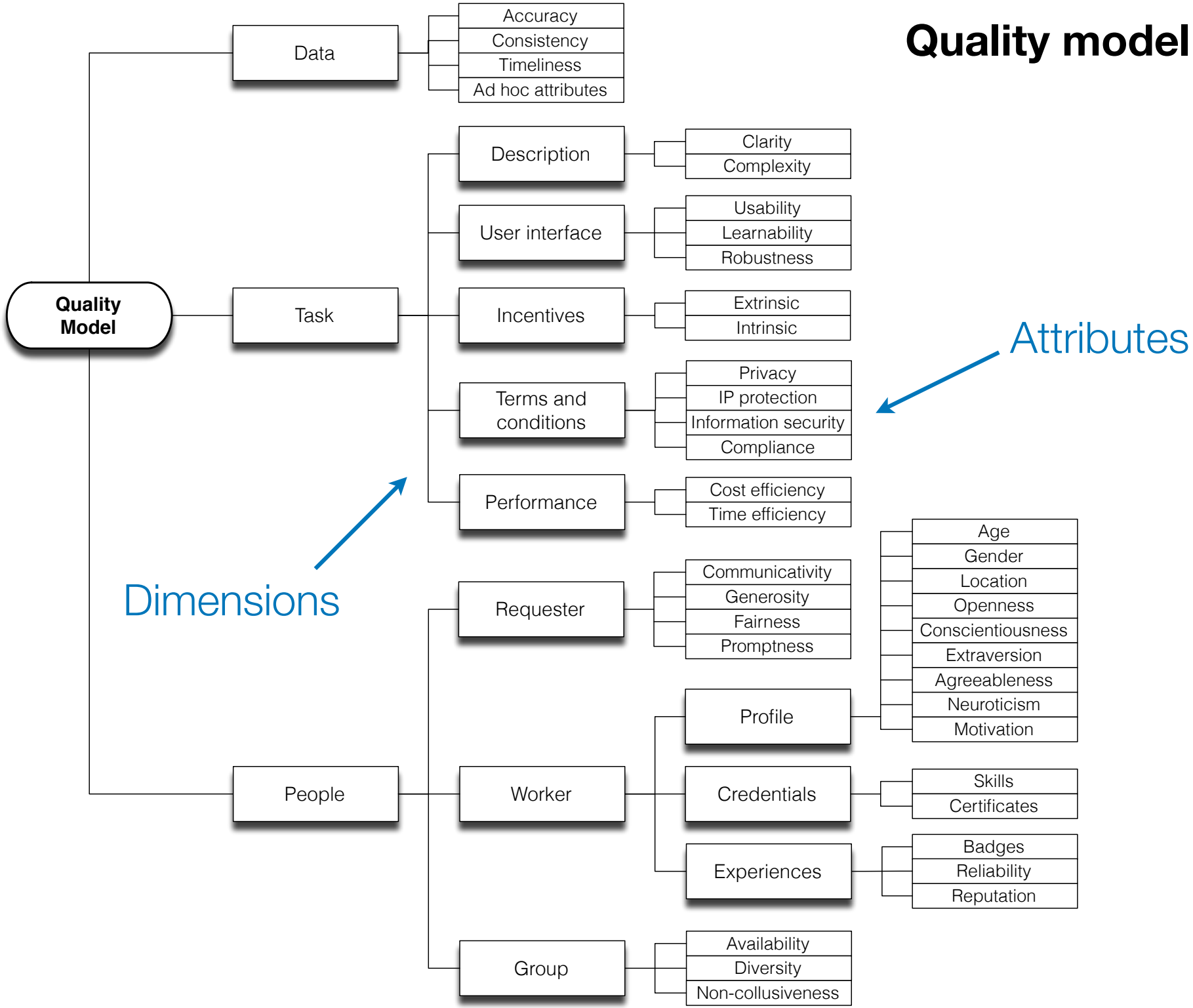

## **Methodology**

Bottom-up construction of models

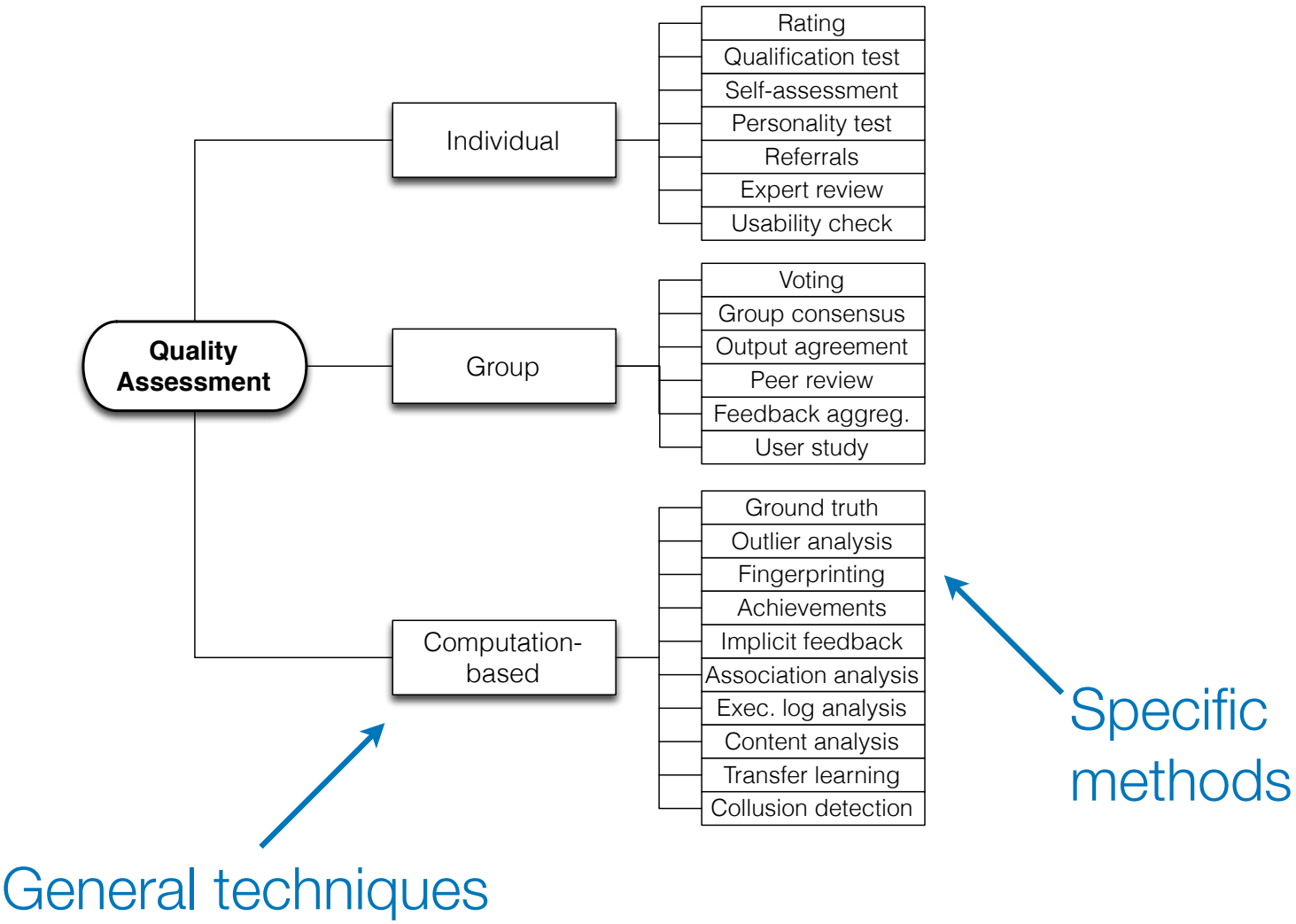
Literature selection: 257 papers analyzed

Analysis of state of the art: 14 platforms positioned inside taxonomy

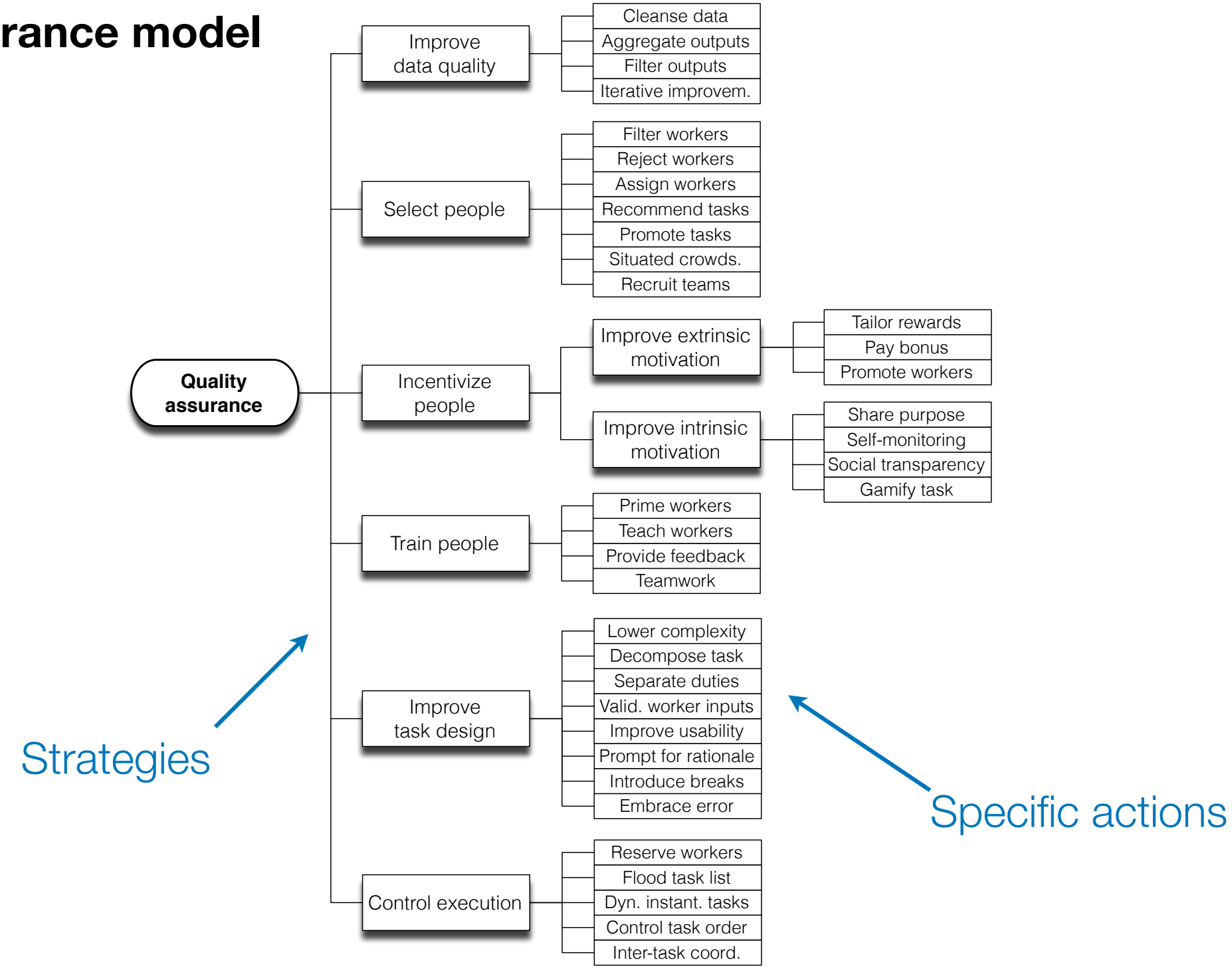
# Quality model



# Assessment model



# Assurance model



# State of the art

(as of end of 2017)



now 



discontinued


now 


**Turkit** (Little et al. 2010c)

**CrowdWeaver** (Kittur et al. 2012)

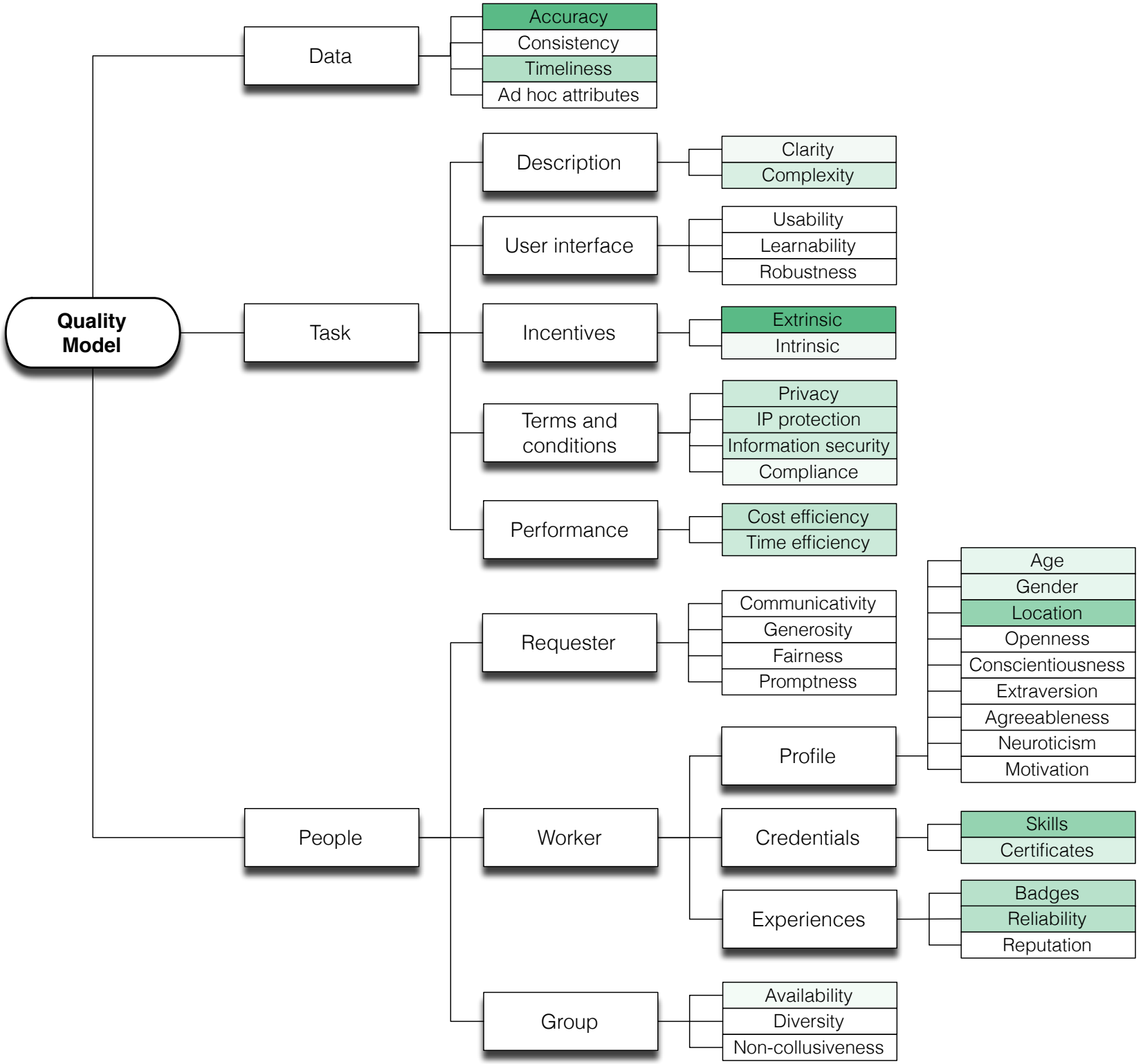
**Jabberwocky** (Ahmad 999 et al. 2011)

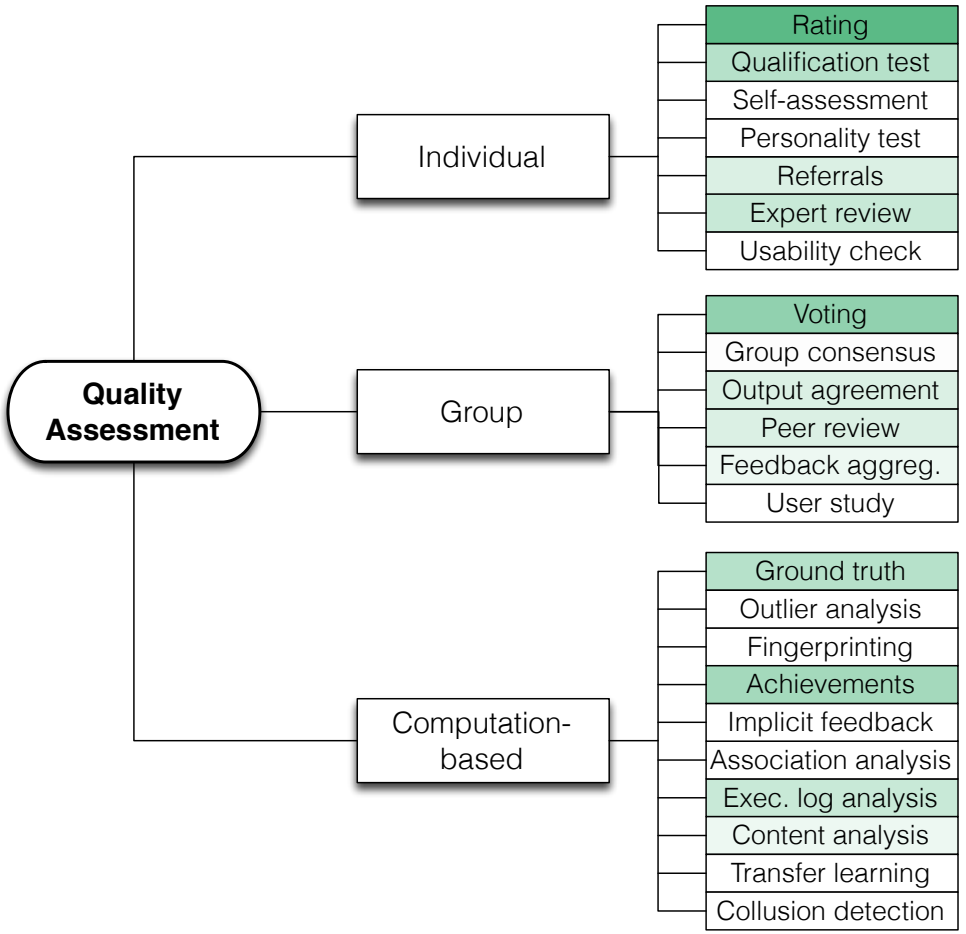
**Turkomatic** (Kulkarni 1003 et al. 2012a)

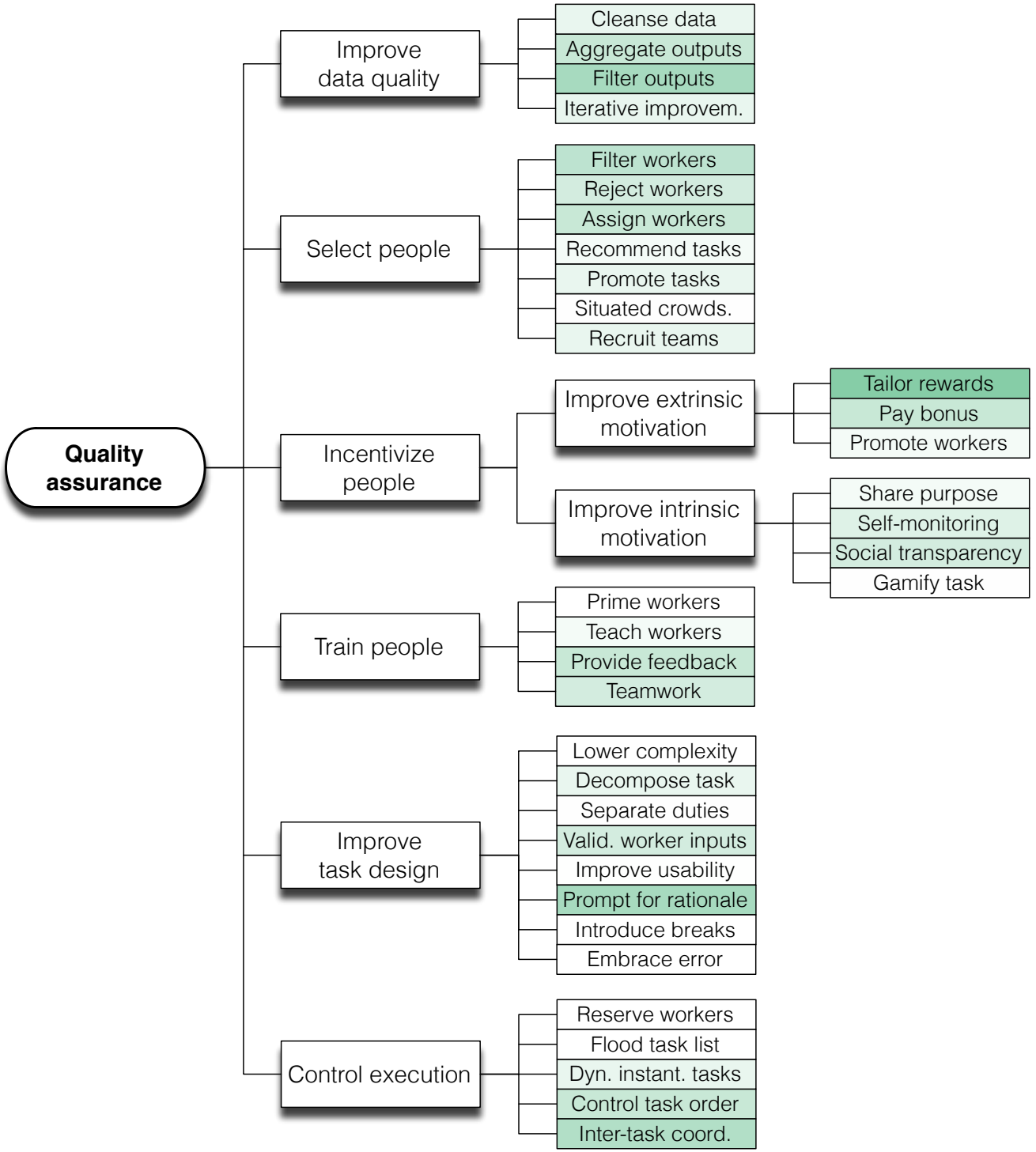


**AskSheet** (Quinn and Bederson 2014)

**CrowdForge** (Kittur et al. 2011)



Quality Model
Data
Accuracy
Consistency
Timeliness
Ad hoc attributes
Task
Description
Clarity
Complexity
User interface
Usability
Learnability
Robustness
Incentives
Extrinsic
Intrinsic
Terms and conditions
Privacy
IP protection
Information security
Compliance
Performance
Cost efficiency
Time efficiency
People
Requester
Communicativity
Generosity
Fairness
Promptness
Worker
Profile
Age
Gender
Location
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism
Motivation
Credentials
Skills
Certificates
Experiences
Badges
Reliability
Reputation
Group
Availability
Diversity
Non-collusiveness


Quality Assessment
Individual
Group
Computation-based
Rating
Qualification test
Self-assessment
Personality test
Referrals
Expert review
Usability check
Voting
Group consensus
Output agreement
Peer review
Feedback aggreg.
User study
Ground truth
Outlier analysis
Fingerprinting
Achievements
Implicit feedback
Association analysis
Exec. log analysis
Content analysis
Transfer learning
Collusion detection

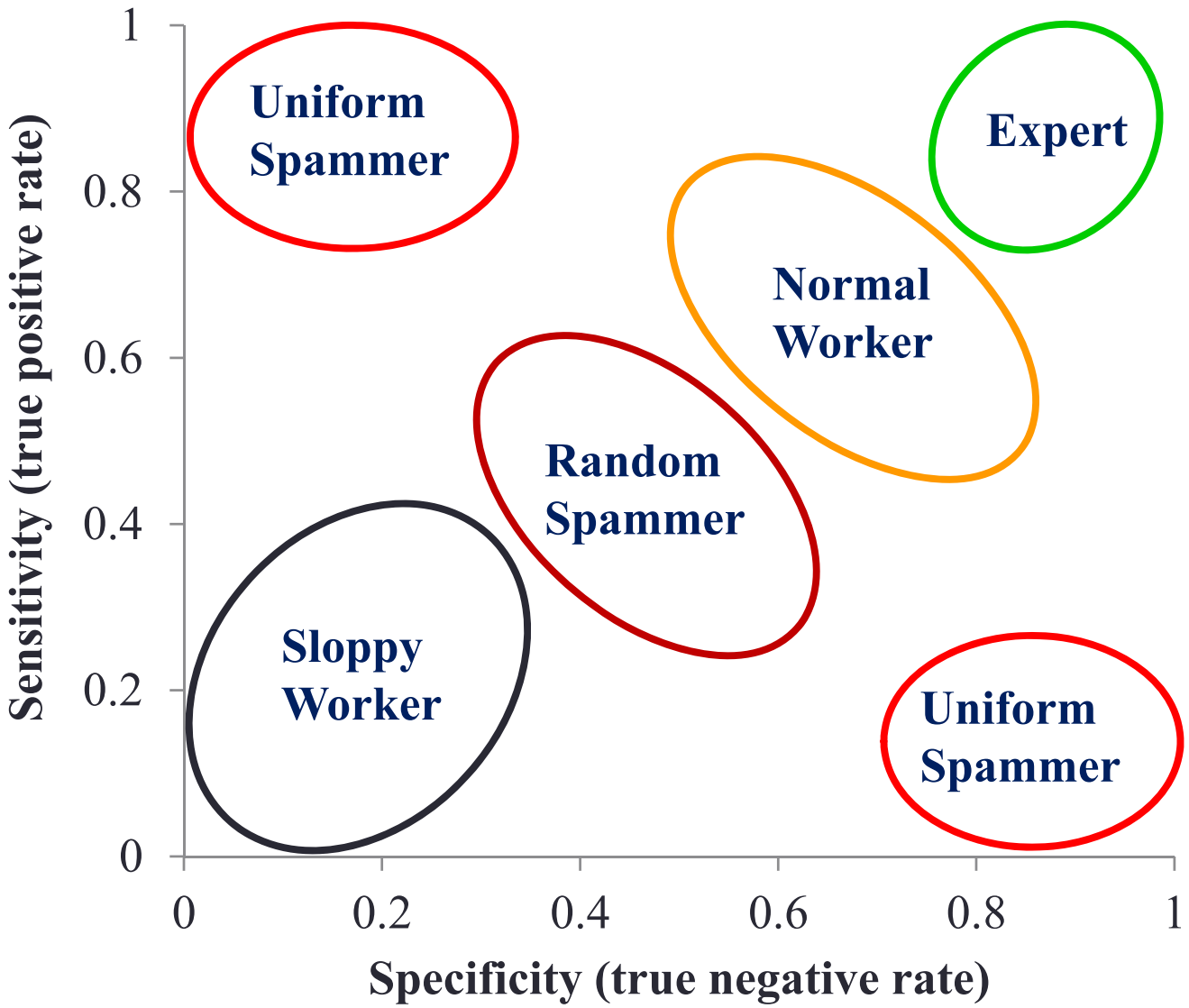

Quality assurance
Improve data quality
Cleanse data
Aggregate outputs
Filter outputs
Iterative improvem.
Select people
Filter workers
Reject workers
Assign workers
Recommend tasks
Promote tasks
Situated crowds.
Recruit teams
Incentivize people
Improve extrinsic motivation
Tailor rewards
Pay bonus
Promote workers
Improve intrinsic motivation
Share purpose
Self-monitoring
Social transparency
Gamify task
Train people
Prime workers
Teach workers
Provide feedback
Teamwork
Improve task design
Lower complexity
Decompose task
Separate duties
Valid. worker inputs
Improve usability
Prompt for rationale
Introduce breaks
Embrace error
Control execution
Reserve workers
Flood task list
Dyn. instant. tasks
Control task order
Inter-task coord.

Quality control in 


Lionel:

> "…**cross-match the answers** of students to questions we don't have the answer for"
>
> "directly or indirectly **ask boolean questions** to the students (e.g. 'Does the student think that this word is a verb?', 'Does the student think that this translation is ok?' etc.)"
>
> "…focus on **aggregation methods** for answers to boolean questions"

# **Binary/Boolean labeling**: worker types




Hung, N. Q. V., Tam, N. T., Tran, L. N., & Aberer, K. (2013, October). An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering* (pp. 1-15). Springer, Berlin, Heidelberg.

# An Evaluation of Aggregation Techniques in Crowdsourcing

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer

École Polytechnique Fédérale de Lausanne
{quocviethung.nguyen,tam.nguyenthanh,ngoc.lam,karl.aberer}@epfl.ch

**Abstract.** As the volumes of AI problems involving human knowledge are likely to soar, crowdsourcing has become essential in a wide range of world-wide-web applications. One of the biggest challenges of crowdsourcing is aggregating the answers collected from the crowd since the workers might have wide-ranging levels of expertise. In order to tackle this challenge, many aggregation techniques have been proposed. These techniques, however, have never been compared and analyzed under the same setting, rendering a 'right' choice for a particular application very difficult. Addressing this problem, this paper presents a benchmark that offers a comprehensive empirical study on the performance comparison of the aggregation techniques. Specifically, we integrated several state-of-the-art methods in a comparable manner, and measured various performance metrics with our benchmark, including *computation time, accuracy, robustness to spammers,* and *adaptivity to multi-labeling*. We then provide in-depth analysis of benchmarking results, obtained by simulating the crowdsourcing process with different types of workers. We believe that the findings from the benchmark will be able to serve as a practical guideline for crowdsourcing applications.

Comparison of **non-iterative** and **iterative** aggregation techniques

>> "For binary labeling, **Expectation Maximization** is the winner"

# Aggregating Crowdsourced Binary Ratings

Nilesh Dalvi
Facebook, Inc.
Menlo Park, CA
nileshdalvi@gmail.com

Anirban Dasgupta
Yahoo! Labs
Sunnyvale, CA
anirban.dasgupta@gmail.com

Ravi Kumar
Google
Mountain View, CA
ravi.k53@gmail.com

Vibhor Rastogi
Google
Mountain View, CA
vibhor.rastogi@gmail.com

## ABSTRACT

In this paper we analyze a crowdsourcing system consisting of a set of users and a set of binary choice questions. Each user has an unknown, fixed, reliability that determines the user's error rate in answering questions. The problem is to determine the truth values of the questions solely based on the user answers. Although this problem has been studied extensively, theoretical error bounds have been shown only for restricted settings: when the graph between users and questions is either random or complete. In this paper we consider a general setting of the problem where the user–question graph can be arbitrary. We obtain bounds on the error rate of our algorithm and show it is governed by the expansion of the graph. We demonstrate, using several synthetic and real datasets, that our algorithm outperforms the state of the art.

Specific focus on
**binary ratings**

# Adaptive Task Assignment for Crowdsourced Classification

**Chien-Ju Ho, Shahin Jabbari** CJHO@CS.UCLA.EDU, SHAHIN@CS.UCLA.EDU
University of California, Los Angeles

**Jennifer Wortman Vaughan** JENN@MICROSOFT.COM
Microsoft Research, New York City and University of California, Los Angeles

## Abstract

Crowdsourcing markets have gained popularity as a tool for inexpensively collecting data from diverse populations of workers. Classification tasks, in which workers provide labels (such as "offensive" or "not offensive") for instances (such as "websites"), are among the most common tasks posted, but due to human error and the prevalence of spam, the labels collected are often noisy. This problem is typically addressed by collecting labels for each instance from multiple workers and combining them in a clever way, but the question of how to choose which tasks to assign to each worker is often overlooked. We investigate the problem of task assignment and label inference for heterogeneous classification tasks. By applying online primal-dual techniques, we derive a provably near-optimal adaptive assignment algorithm. We show that adaptively assigning workers to tasks can lead to more accurate predictions at a lower cost when the available workers are diverse.

Adaptively assigns tasks to workers to **optimize** overall budget spent >> requires ability to **assign tasks** directly to workers + worker profiles

# Efficient Budget Allocation with Accuracy Guarantees for Crowdsourcing Classification Tasks

Long Tran-Thanh, Matteo Venanzi, Alex Rogers & Nicholas R. Jennings
University of Southampton
{ltt08r,mv1g10,acr,nrj}@ecs.soton.ac.uk

## ABSTRACT

In this paper we address the problem of budget allocation for redundantly crowdsourcing a set of classification tasks where a key challenge is to find a trade–off between the total cost and the accuracy of estimation. We propose CrowdBudget, an agent–based budget allocation algorithm, that efficiently divides a given budget among different tasks in order to achieve low estimation error. In particular, we prove that CrowdBudget can achieve at most $\max\left\{0, \frac{K}{2} - O\left(\sqrt{B}\right)\right\}$ estimation error with high probability, where $K$ is the number of tasks and $B$ is the budget size. This result significantly outperforms the current best theoretical guarantee from Karger *et al.* In addition, we demonstrate that our algorithm outperforms existing methods by up to 40% in experiments based on real–world data from a prominent database of crowdsourced classification responses.

**Majority voting** based optimization without the need for direct task assignment

# My impression

The problem is **not just aggregating** outputs!

Quality is a holistic problem that is determined by **all aspects** of a crowdsourced task

- Quality of input data
- Quality of task design
- Quality of people
- Quality of output processing

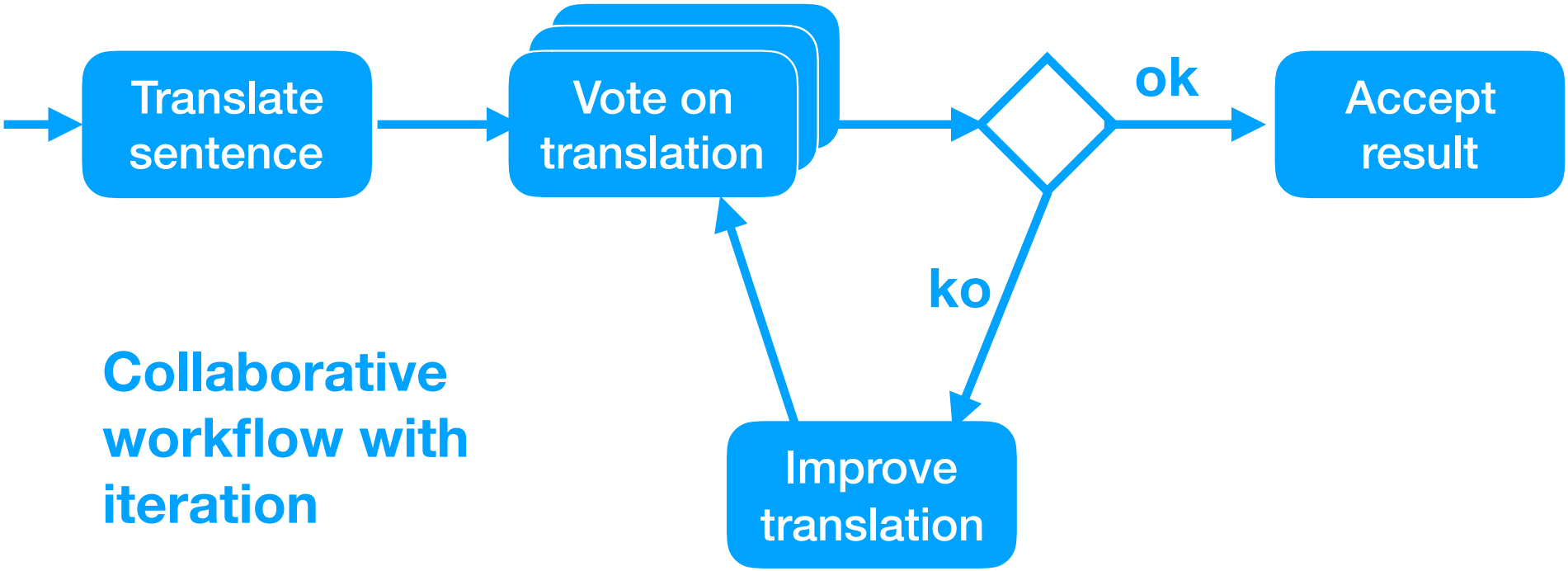
Each crowdsourced task is an own **experiment** and has own quality control requirements

>> iterative development of tasks

“Does the student think that this **translation is ok**?”


(and hope for it)

vs.




**Collaborative workflow with iteration**

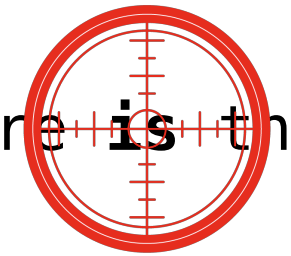
“Does the student think that this **word is a verb**?”

```
Is the following word a verb?
“is” [yes|no]
```

vs.

**Gamification**

Where **is** the verb in this sentence?



Assign **1 point** if the verb is correctly identified.

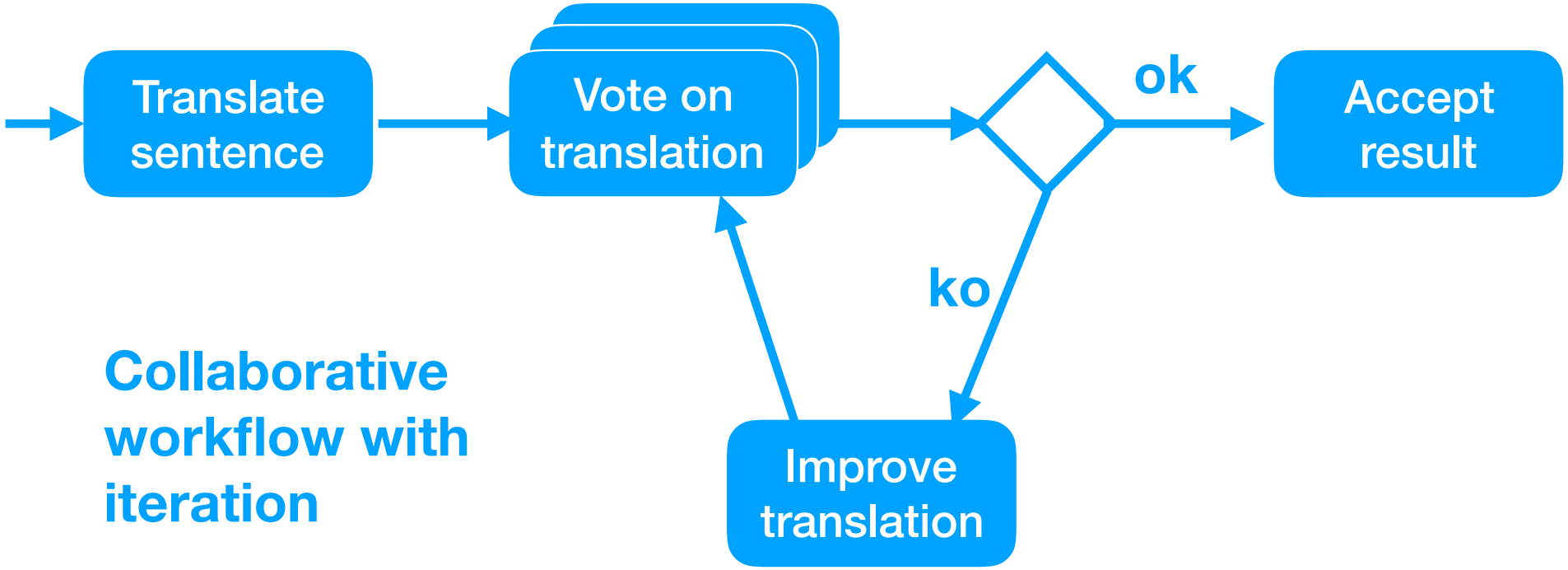
Use common **MV** to decide on correctness of verb.

Assign **3 more points** to the player who identified it first.

Publish **top scorers / a leaderboard** (fosters competition).

Award **badges** for achieved milestones.

# Mixed collaboration and gamification




Give **n points** for a translation that obtains n positive votes.

Give **1 point** to votes that are the majority vote, 0 otherwise.

Publish **ranking**.

In short, use **all registers** you have to assure quality!

And keep measuring and checking…