Towards Human-Enhanced Data Management Systems

Alessandro Bozzon - TU Delft
Software and Computer Technology Department
Faculty of Electrical Engineering, Mathematics, and Computer Science (EEMCS)

Website: www.alessandrobozzon.com
E-mail: a.bozzon@tudelft.nl
Twitter: @aleboz

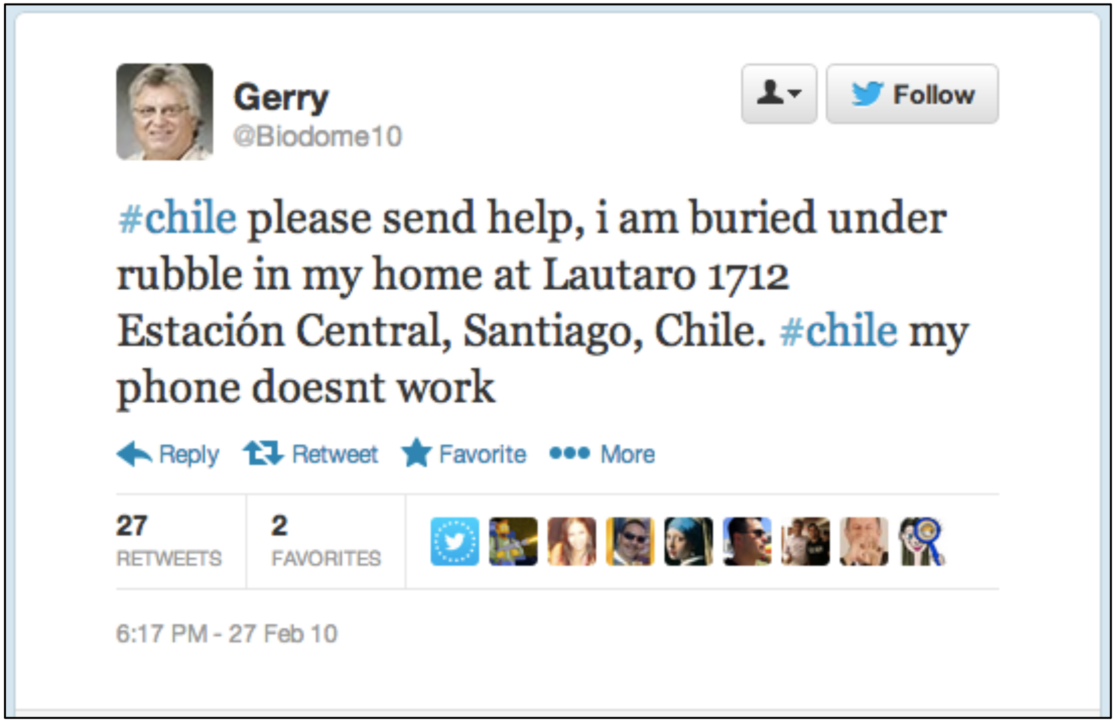
How can data management systems meet the challenges of next generation knowledge- and data-intensive applications?

Big (Ambiguous) Data

Flash Mob Vs. Riot

http://www.telegraph.co.uk/technology/facebook/4542840/Flash-mob-mimicks-T-Mobile-advert-and-closes-train-station.html

Big (False) Data



Gerry
@Biodome10

#chile please send help, i am buried under rubble in my home at Lautaro 1712 Estación Central, Santiago, Chile. #chile my phone doesnt work

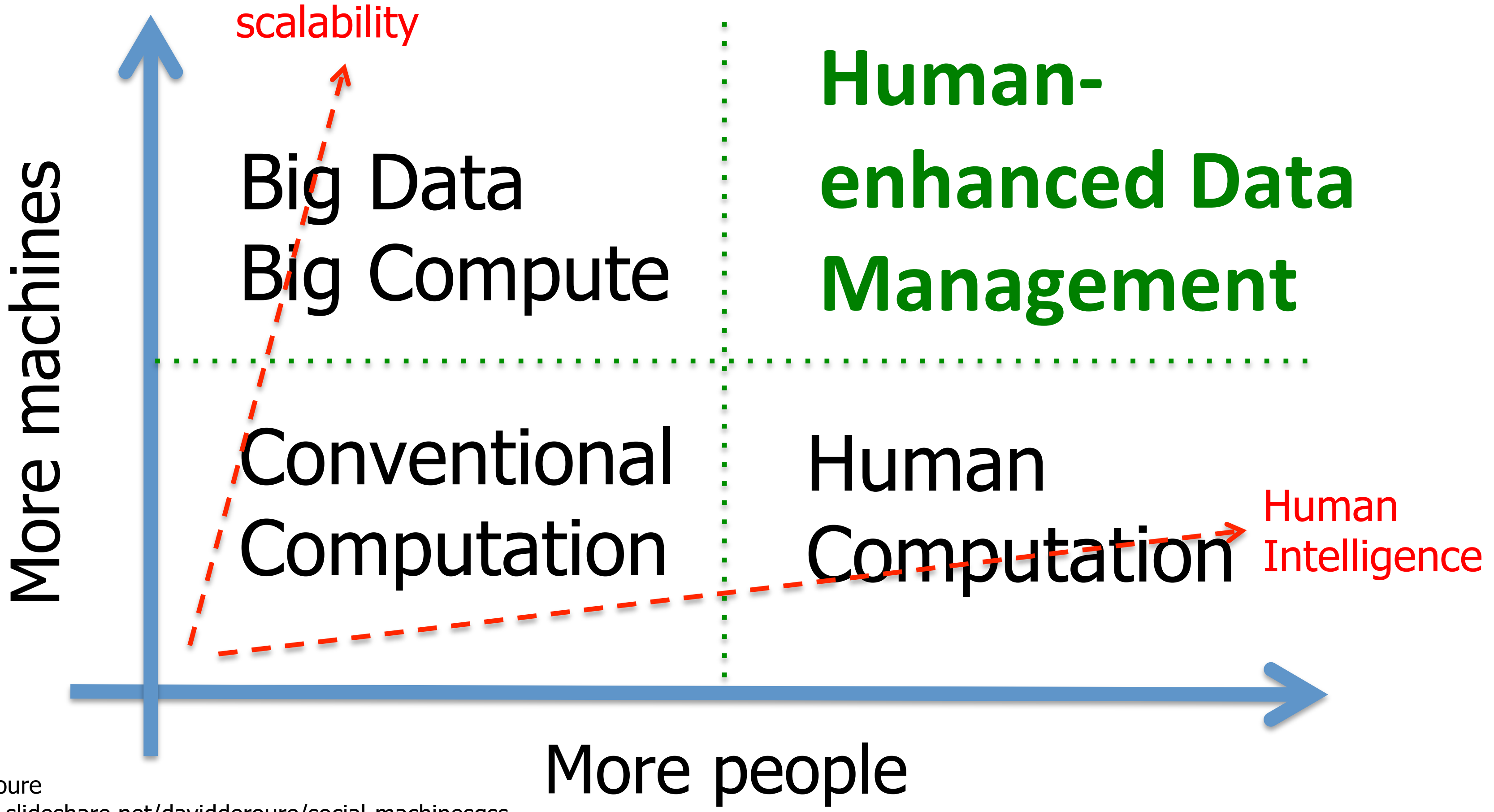
27 RETWEETS

2 FAVORITES

6:17 PM - 27 Feb 10

Chile Earthquake 2010

Big Data (Smart) Sense-making



Big Data
Big Compute

Human-enhanced Data Management

Conventional Computation

Human Computation

David De Roure
http://www.slideshare.net/davidderoure/social-machinesgss

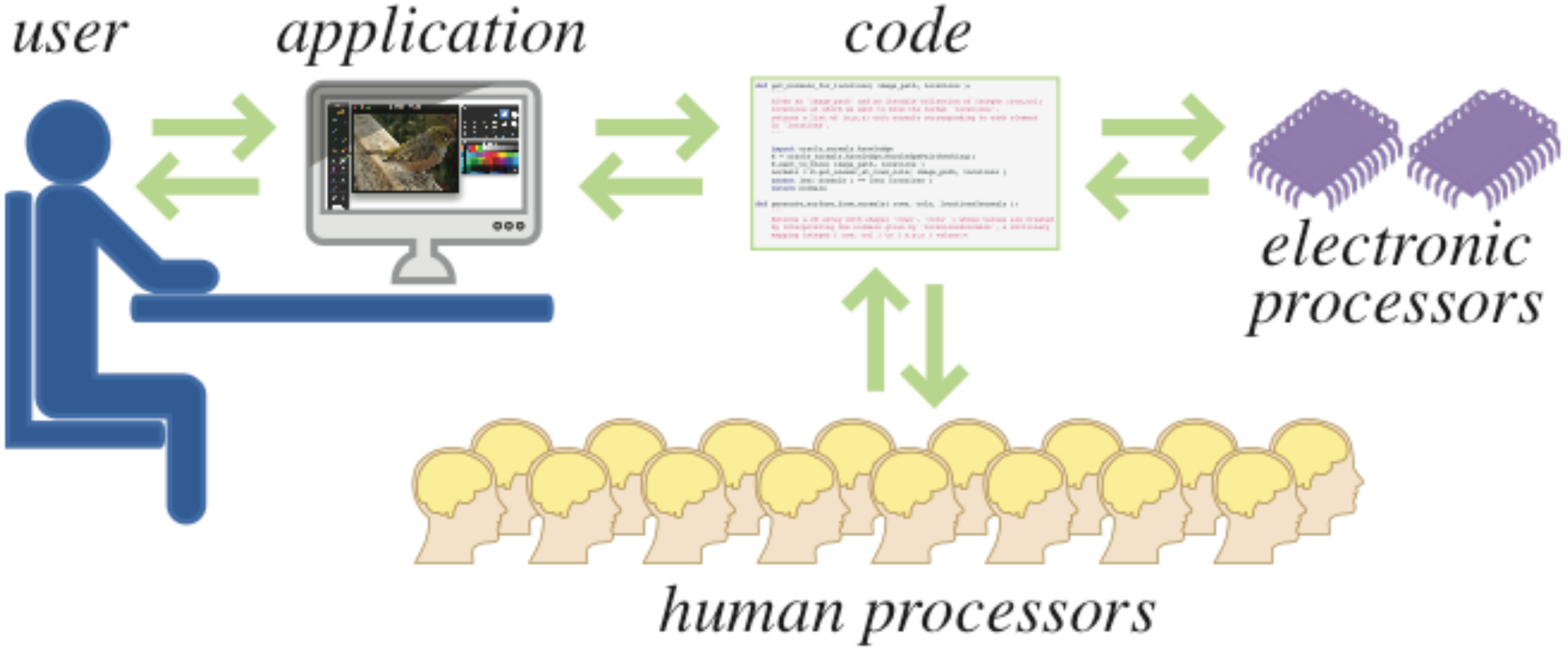
What do we need?

- A mature theory/practice of automatic and human data management
- A better understanding of the available workforce
- A better understanding of workers engagement mechanisms

A mature theory/practice of automatic and human data management

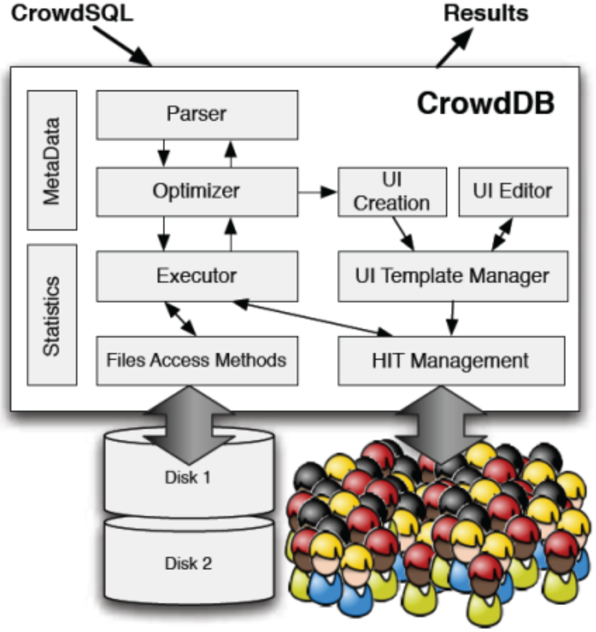
A theory of hybrid computation

The human co-processing units (HPU)



- Abstractions?
- Instruction set?
- Design tools?
- Debug?
- Control?

CrowdDB



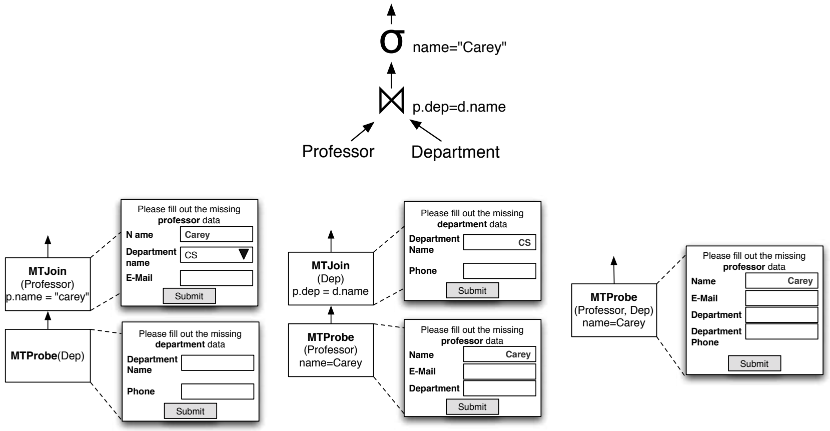
- CROWD columns
  - entities known, properties of entities may be unknown

```
CREATE TABLE company (
name STRING PRIMARY KEY,
hq_address CROWD STRING);
```

- CROWD table
  - entities unknown, crowd-source new entities

```
CREATE CROWD TABLE department (
    university STRING,
    department STRING,
    phone_no STRING)
    PRIMARY KEY(university, department) );
```

- GOAL: crowd-source comparisons, missing data
  - SQL with extensions to the DML and the Query Language


σ name="Carey"
⋈ p.dep=d.name
Professor
Department
MTJoin (Professor) p.name = "carey"
MTProbe(Dep)
MTJoin (Dep) p.dep = d.name
MTProbe (Professor) name=Carey
MTProbe (Professor, Dep) name=Carey

Please fill out the missing **professor** data
N ame: Carey
Department name: CS
E-Mail:
Submit

Please fill out the missing **department** data
Department Name:
Phone:
Submit

(Department first)

Please fill out the missing **department** data
Department Name: CS
Phone:
Submit

Please fill out the missing **professor** data
Name: Carey
E-Mail:
Department:
Submit

(Professor first)

≈10% Error-Rate

Please fill out the missing **professor** data
Name: Carey
E-Mail:
Department:
Department Phone:
Submit

(De-normalized Probe)

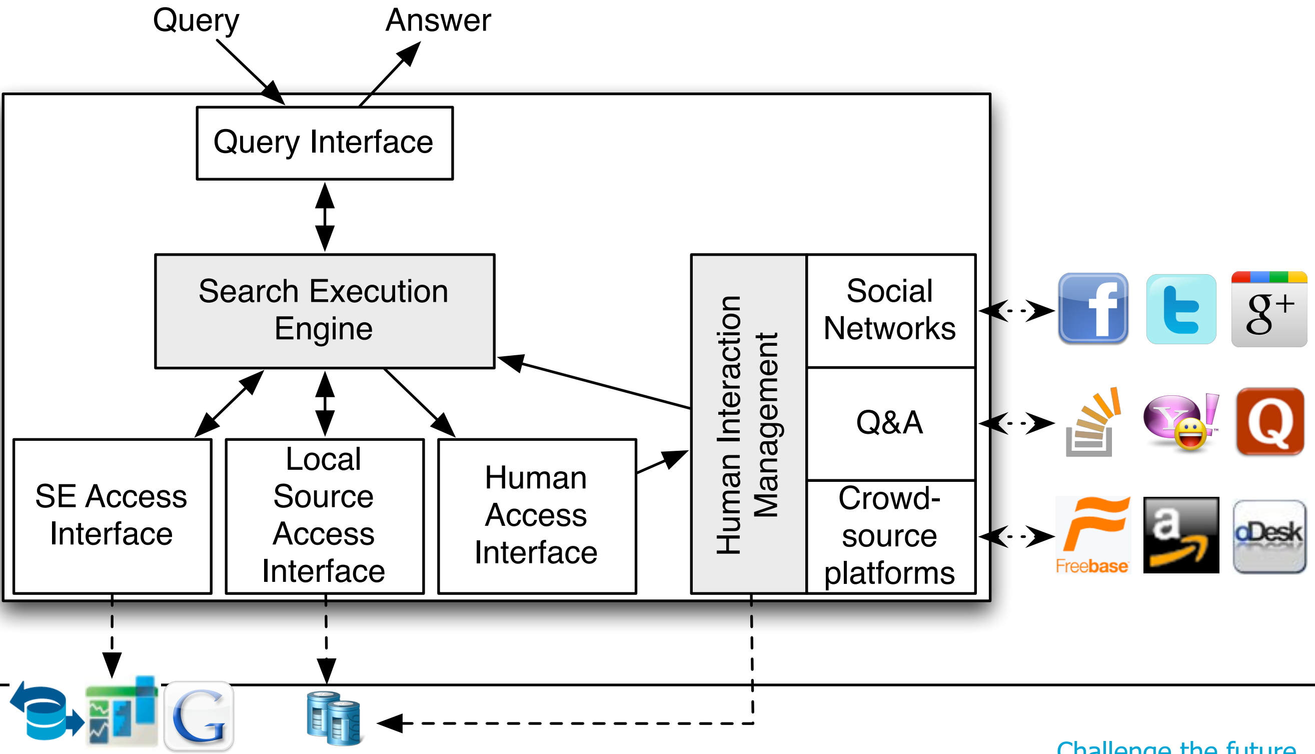
≈80% Error-Rate

Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: answering queries with crowdsourcing. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11). ACM, New York, NY, USA, 61-72.

CrowdSearcher

[Bozzon2012,WWW][Bozzon2013,WWW]

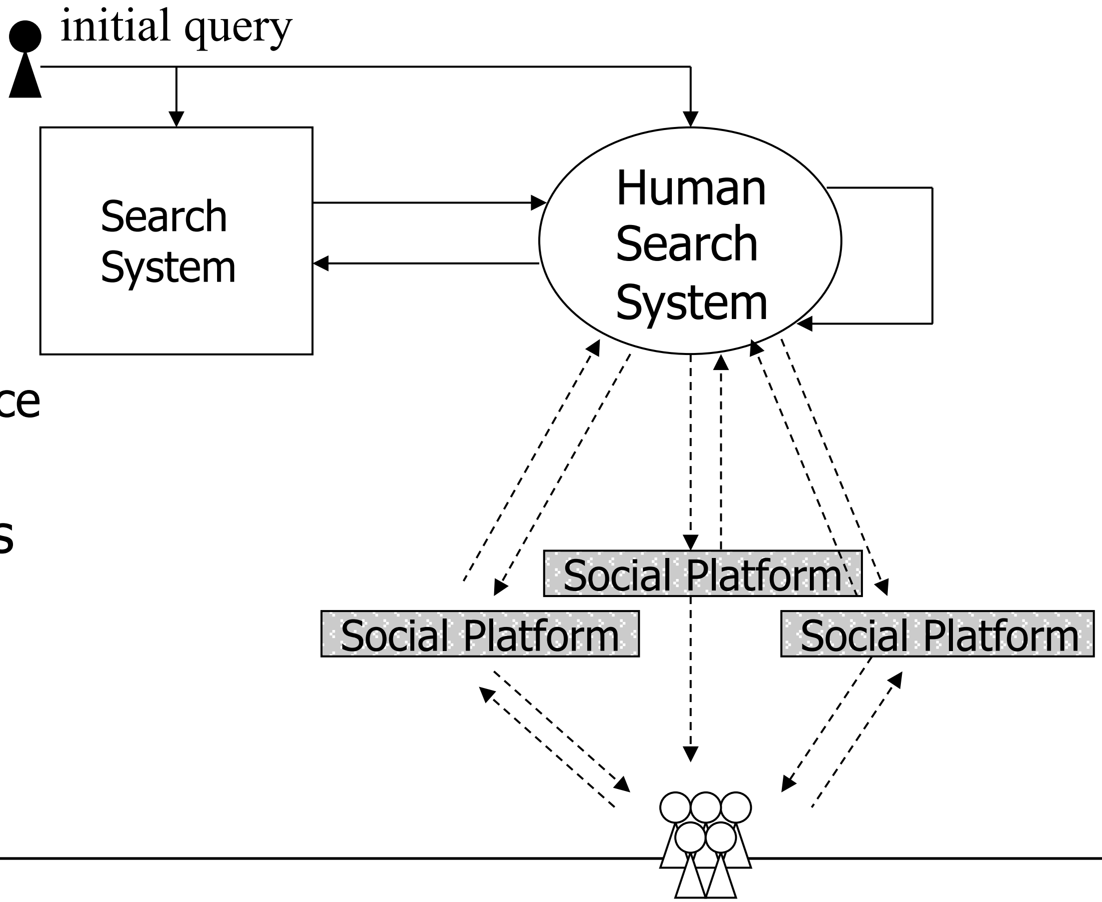
- Multi-platform, reactive, social-network-enabled crowdsourcing
- Our approach: a coordination engine which keeps an overall control on the application deployment and execution



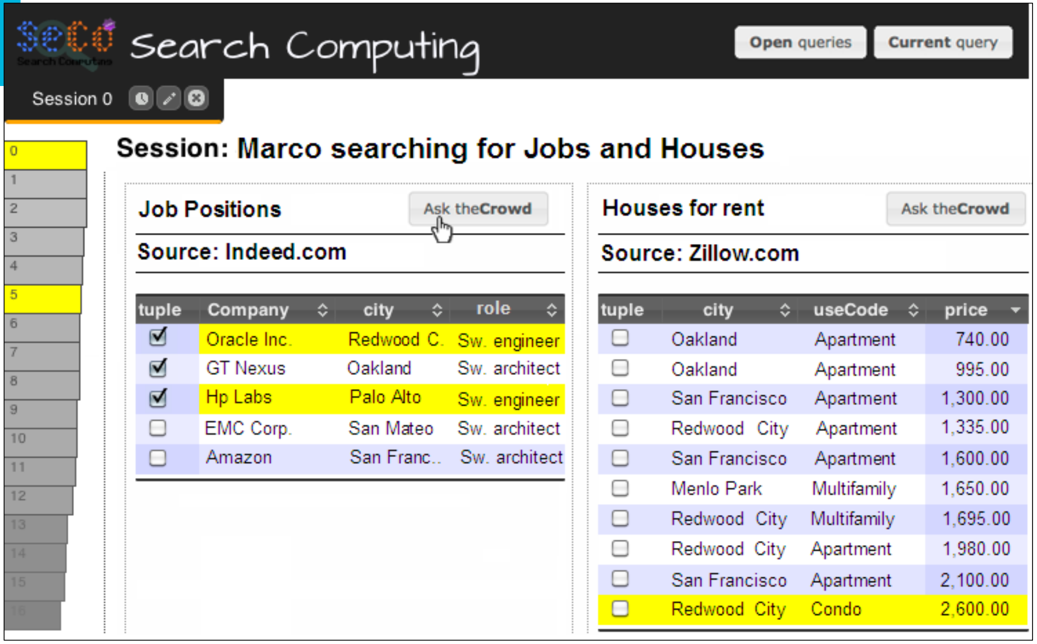
An example of crowd-based application: crowd-search

- **People do not trust web search completely**
  - Want to get direct feedback from people
  - Expect recommendations, insights, opinions, reassurance

- From search results to friends and experts feedback



Example: Find your next job (exploration)



Session: Marco searching for Jobs and Houses

Job Positions

Source: Indeed.com

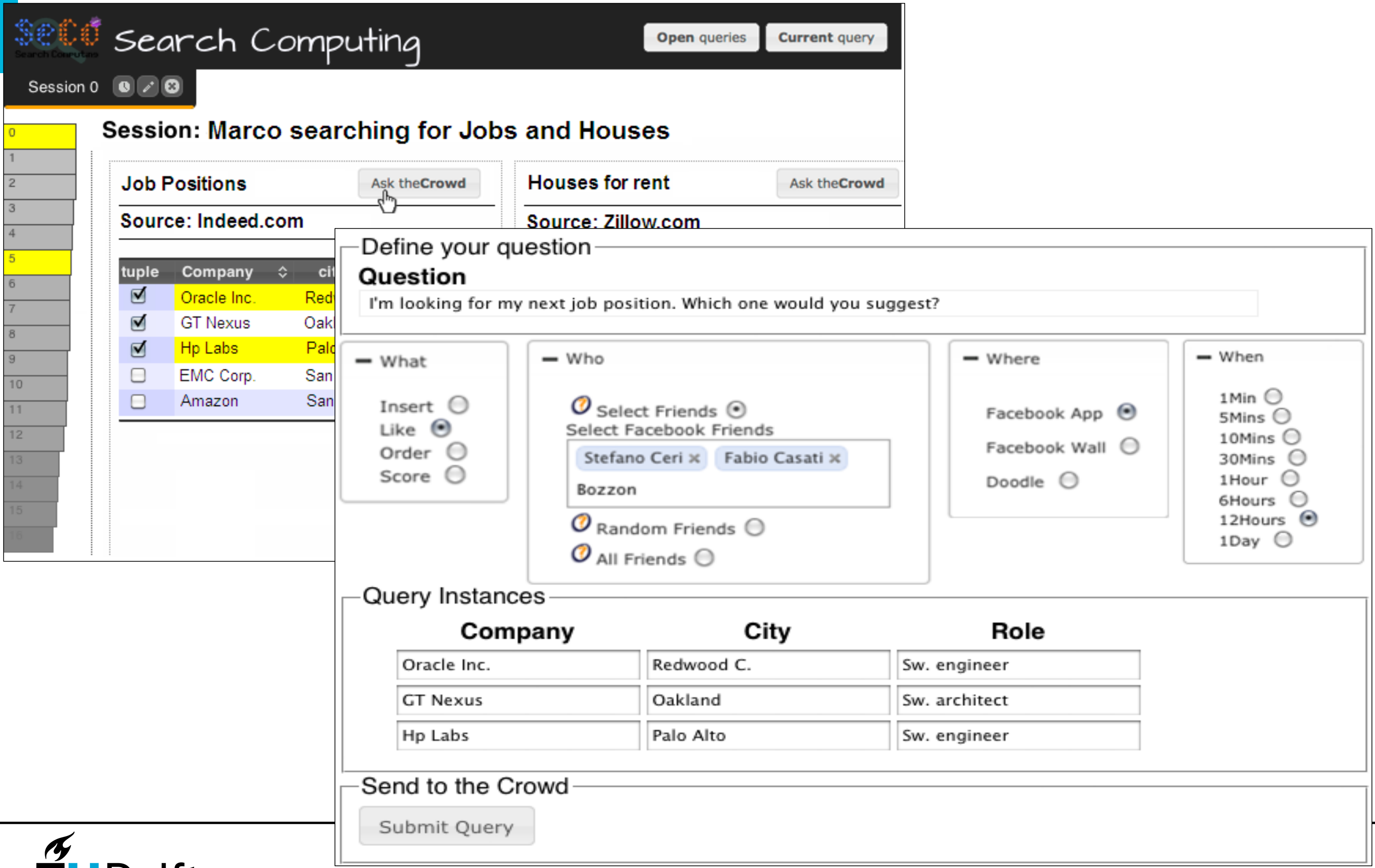
tuple	Company	city	role
☑	Oracle Inc.	Redwood C.	Sw. engineer
☑	GT Nexus	Oakland	Sw. architect
☑	Hp Labs	Palo Alto	Sw. engineer
□	EMC Corp.	San Mateo	Sw. architect
□	Amazon	San Franc..	Sw. architect

Houses for rent

Source: Zillow.com

tuple	city	useCode	price
□	Oakland	Apartment	740.00
□	Oakland	Apartment	995.00
□	San Francisco	Apartment	1,300.00
□	Redwood City	Apartment	1,335.00
□	San Francisco	Apartment	1,600.00
□	Menlo Park	Multifamily	1,650.00
□	Redwood City	Multifamily	1,695.00
□	Redwood City	Apartment	1,980.00
□	San Francisco	Apartment	2,100.00
□	Redwood City	Condo	2,600.00

Example: Find your job (social invitation)



Session: Marco searching for Jobs and Houses

Job Positions

Source: Indeed.com

Houses for rent

Source: Zillow.com

tuple	Company	ci
☑	Oracle Inc.	Red
☑	GT Nexus	Oak
☑	Hp Labs	Palo
□	EMC Corp.	San
□	Amazon	San

Define your question

Question

I'm looking for my next job position. Which one would you suggest?

What

- Insert ○
- Like ●
- Order ○
- Score ○

Who

- Select Friends ●
  Select Facebook Friends: Stefano Ceri × Fabio Casati × Bozzon
- Random Friends ○
- All Friends ○

Where

- Facebook App ●
- Facebook Wall ○
- Doodle ○

When

- 1Min ○
- 5Mins ○
- 10Mins ○
- 30Mins ○
- 1Hour ○
- 6Hours ○
- 12Hours ●
- 1Day ○

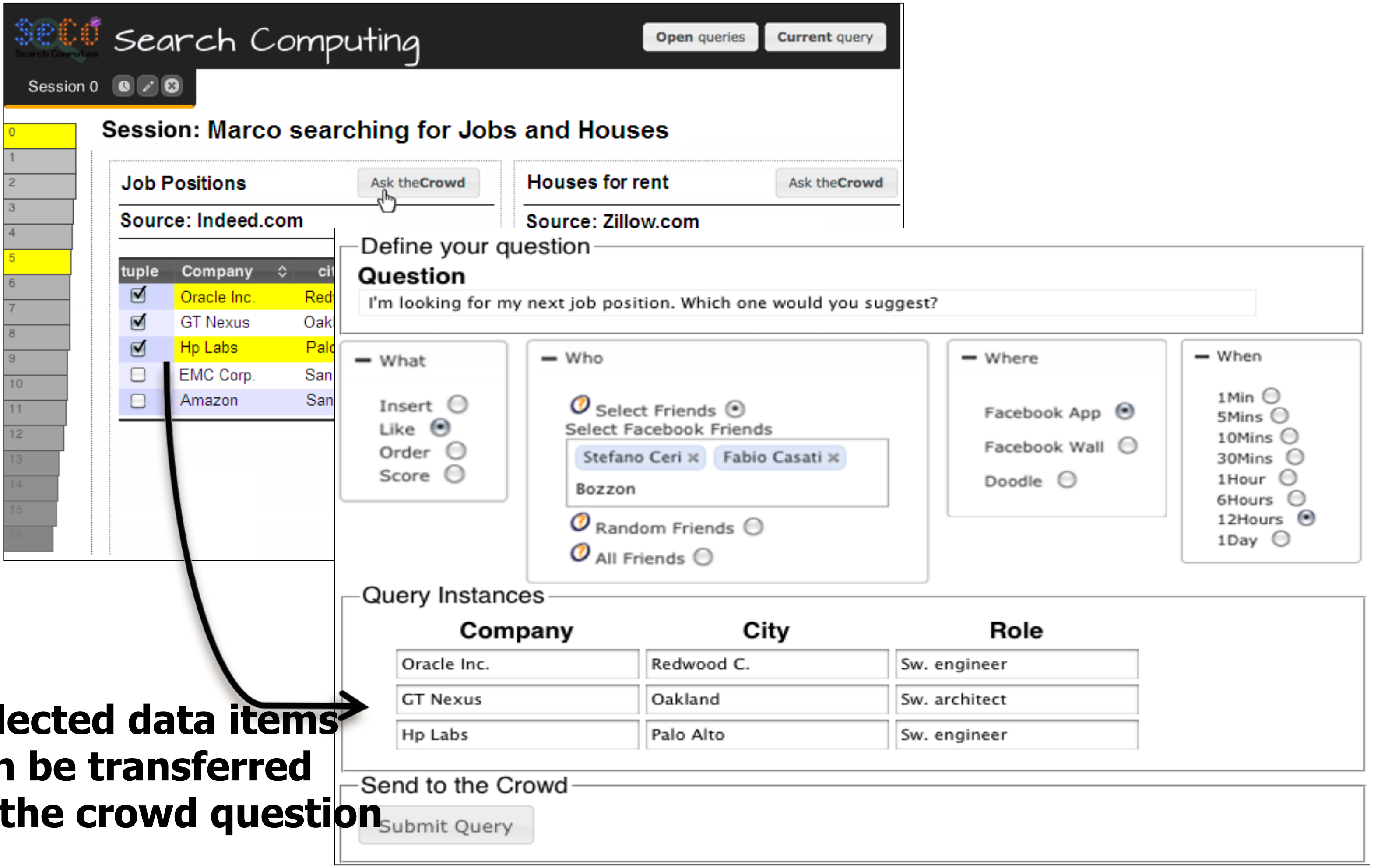
Query Instances

Company	City	Role
Oracle Inc.	Redwood C.	Sw. engineer
GT Nexus	Oakland	Sw. architect
Hp Labs	Palo Alto	Sw. engineer

Send to the Crowd

Submit Query

Example: Find your job (social invitation)



Session: Marco searching for Jobs and Houses

Job Positions

Source: Indeed.com

tuple	Company	cit
☑	Oracle Inc.	Red
☑	GT Nexus	Oak
☑	Hp Labs	Palo
□	EMC Corp.	San
□	Amazon	San

Houses for rent

Source: Zillow.com

Define your question

Question

I'm looking for my next job position. Which one would you suggest?

What: Insert ○ Like ● Order ○ Score ○

Who: Select Friends ●
Select Facebook Friends: Stefano Ceri × Fabio Casati × Bozzon
Random Friends ○
All Friends ○

Where: Facebook App ● Facebook Wall ○ Doodle ○

When: 1Min ○ 5Mins ○ 10Mins ○ 30Mins ○ 1Hour ○ 6Hours ○ 12Hours ● 1Day ○

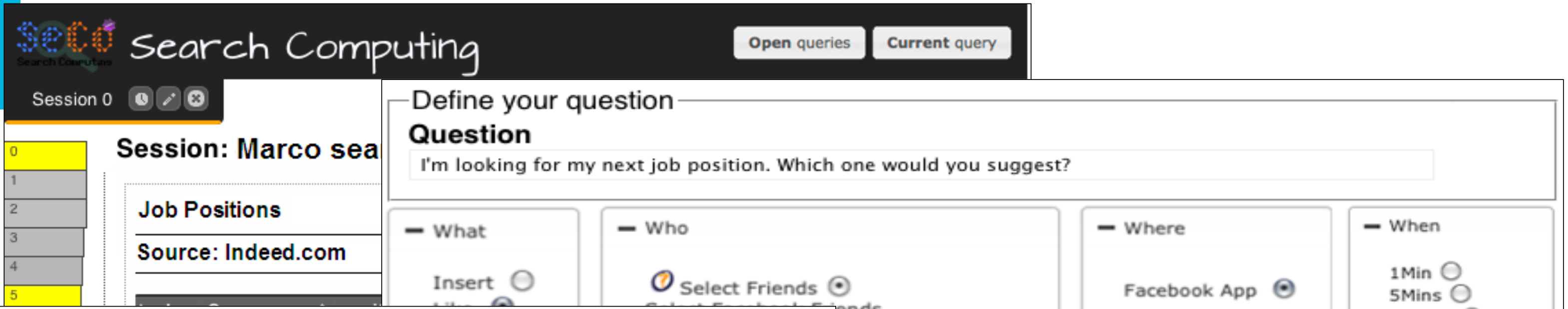
Query Instances

Company	City	Role
Oracle Inc.	Redwood C.	Sw. engineer
GT Nexus	Oakland	Sw. architect
Hp Labs	Palo Alto	Sw. engineer

Send to the Crowd

Selected data items can be transferred to the crowd question

Find your job (response submission)



Search Computing

Session: Marco sea

Job Positions

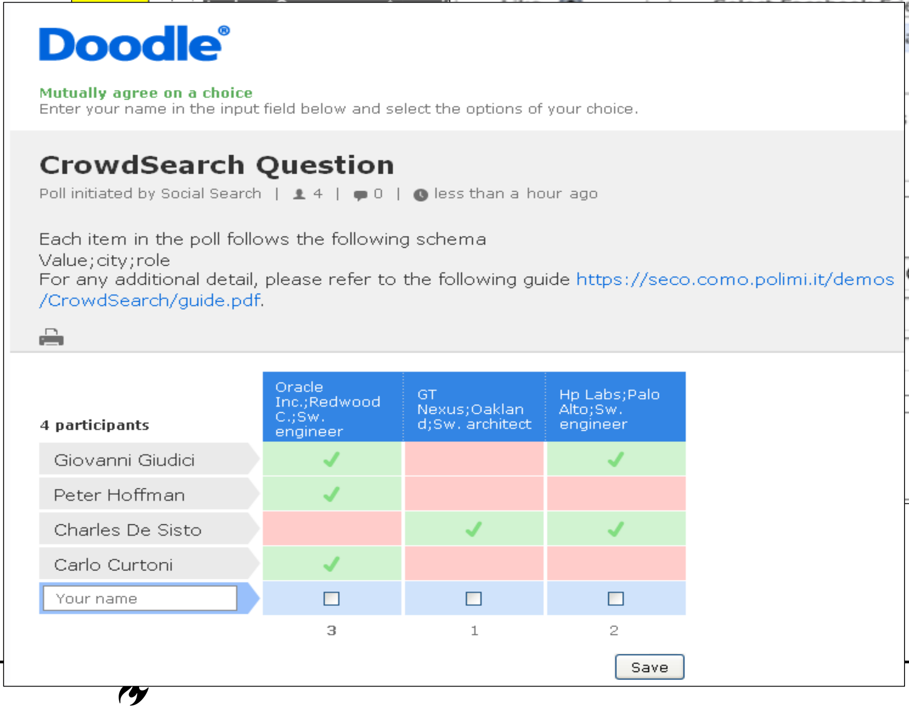
Source: Indeed.com

Define your question

Question

I'm looking for my next job position. Which one would you suggest?

What	Who	Where	When
Insert ○	Select Friends ●	Facebook App ●	1Min ○ 5Mins ○



Doodle®

Mutually agree on a choice

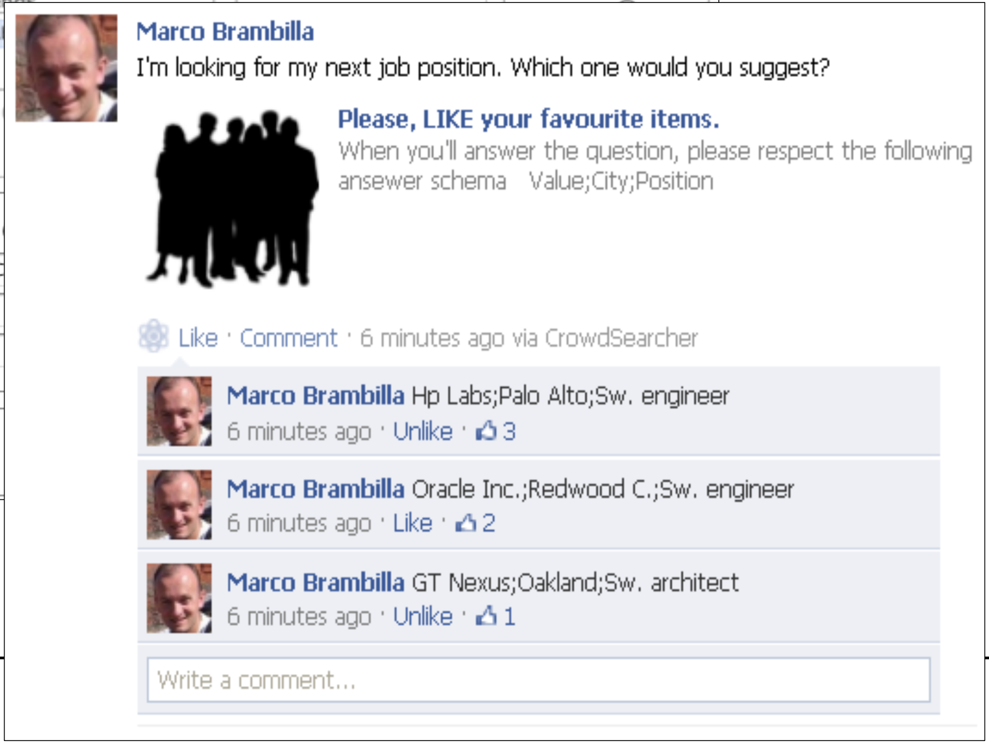
Enter your name in the input field below and select the options of your choice.

CrowdSearch Question

Poll initiated by Social Search | 4 | 0 | less than a hour ago

Each item in the poll follows the following schema
Value;city;role
For any additional detail, please refer to the following guide https://seco.como.polimi.it/demos/CrowdSearch/guide.pdf.

4 participants	Oracle Inc.;Redwood C.;Sw. engineer	GT Nexus;Oakland;Sw. architect	Hp Labs;Palo Alto;Sw. engineer
Giovanni Giudici	✓		✓
Peter Hoffman	✓		
Charles De Sisto		✓	✓
Carlo Curtoni	✓		
Your name	□	□	□
	3	1	2



Marco Brambilla

I'm looking for my next job position. Which one would you suggest?

Please, LIKE your favourite items.

When you'll answer the question, please respect the following answer schema Value;City;Position

Like · Comment · 6 minutes ago via CrowdSearcher

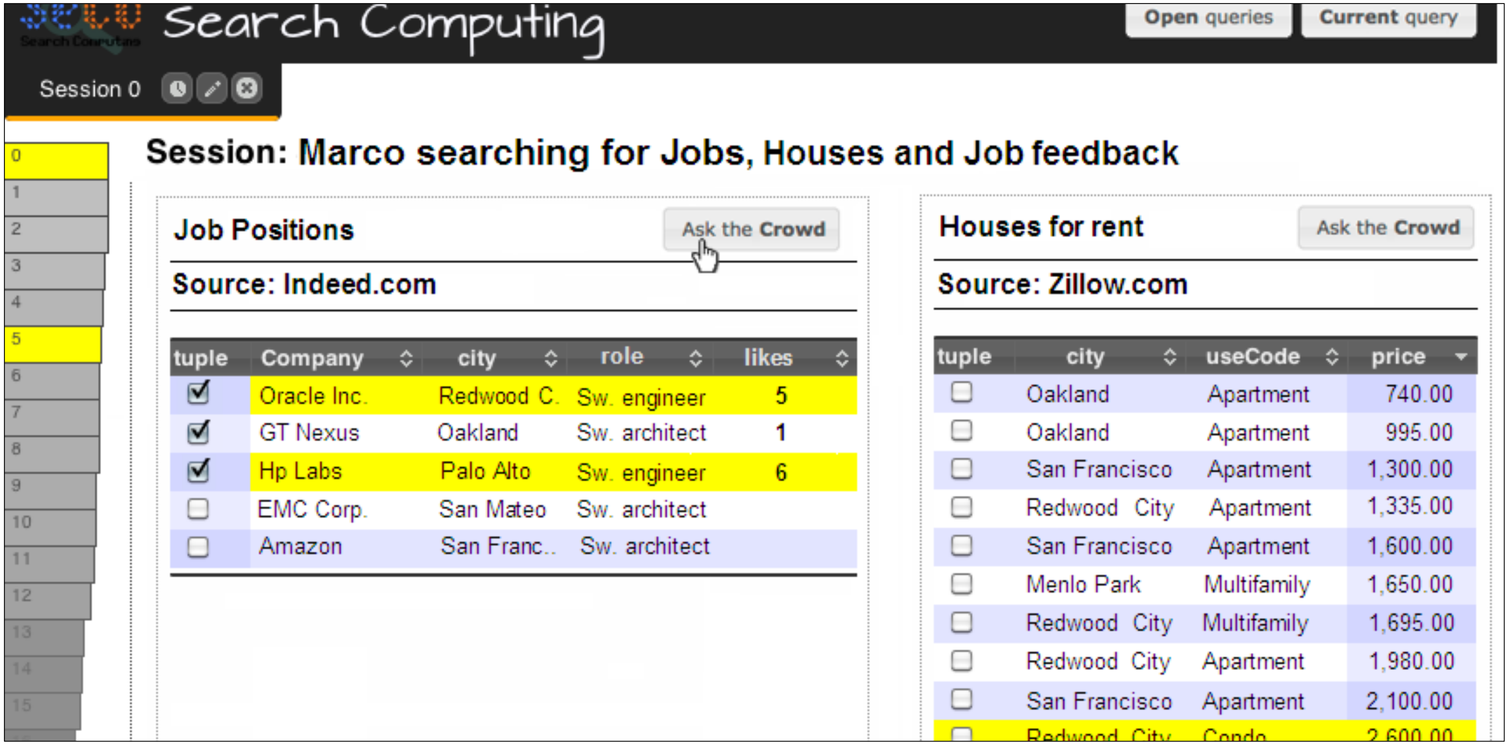
Marco Brambilla Hp Labs;Palo Alto;Sw. engineer
6 minutes ago · Unlike · 3

Marco Brambilla Oracle Inc.;Redwood C.;Sw. engineer
6 minutes ago · Like · 2

Marco Brambilla GT Nexus;Oakland;Sw. architect
6 minutes ago · Unlike · 1

Write a comment...

Crowdsearcher results (in the loop)



Session: Marco searching for Jobs, Houses and Job feedback

Job Positions

Source: Indeed.com

tuple	Company	city	role	likes
☑	Oracle Inc.	Redwood C.	Sw. engineer	5
☑	GT Nexus	Oakland	Sw. architect	1
☑	Hp Labs	Palo Alto	Sw. engineer	6
☐	EMC Corp.	San Mateo	Sw. architect	
☐	Amazon	San Franc..	Sw. architect	

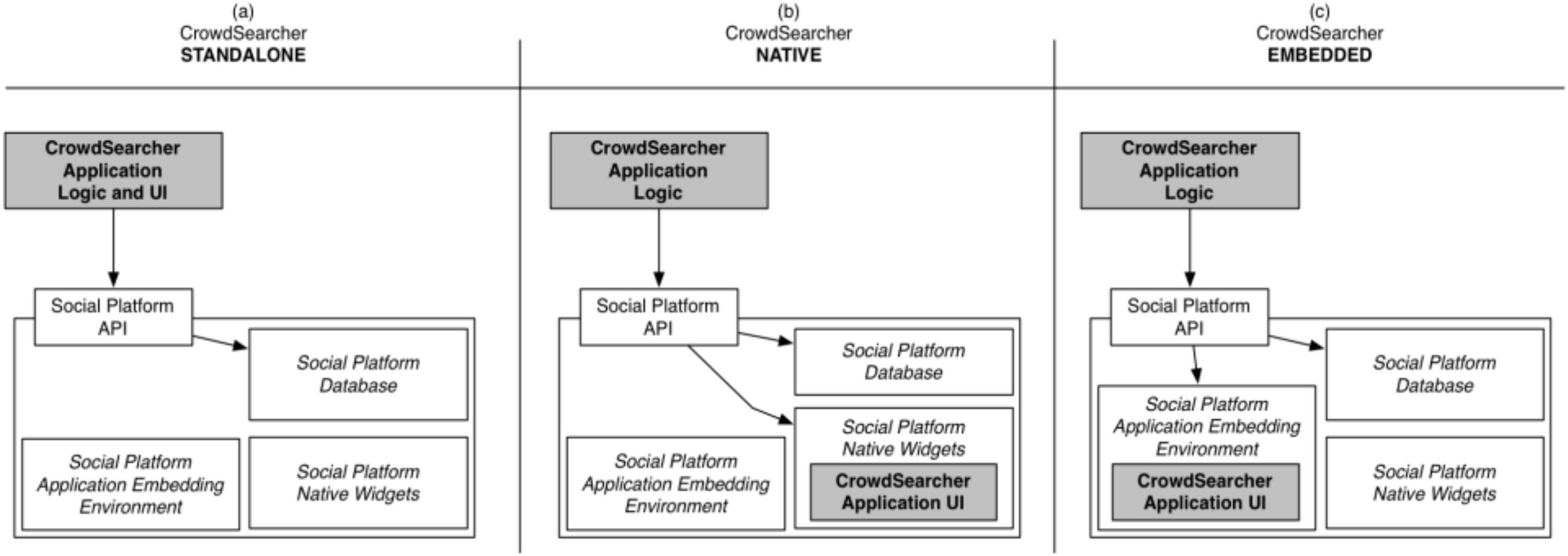
Houses for rent

Source: Zillow.com

tuple	city	useCode	price
☐	Oakland	Apartment	740.00
☐	Oakland	Apartment	995.00
☐	San Francisco	Apartment	1,300.00
☐	Redwood City	Apartment	1,335.00
☐	San Francisco	Apartment	1,600.00
☐	Menlo Park	Multifamily	1,650.00
☐	Redwood City	Multifamily	1,695.00
☐	Redwood City	Apartment	1,980.00
☐	San Francisco	Apartment	2,100.00
☐	Redwood City	Condo	2,600.00

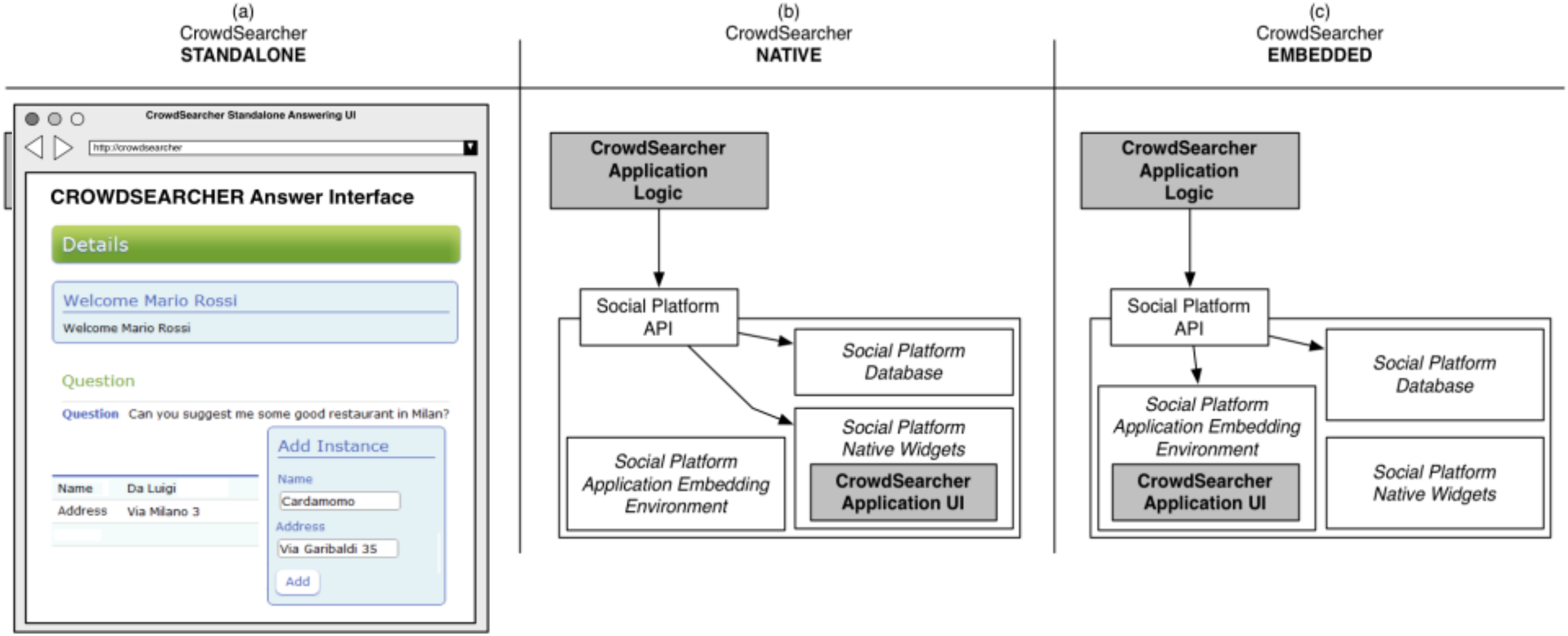
Deployment: search on the social network

- Multi-platform deployment


(a)
CrowdSearcher
STANDALONE
(b)
CrowdSearcher
NATIVE
(c)
CrowdSearcher
EMBEDDED
CrowdSearcher Application Logic and UI
Social Platform API
Social Platform Database
Social Platform Application Embedding Environment
Social Platform Native Widgets
CrowdSearcher Application Logic
Social Platform API
Social Platform Database
Social Platform Native Widgets
CrowdSearcher Application UI
Social Platform Application Embedding Environment
CrowdSearcher Application Logic
Social Platform API
Social Platform Application Embedding Environment
CrowdSearcher Application UI
Social Platform Database
Social Platform Native Widgets

Deployment: search on the social network

- Multi-platform deployment



(a)
CrowdSearcher
STANDALONE
(b)
CrowdSearcher
NATIVE
(c)
CrowdSearcher
EMBEDDED
CrowdSearcher Standalone Answering UI
http://crowdsearcher
CrowdSearcher Application Logic
Social Platform API
Social Platform Database
Social Platform Native Widgets
CrowdSearcher Application UI
Social Platform Application Embedding Environment
CrowdSearcher Application Logic
Social Platform API
Social Platform Application Embedding Environment
CrowdSearcher Application UI
Social Platform Database
Social Platform Native Widgets

CROWDSEARCHER Answer Interface

Details

Welcome Mario Rossi

Welcome Mario Rossi

Question

Question Can you suggest me some good restaurant in Milan?

Name	Da Luigi
Address	Via Milano 3

Add Instance

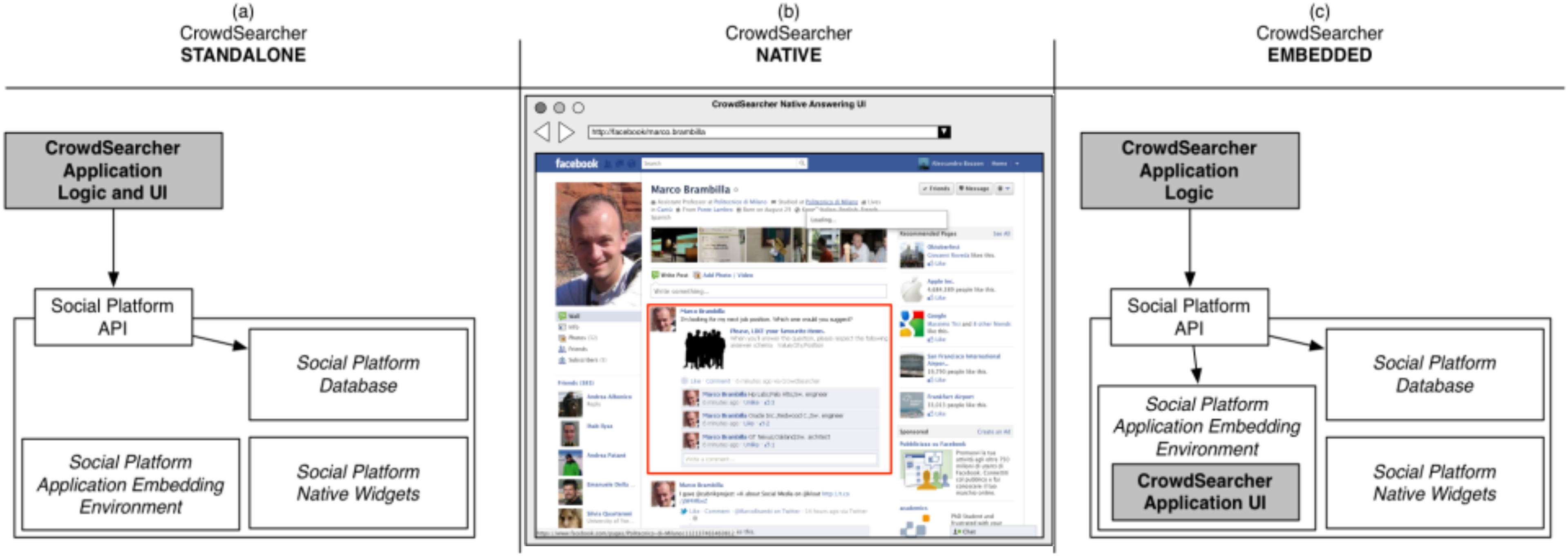
Name: Cardamomo

Address: Via Garibaldi 35

Add

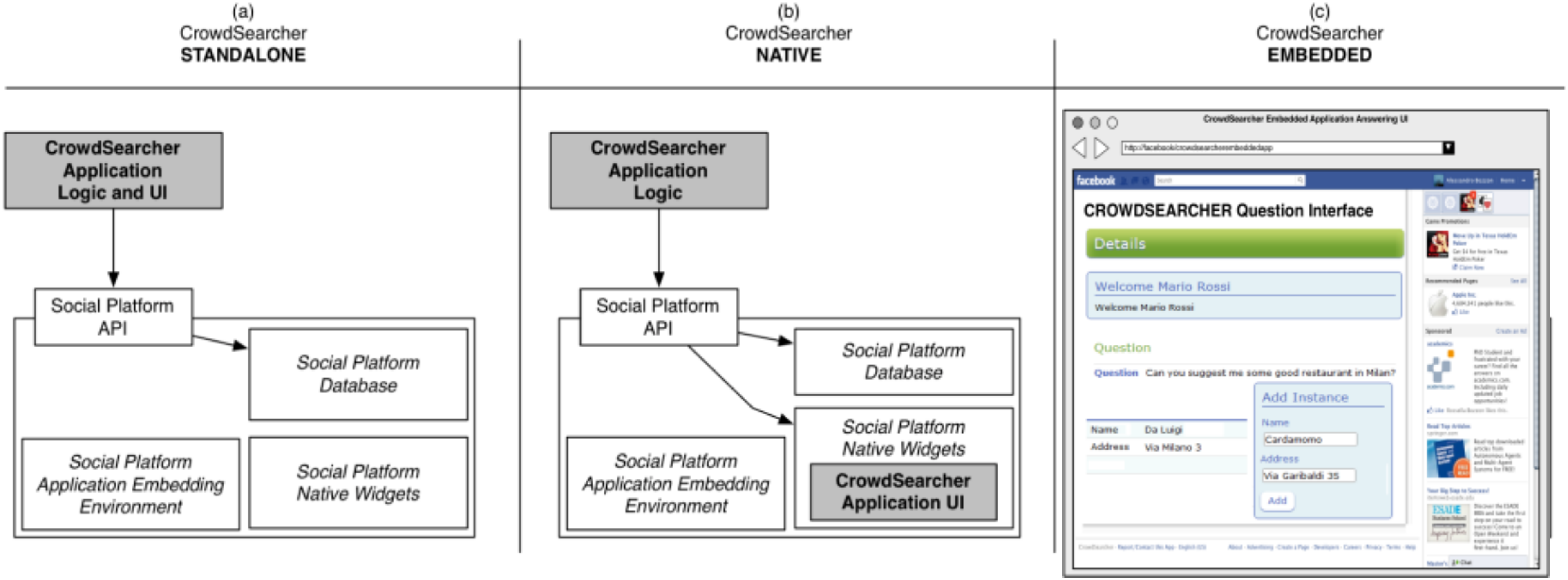
Deployment: search on the social network

- Multi-platform deployment


(a)
CrowdSearcher
STANDALONE
CrowdSearcher Application Logic and UI
Social Platform API
Social Platform Database
Social Platform Application Embedding Environment
Social Platform Native Widgets
(b)
CrowdSearcher
NATIVE
CrowdSearcher Native Answering UI
http://facebook/marco.brambilla
facebook
Marco Brambilla
(c)
CrowdSearcher
EMBEDDED
CrowdSearcher Application Logic
Social Platform API
Social Platform Application Embedding Environment
CrowdSearcher Application UI
Social Platform Database
Social Platform Native Widgets

Deployment: search on the social network

- Multi-platform deployment


(a)
CrowdSearcher
STANDALONE
(b)
CrowdSearcher
NATIVE
(c)
CrowdSearcher
EMBEDDED
CrowdSearcher Application Logic and UI
Social Platform API
Social Platform Database
Social Platform Application Embedding Environment
Social Platform Native Widgets
CrowdSearcher Application Logic
Social Platform API
Social Platform Database
Social Platform Native Widgets
CrowdSearcher Application UI
Social Platform Application Embedding Environment
CrowdSearcher Embedded Application Answering UI
http://facebook/crowdsearcherembeddedapp
facebook

CROWDSEARCHER Question Interface

Details

Welcome Mario Rossi

Welcome Mario Rossi

Question

Question Can you suggest me some good restaurant in Milan?

Name	Da Luigi
Address	Via Milano 3

Add Instance

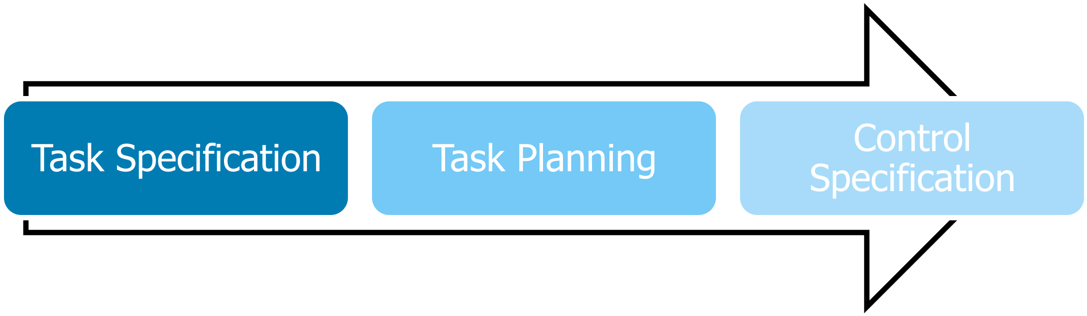
Name: Cardamomo

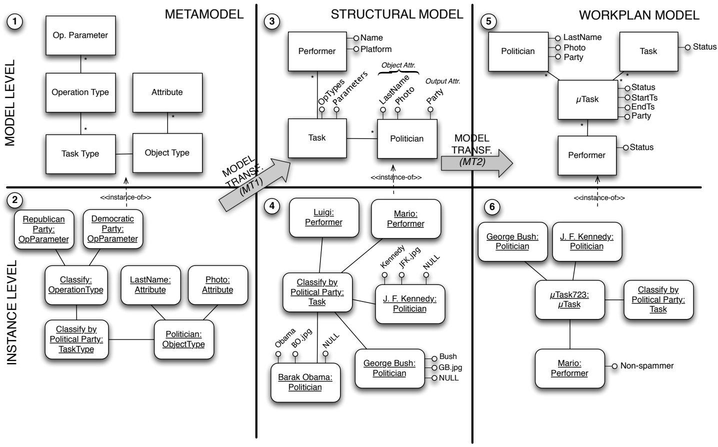
Address: Via Garibaldi 35

Add

Design Process

- A simple task design and deployment process, based on specific data structures
  - created using model-driven transformations
  - driven by the task specification





- **Task Specification:** task operations, objects, and performers
- **Task Planning:** work distribution
- **Control Specification**: task control policies

Design Dimensions

- Support for several types of task operations
  - Like, Comment, Tag, Classify, Add, Modify, Order, etc.

- Several strategies for
  - **Task splitting**: the input data collection is too complex relative to the cognitive capabilities of users.
  - **Task structuring**: the query is too complex or too critical to be executed in one shot.
  - **Task routing**: a query can be distributed according to the values of some attribute of the collection
  - Output aggregation

- Platform/community assignment
  - a task can be assigned to different communities or social platforms based on its focus

- Control => time, quality, money
  - A reactive execution environment for requirement enforcement and reactive execution

Reactive Crowdsourcing

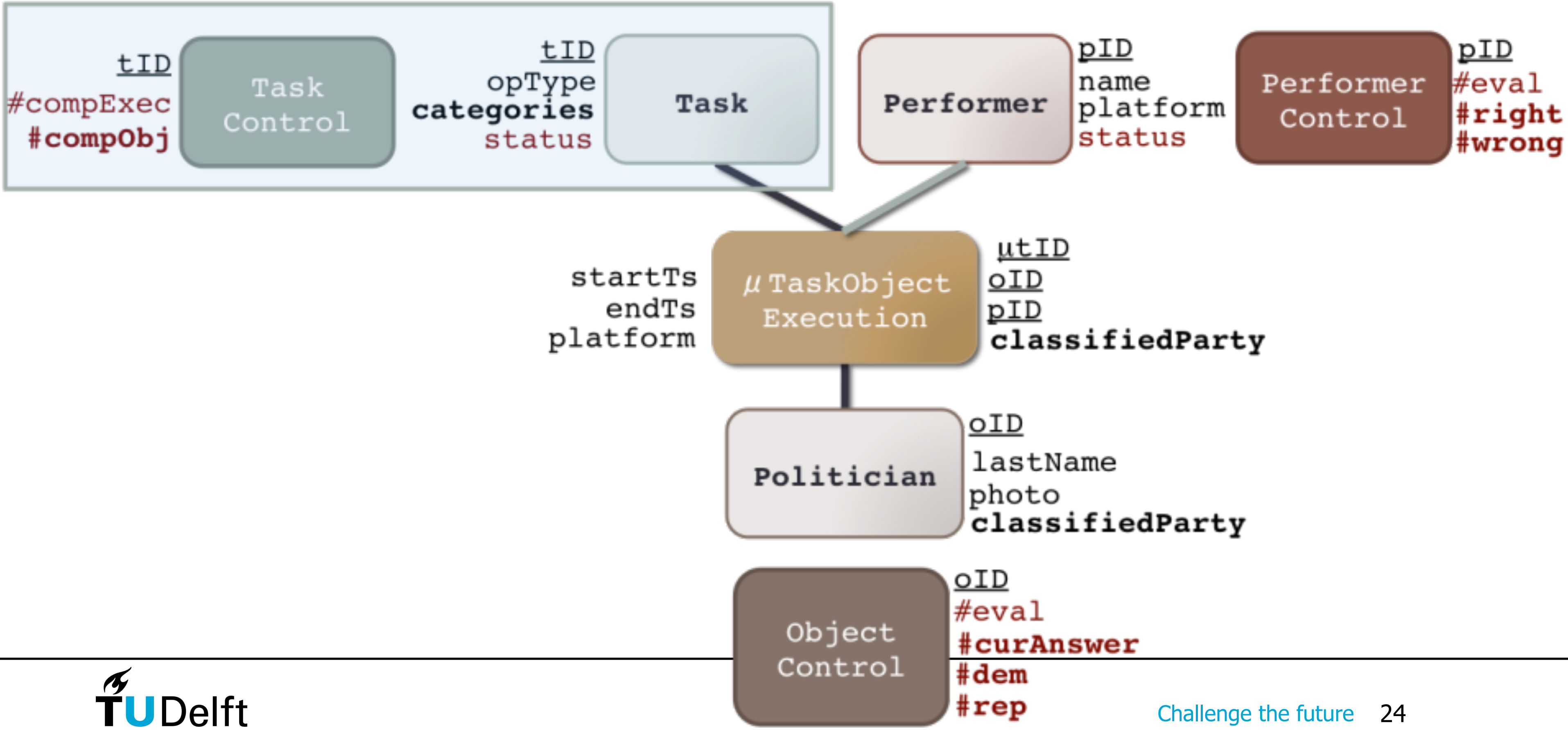
- A **conceptual framework** for controlling the execution of crowd-based computations. Based on:
    - Control Marts
    - Active Rules

- Classical forms of controls:
    - Majority control (to close object computations)
    - Quality control (to check that quality constraints are met)
    - Spam detection (to detect / eliminate some performers)
    - Multi-platform adaptation (to change the deployment platform)
    - Social adaptation (to change the community of performers)

- Why Active Rules?

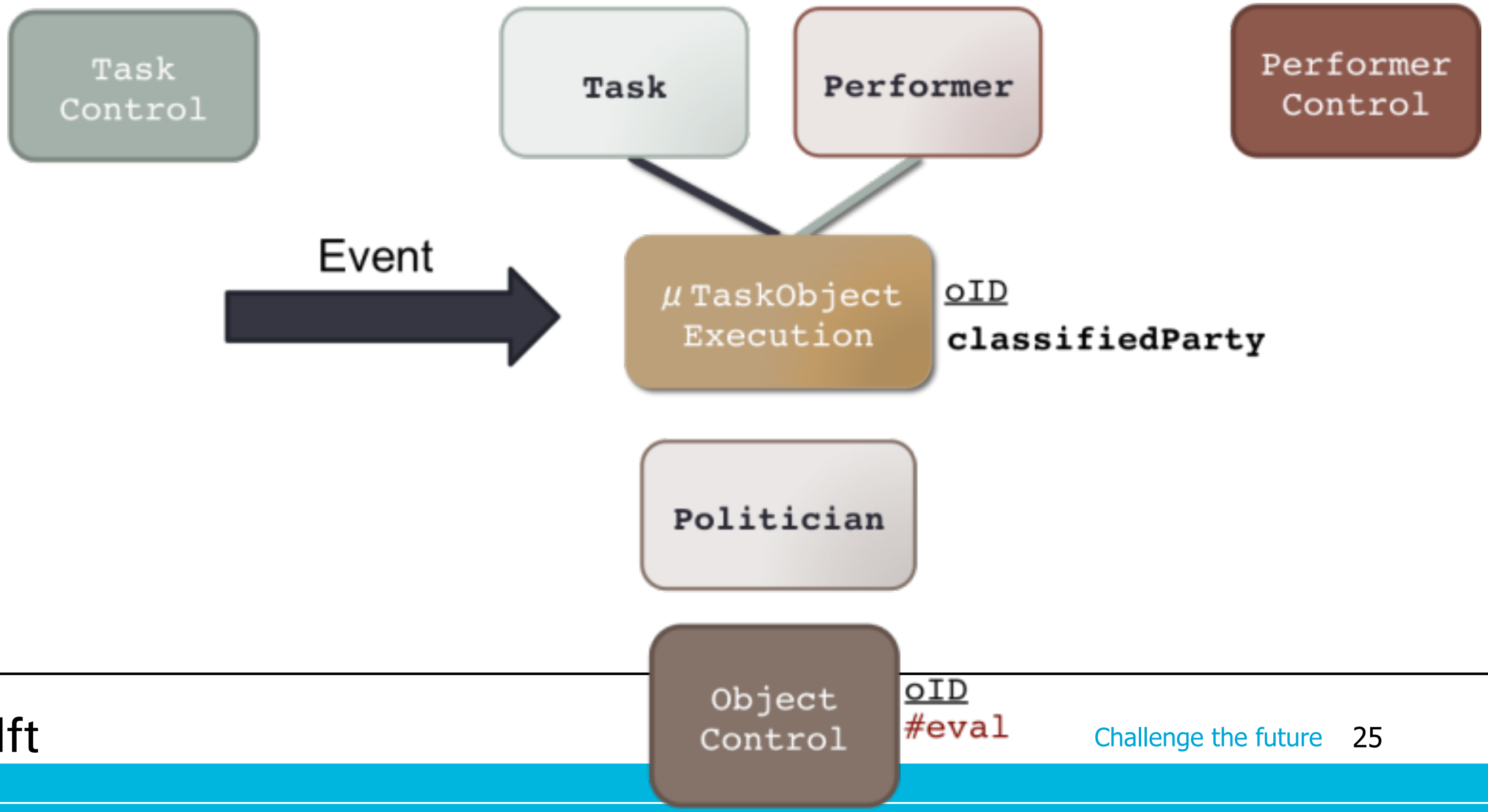
Auxiliary Structures

- **Object :** tracking object responses
- **Performer:** tracking performer behavior (e.g. spammers)
- **Task:** tracking task status



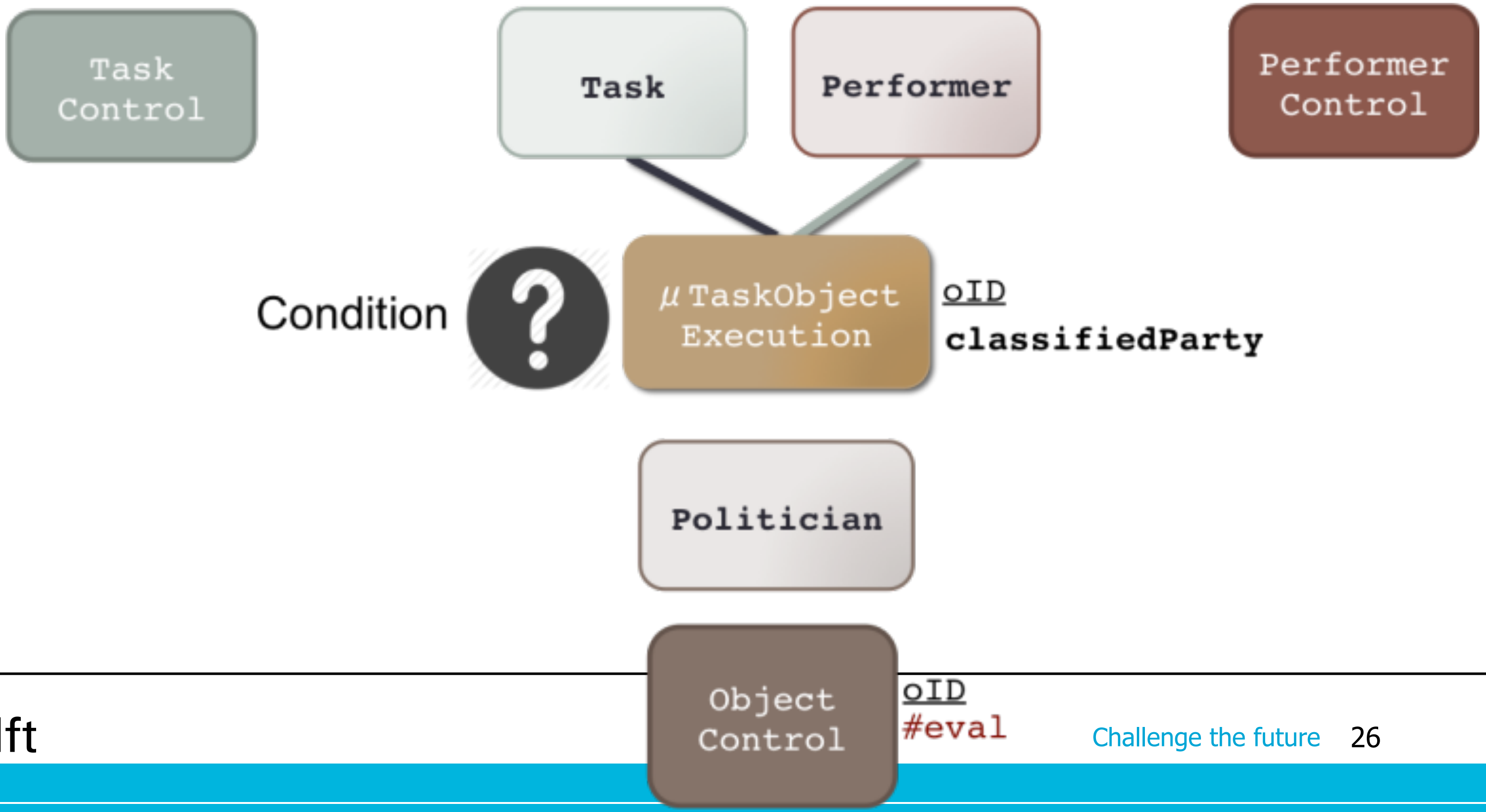
Rule Example

```
e: UPDATE FOR μTaskObjectExecution[ClassifiedParty]
c: NEW.ClassifiedParty == 'Republican'
a: SET ObjectControl[oID == NEW.oID].#Eval+= 1
```



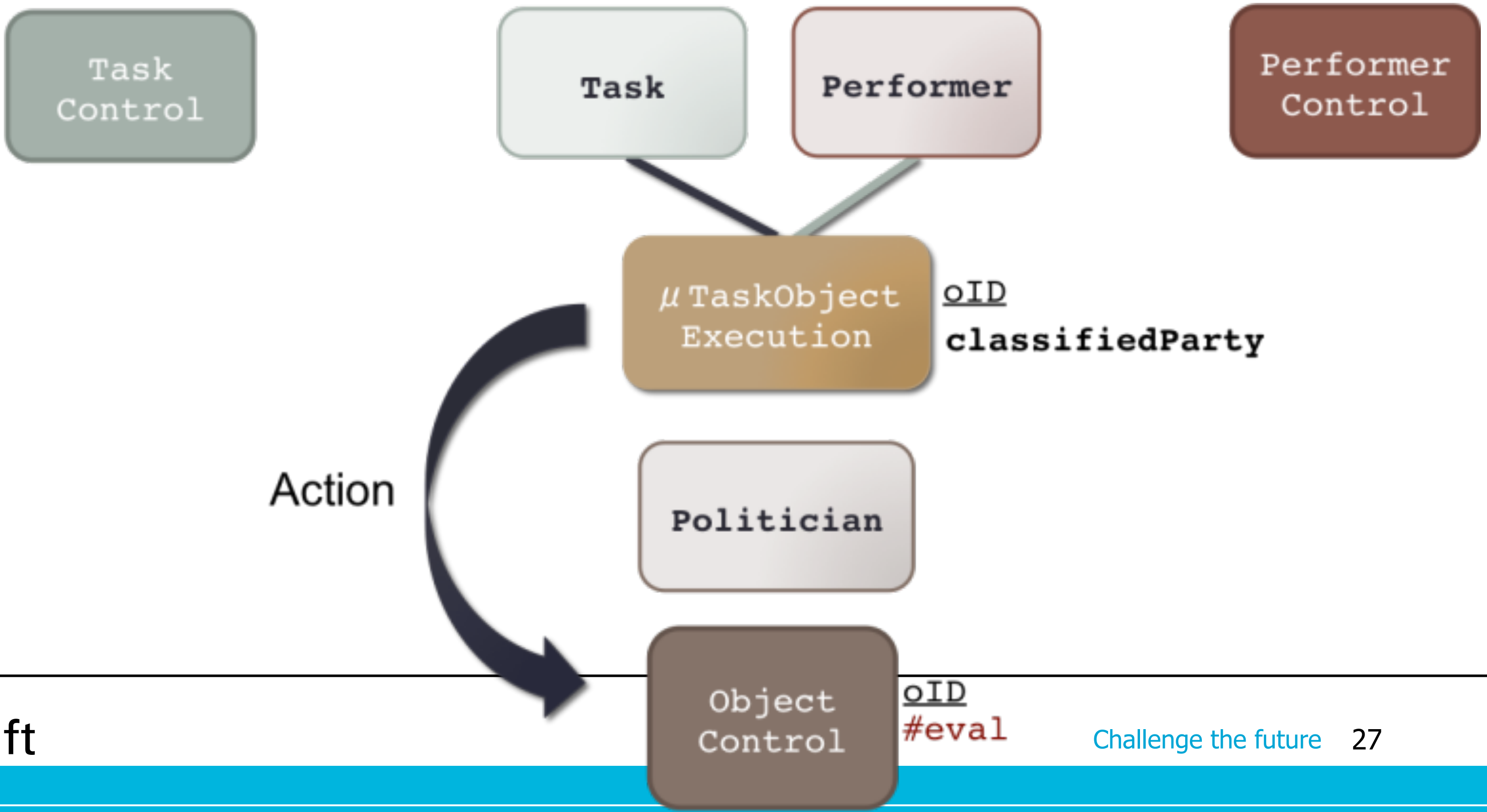
Rule Example

```
e: UPDATE FOR μTaskObjectExecution[ClassifiedParty]
c: NEW.ClassifiedParty == 'Republican'
a: SET ObjectControl[oID == NEW.oID].#Eval+= 1
```

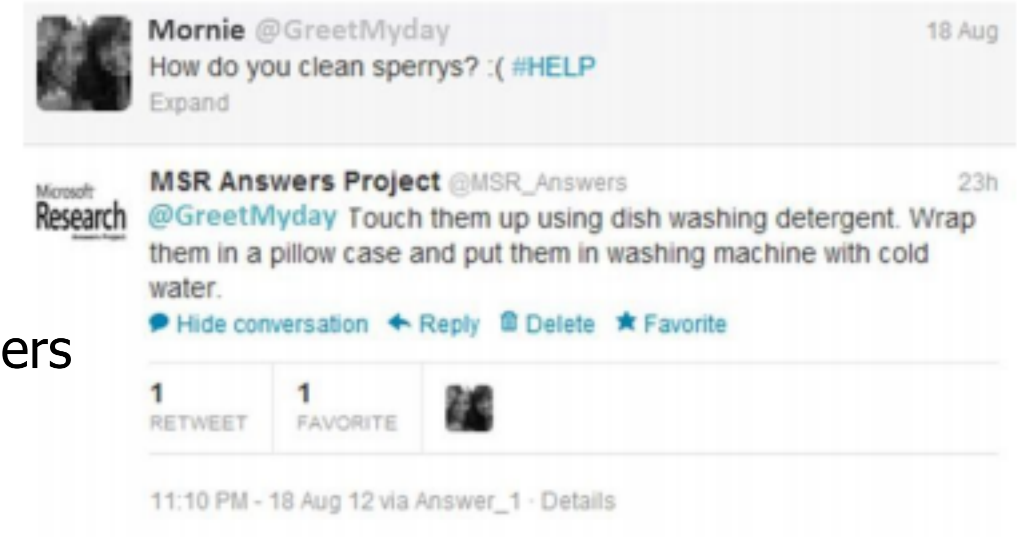


Rule Example

```
e: UPDATE FOR μTaskObjectExecution[ClassifiedParty]
c: NEW.ClassifiedParty == ’Republican’
a: SET ObjectControl[oID == NEW.oID].#Eval+= 1
```



CROWD-POWERED SEARCH



Mornie @GreetMyday 18 Aug
How do you clean sperrys? :(#HELP
Expand

MSR Answers Project @MSR_Answers 23h
@GreetMyday Touch them up using dish washing detergent. Wrap them in a pillow case and put them in washing machine with cold water.

11:10 PM - 18 Aug 12 via Answer_1 · Details

- Users ask questions on Twitter
- An hybrid system provide answers
- Workers used for
  - label tweets as "rhetorical question" or not
    - Median 3.02 mins
  - produce responses to question
    - Median 77.4 mins
  - Voting responses
    - Median 82.1 mis

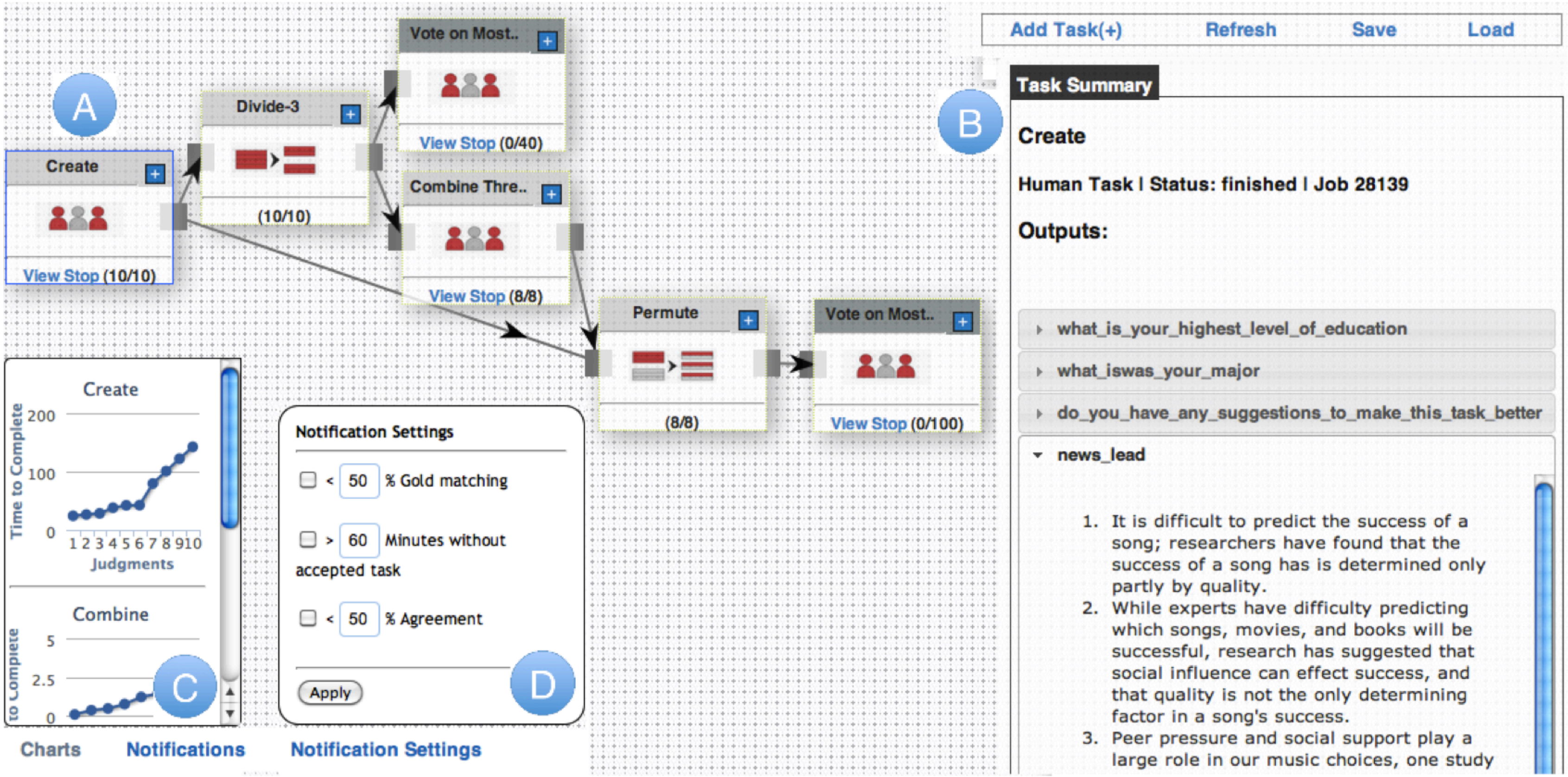
Median time =>162.5 minutes

Cost => $0.95 per tweet

A Crowd-Powered Socially Embedded Search Engine. Jin-Woo Jeong, Meredith Ringel Morris, Jaime Teevan, Daniel Liebling. ICWSM 2013

Crowdweaver

- [Kittur et al. 2012]



Human Task | Status: finished | Job 28139

Outputs:

▸ what_is_your_highest_level_of_education

▸ what_iswas_your_major

▸ do_you_have_any_suggestions_to_make_this_task_better

▾ news_lead

1. It is difficult to predict the success of a song; researchers have found that the success of a song has is determined only partly by quality.
2. While experts have difficulty predicting which songs, movies, and books will be successful, research has suggested that social influence can effect success, and that quality is not the only determining factor in a song's success.
3. Peer pressure and social support play a large role in our music choices, one study

Notification Settings

□ < 50 % Gold matching

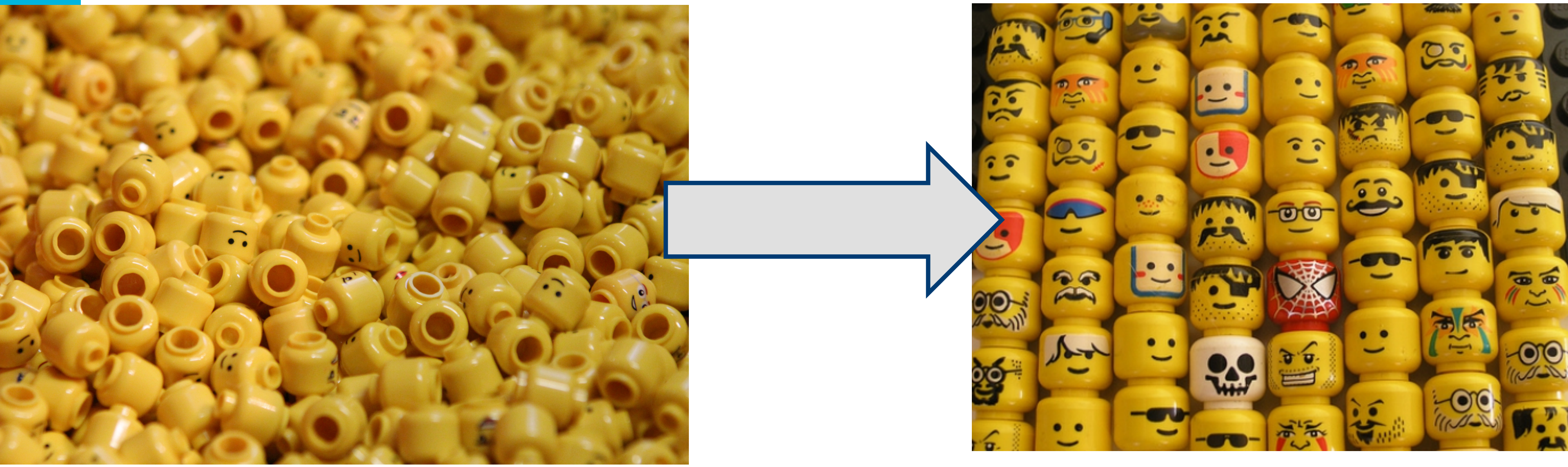
□ > 60 Minutes without accepted task

□ < 50 % Agreement

Apply

Better Understanding of the Available Workforce

Worker Modeling



- Implicit/Explicit Knowledge
- Interests
- Hard/Soft Skills

- Attitude
- Task-type/domain specific performance
- Availability

- Capacity
- Teams/Groups
- Trust

Problem

- Ranking the members of a social group according to the level of knowledge that they have about a given topic
- Application: crowd selection (for Crowd Searching or Sourcing)

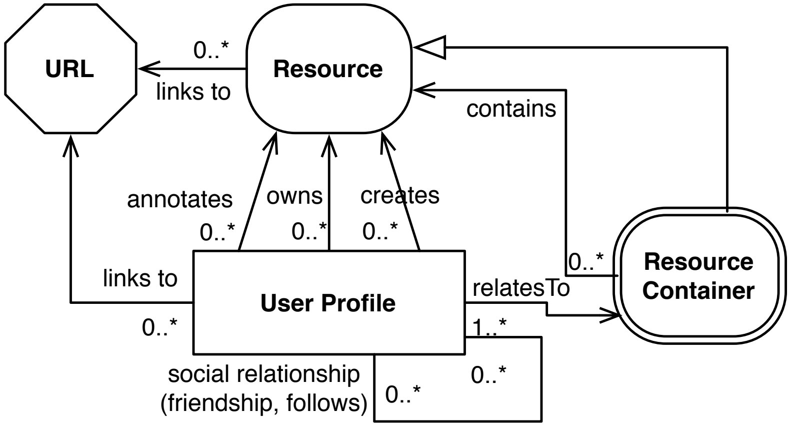
- Available data
  - User profile
  - behavioral trace that users leave behind them through their social activities

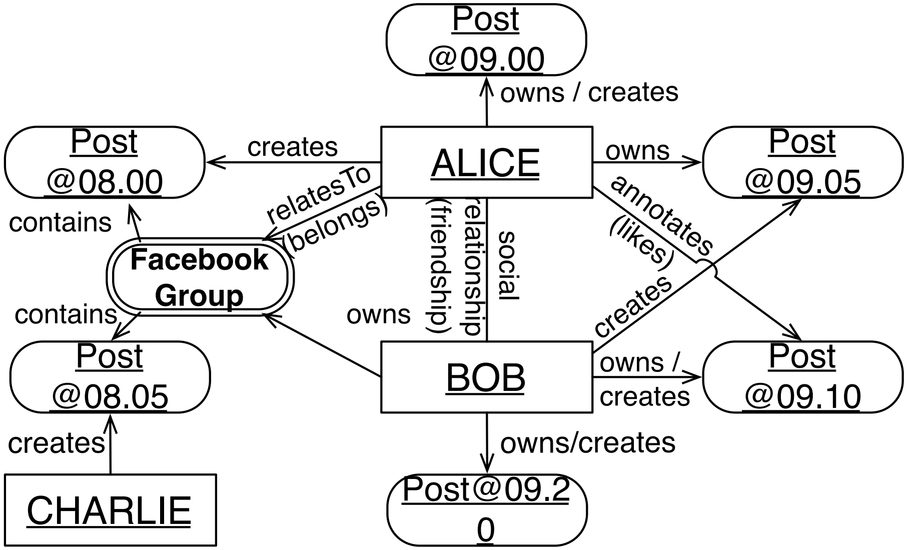
Finding the right crowd

[Bozzon2013,EDBT]

- Ranking the members of a social group according to the level of knowledge that they have about a given topic
    - Application: crowd selection (for Crowd Searching or Sourcing)

- Available data
    - User profile
    - behavioral trace that users leave behind them through their social activities





Main Results

- **Profiles** are **less effective** than level-1 resources
  - Resources **produced by others help** in describing each individual's expertise

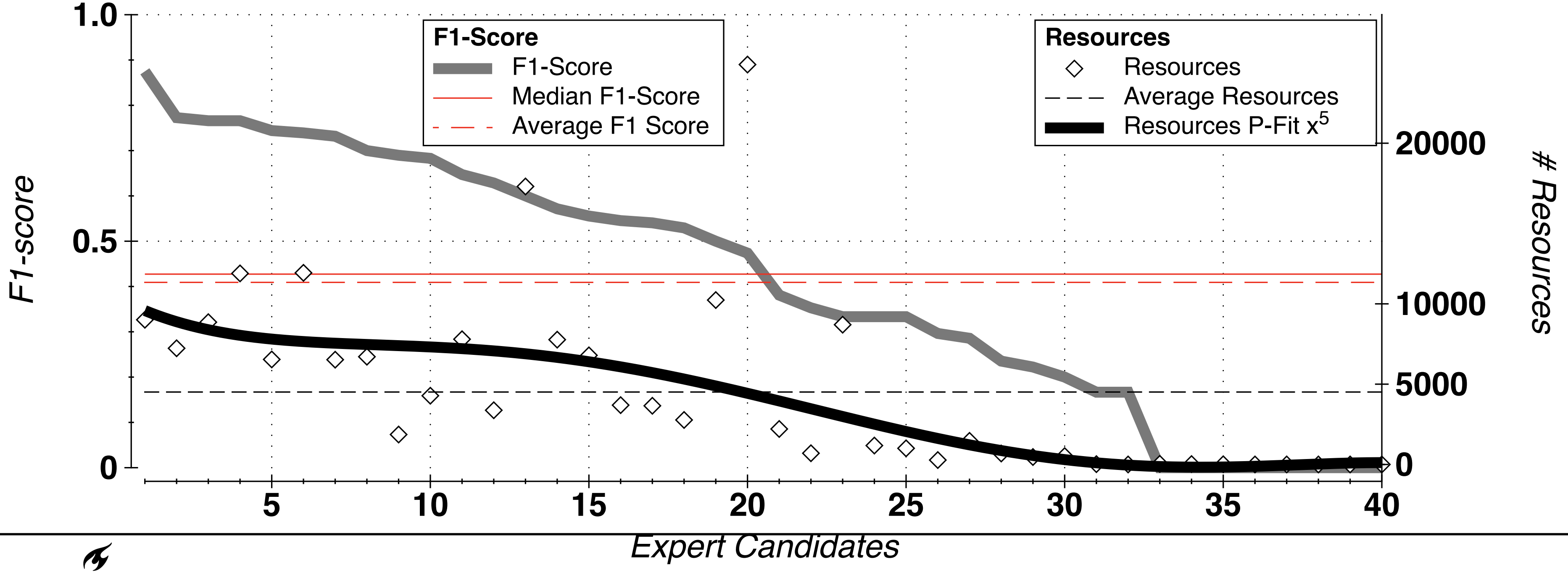
- **Twitter** is the **most effective** social network for expertise matching – sometimes it outperforms the other social networks
  - Twitter most effective in Computer Engineering, Science, Technology & Games, Sport

- **Facebook** effective in Locations, Sport, Movies & TV, Music

- **Linked-in** never very helpful in locating expertise

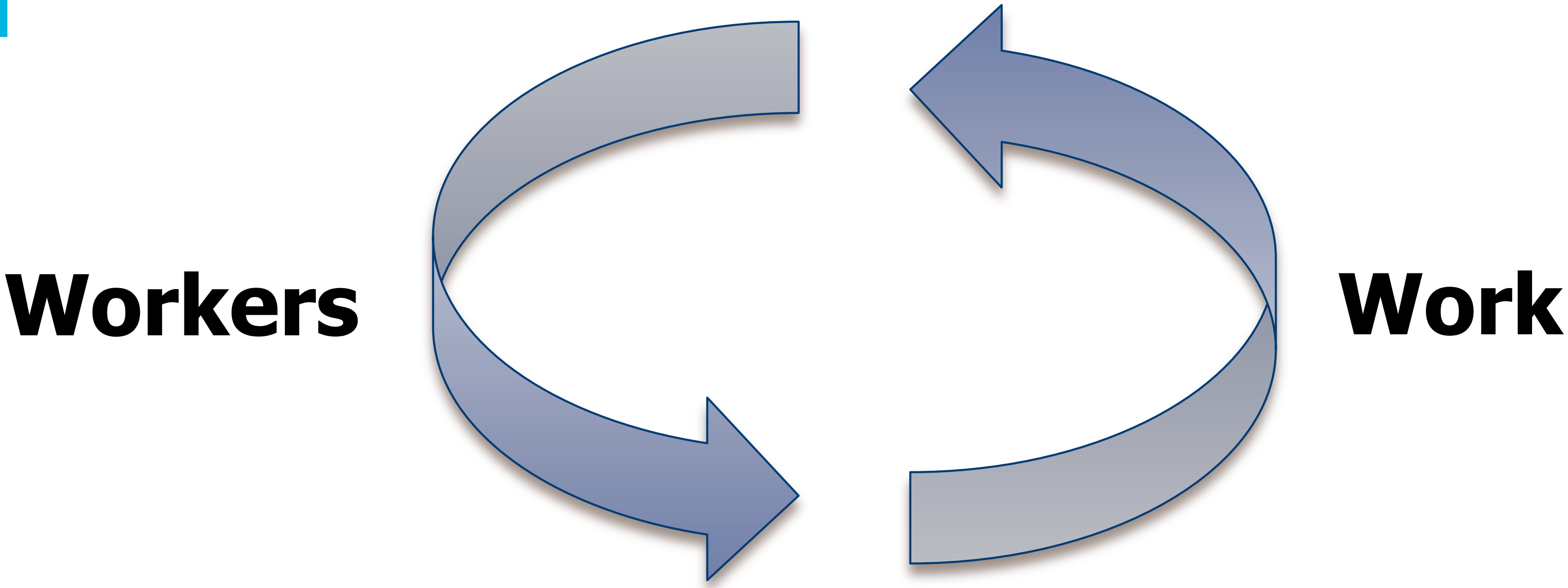
Trustworthiness of Social Information

- Strong correlation between the number of resources and the retrieval performance
- Users DO NOT expose all their interests on social networks



Better understanding of worker engagement mechanisms

Critical Mass



Worker Engagement

“money, love, or glory”

T. W. Malone, R. Laubacher, and C. Dellarocas. Harnessing Crowds: Mapping the Genome of Collective Intelligence. Working paper no. 2009-001, MIT Center for Collective Intelligence, Feb. 2009.

Thank You

Alessandro Bozzon - TU Delft

Software and Computer Technology Department

Faculty of Electrical Engineering, Mathematics, and Computer Science (EEMCS)

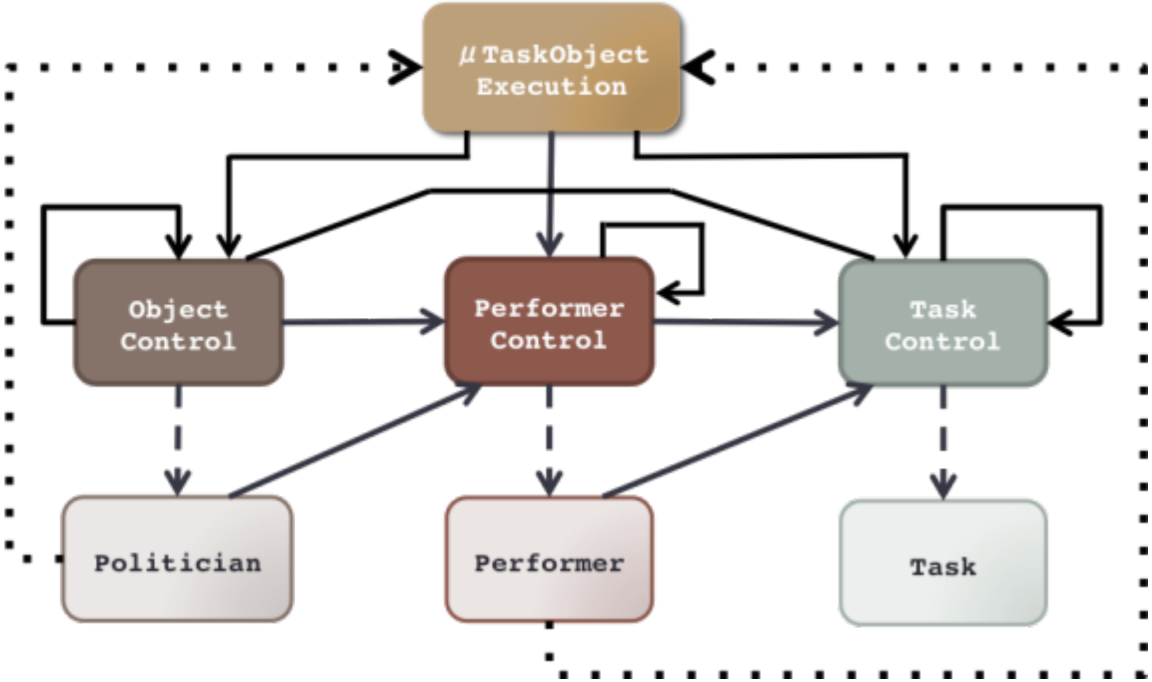
Website: www.alessandrobozzon.com

E-mail: a.bozzon@tudelft.nl

Twitter: @aleboz

Rule Programming Best Practice

- We define **three** classes of rules
  - → **Control rules**: modifying the *control* tables;
  - - - → **Result rules:** modifying the *dimension* tables (object, performer, task);
  - ···→ **Execution rules:** modifying the *execution* table, either directly or through re-planning

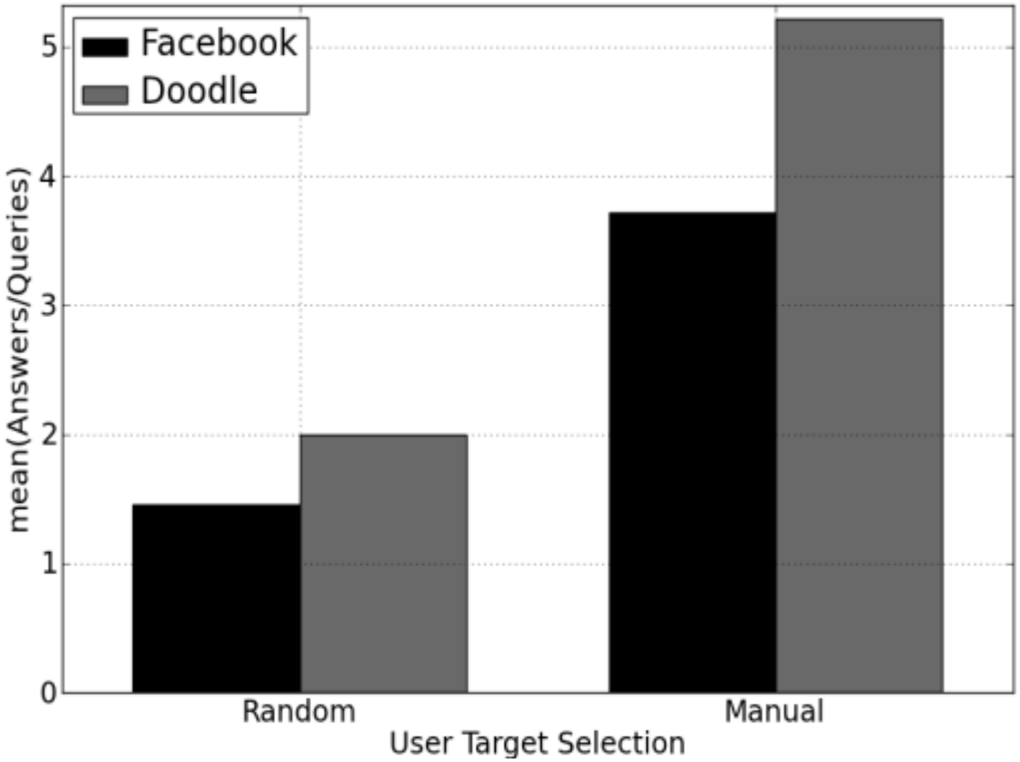


- Termination must be proven (rule precedence graph has cycles)

Crowdsearcher Experiment 1

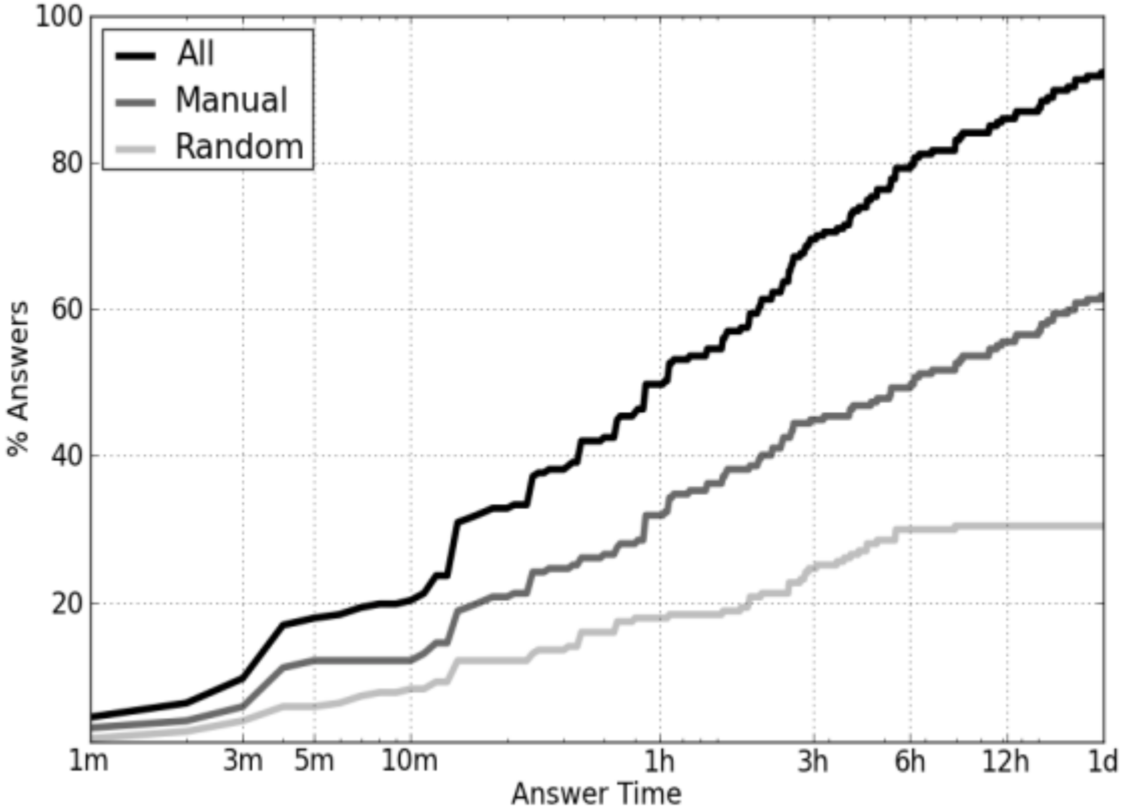
- **Goal:** Test engagement on social networks
- Some 150 users
- Two classes of experiments:
  - Random questions on fixed topics: interests (e.g. restaurants in the vicinity of Politecnico), to famous 2011 songs, or to top-quality EU soccer teams
  - Questions manually submitted by the users
- Different invitation strategies:
  - Random invitation
  - Explicit selection of responders by the asker
- Outcome
  - 175 *like* and *insert* queries
  - 1536 invitations to friends
  - 230 answers
  - 95 questions (~55%) got at least one answer

Manual and Random Questions



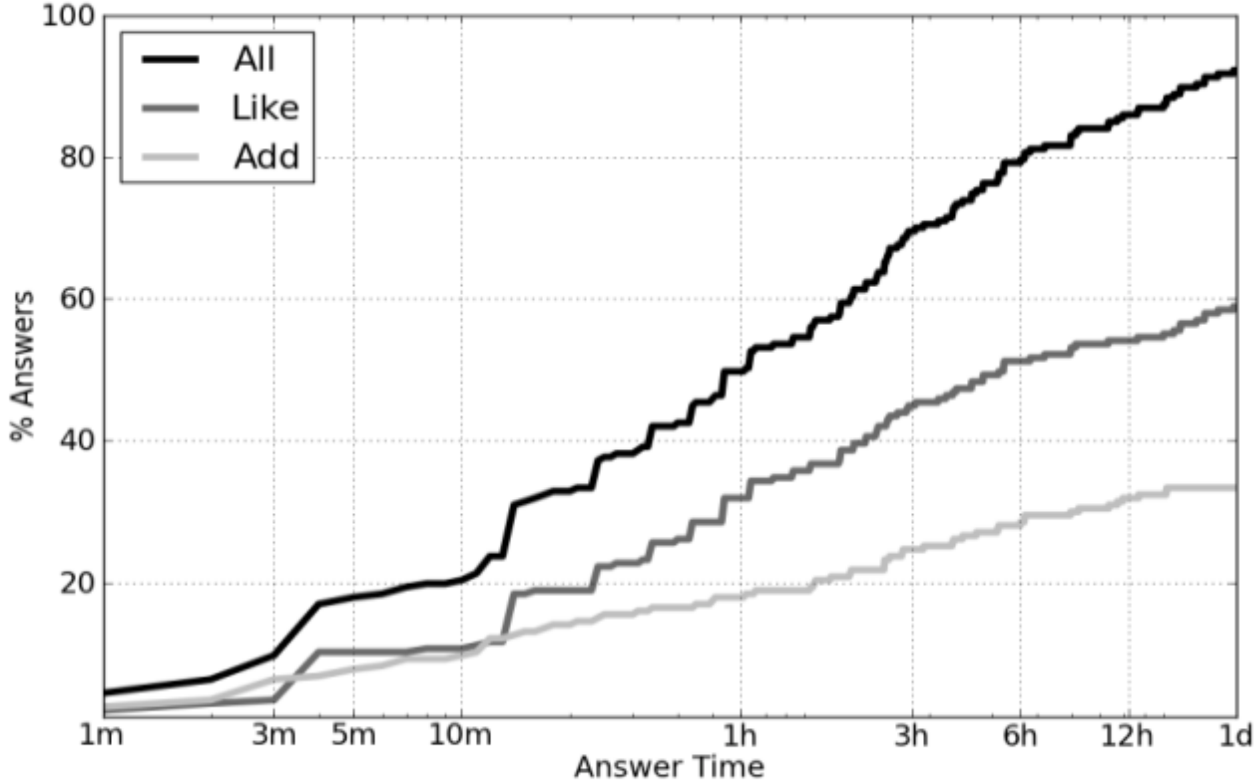
Interest / Rewarding Factor

- Manually written and assigned questions are consistently more responded in time

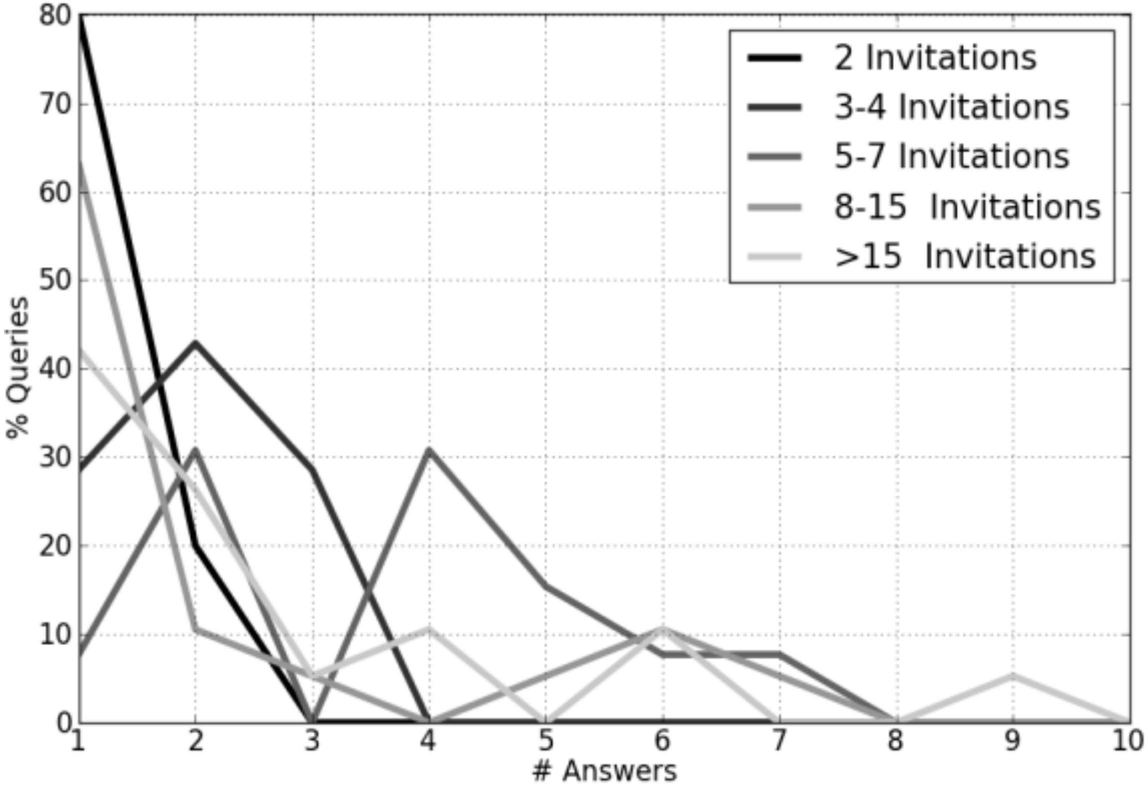


Query Type

- Engagement depends on the difficulty of the task
- Like vs. Add tasks:



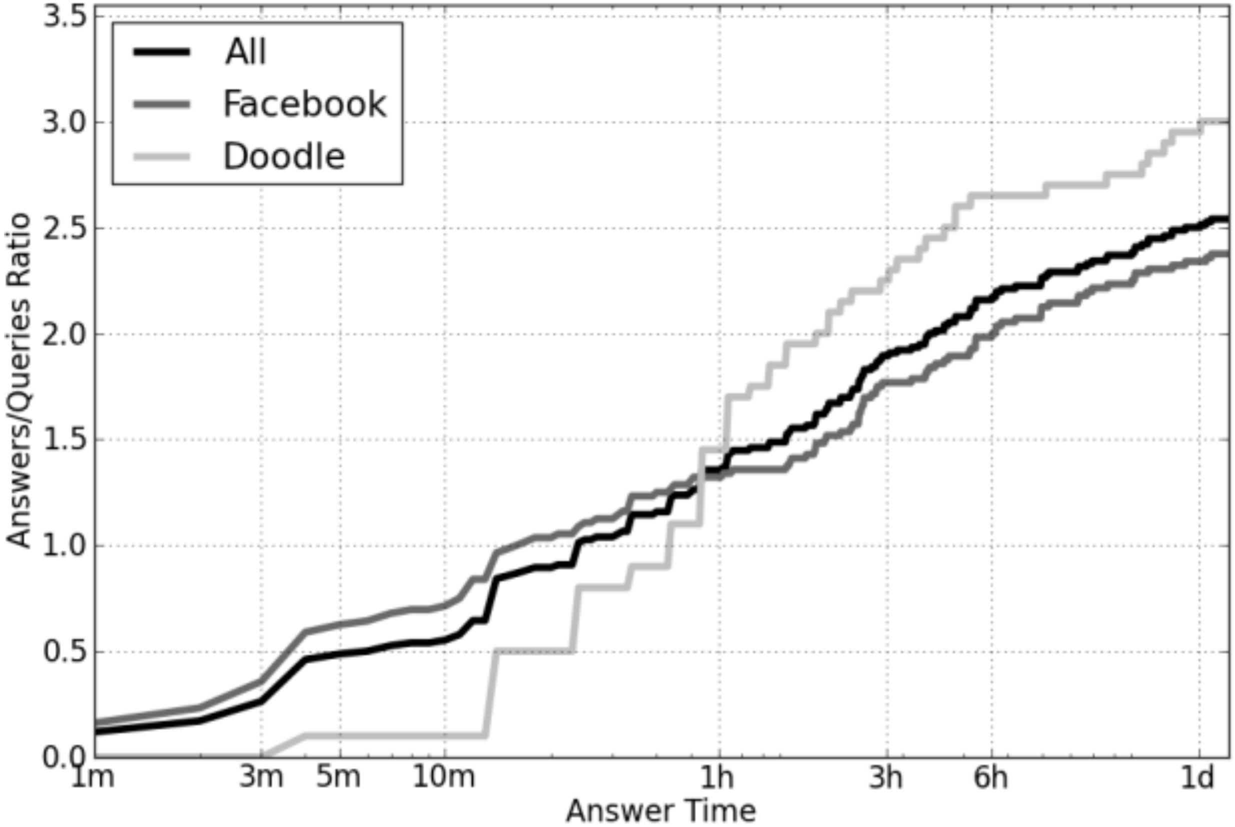
Experiments: Distribution of answers/invitation



Sometimes: more answers than invitations (limited cases)

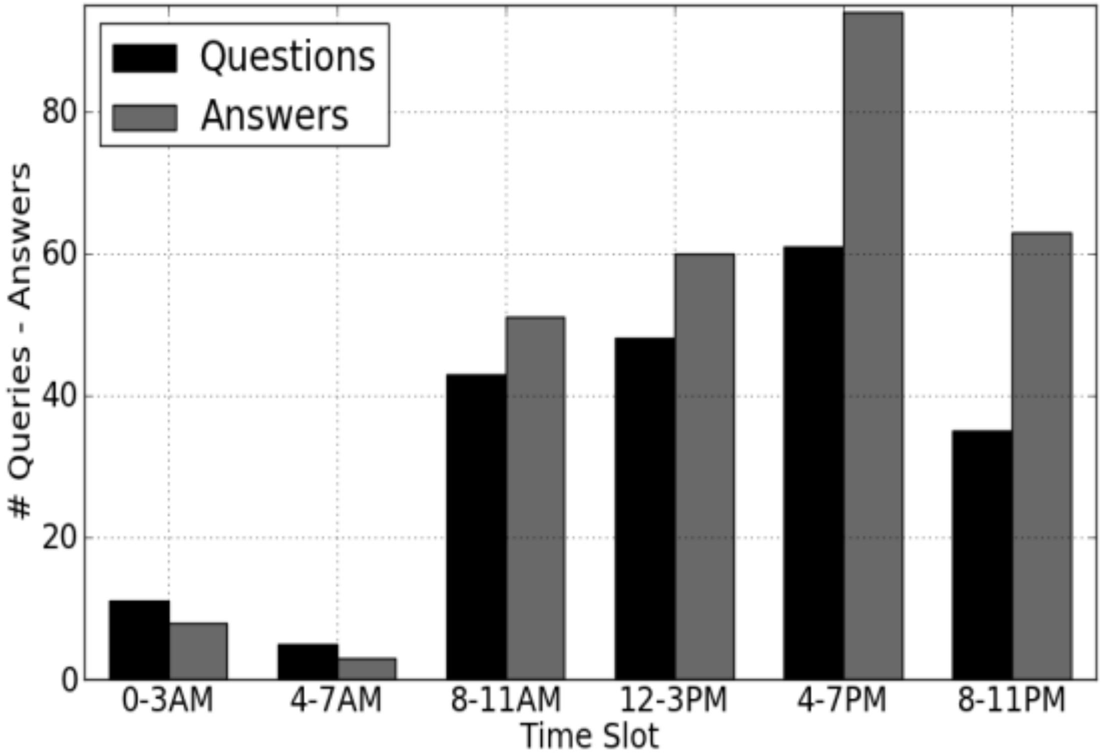
Comparison of Execution Platforms

- Facebook vs. Doodle



Posting Time

- Facebook vs. Doodle



Crowdsearcher Experiment 2

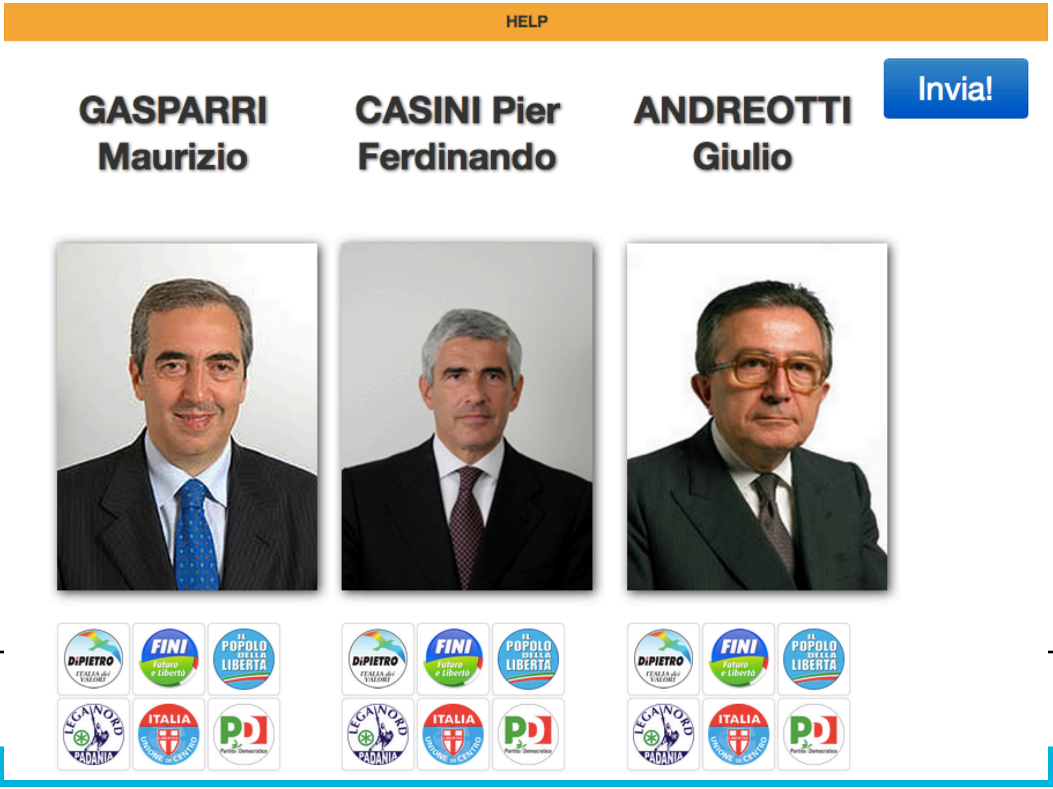
- **GOAL**: demonstrate the *flexibility* and *expressive power* of reactive crowdsourcing

- 3 experiments, focused on Italian politicians
  - **Parties**: Human Computation → affiliation classification
  - **Law**: Game With a Purpose → guess the convicted politician
  - **Order**: Pure Game → hot or not
- 1 week (November 2012)
- 284 distinct performers
  - Recruited through public mailing lists and social networks announcements
- 3500 Micro Tasks

Politician Affiliation

- Given the picture and name of a politician, specify his/her political affiliation
  - No time limit
  - Performers are encouraged to look up online

- 2 set of rules
  - Majority Evaluation
  - Spammer Detection

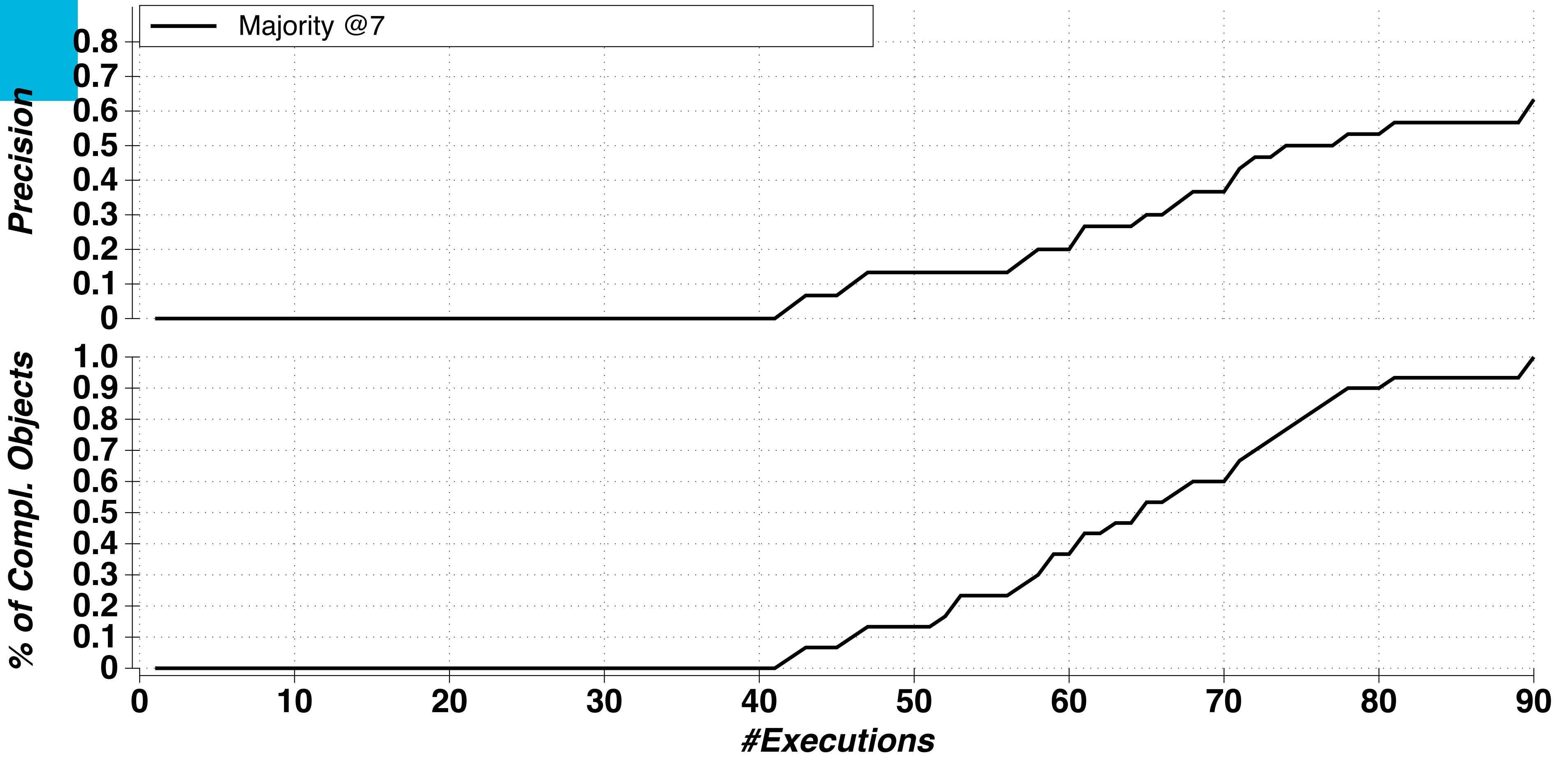


GASPARRI Maurizio

CASINI Pier Ferdinando

ANDREOTTI Giulio

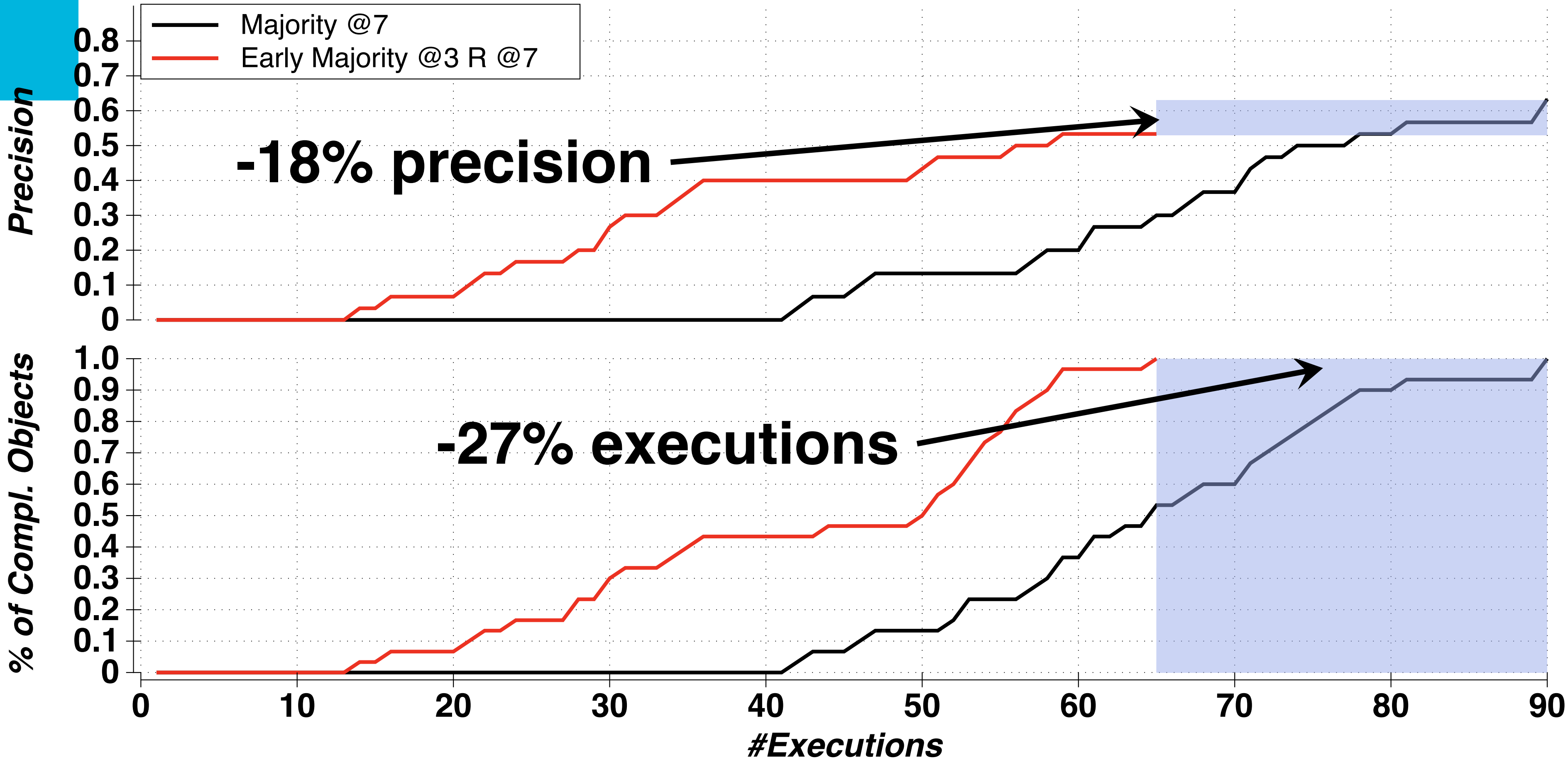
Results – Majority Evaluation_1/3



30 object; object redundancy = 9;

Final object classification as simple majority after 7 evaluations

Results - Majority Evaluation$_{2/3}$



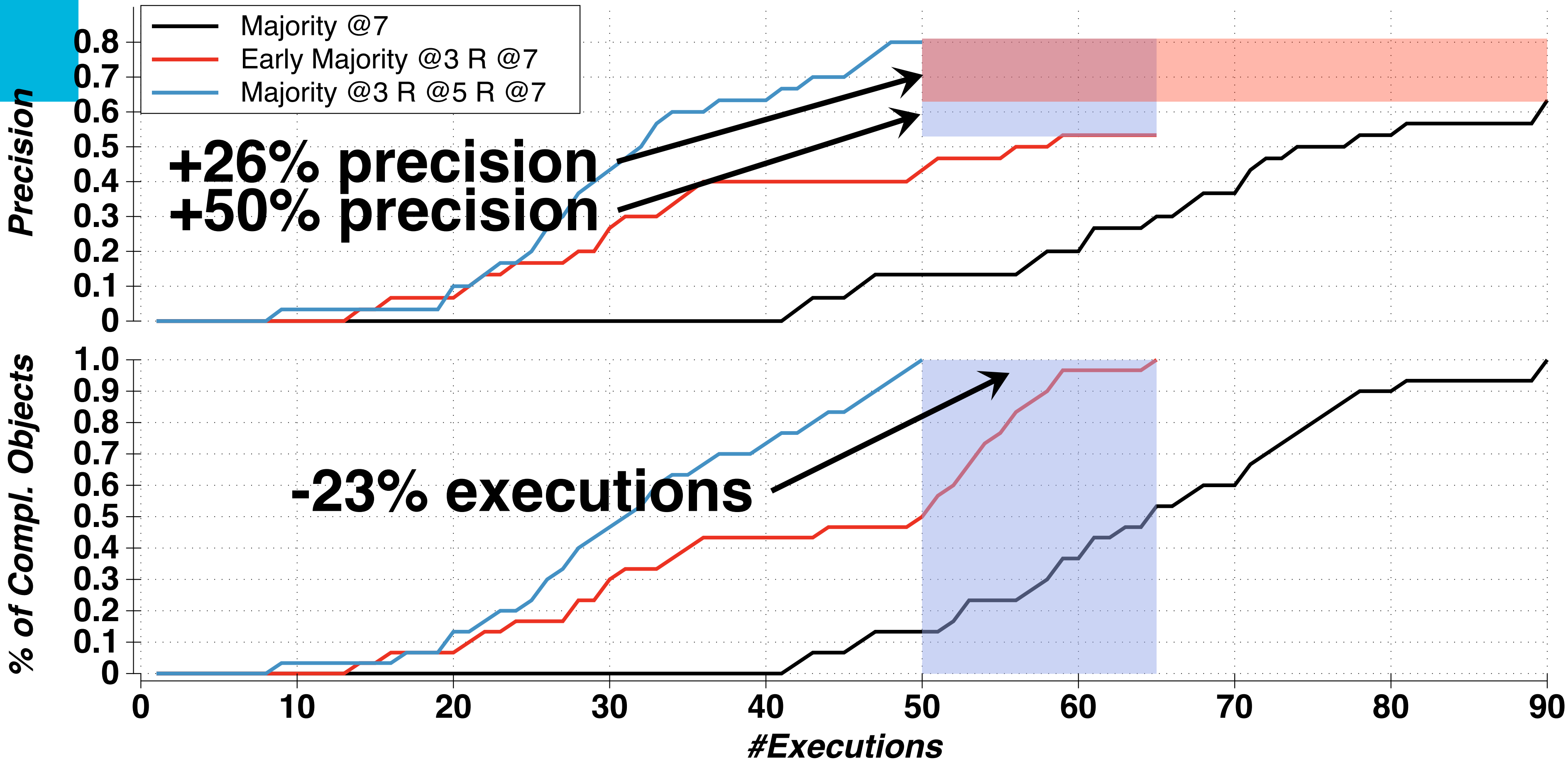
-18% precision

-27% executions

Final object classification as **total** majority after 3 evaluations

Otherwise, re-plan of 4 additional evaluations. Then simple majority at 7

Results - Majority Evaluation_3/3



+26% precision

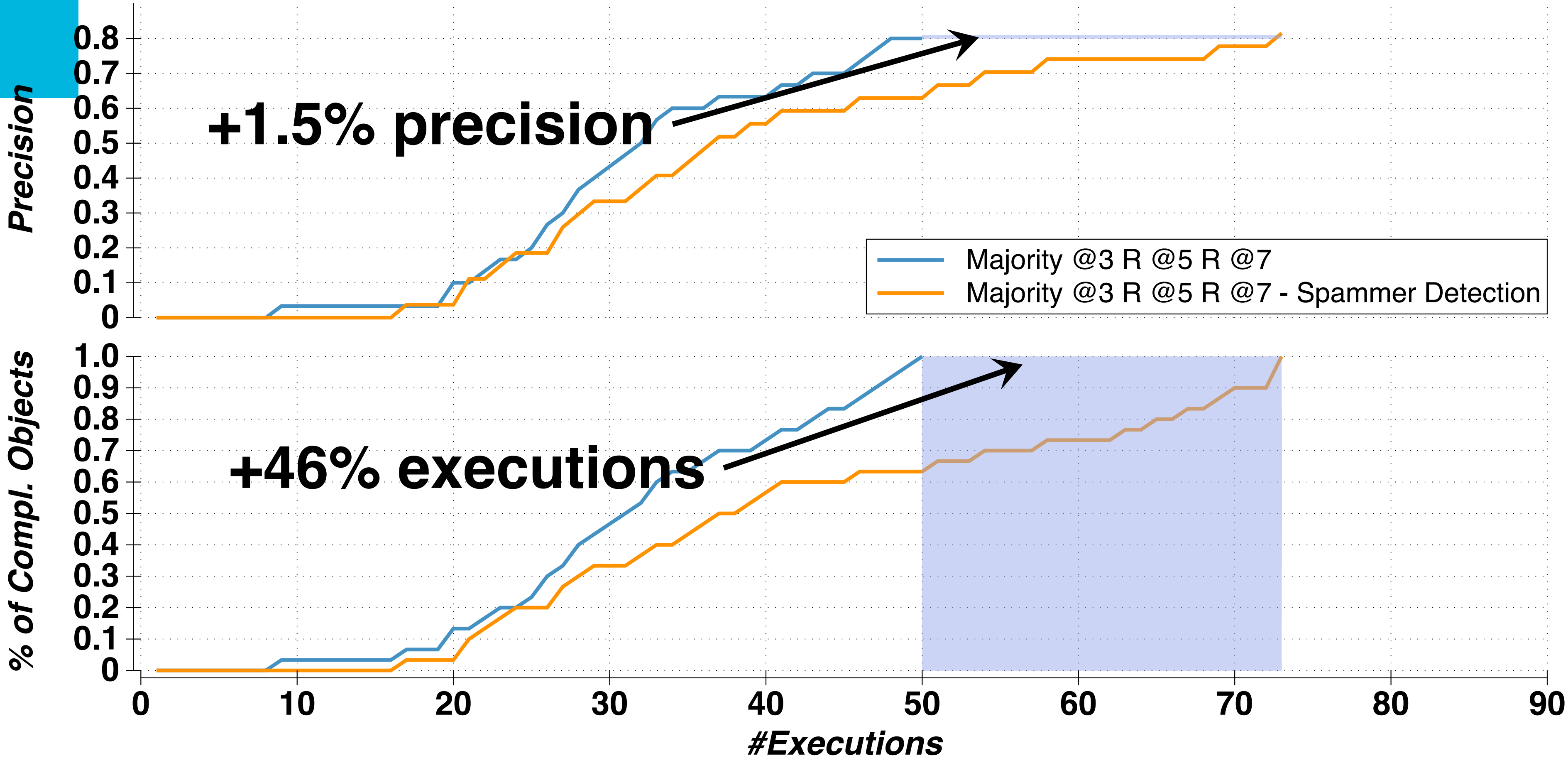
+50% precision

-23% executions

Final object classification as **total** majority after 3 evaluations

Otherwise, **simple** majority at 5 or at 7 (with replan)

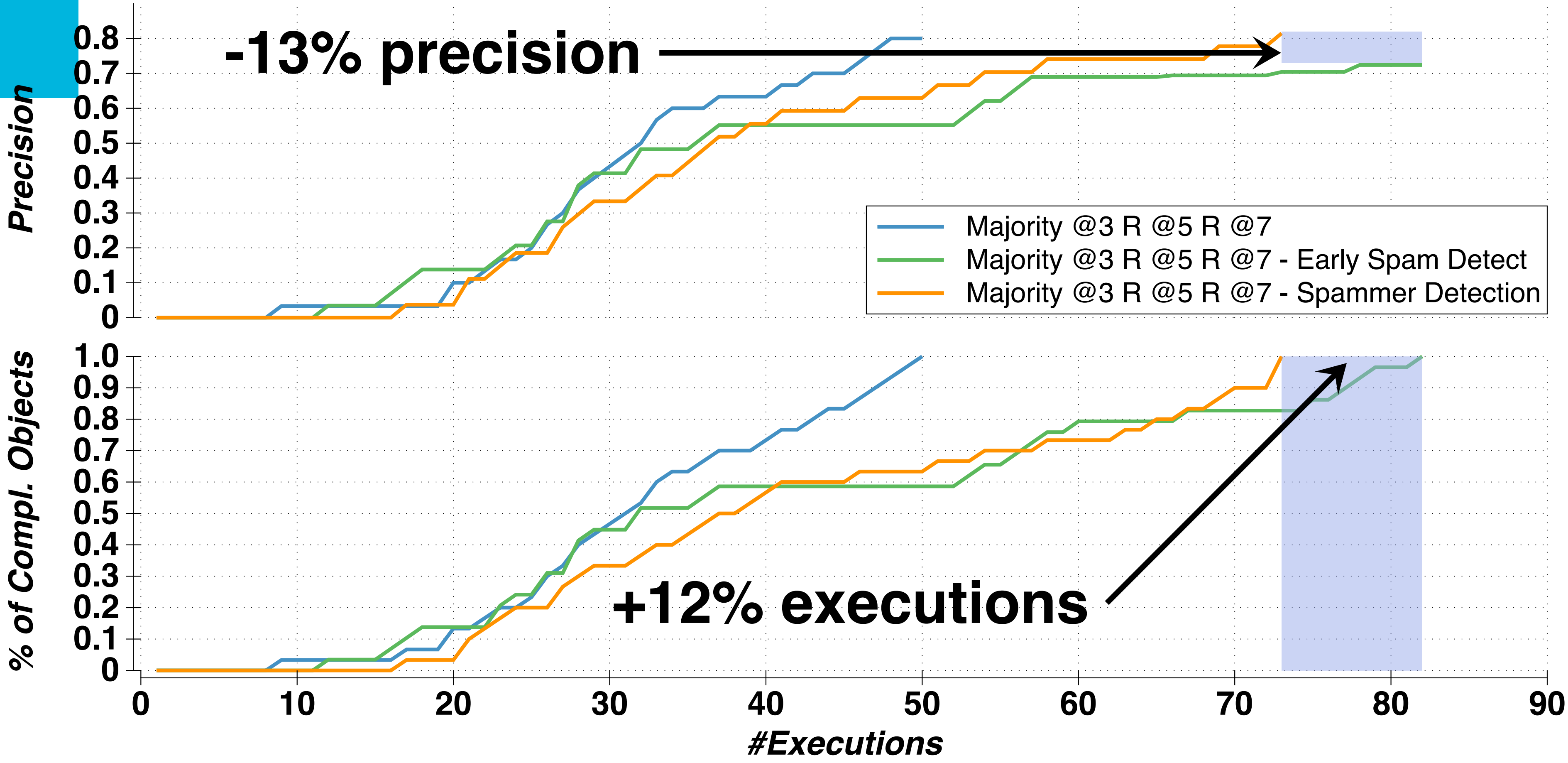
Results – Spammer Detection_1/2



New rule for spammer detection **without ground truth**

Performer correctness on **final** majority. Spammer if > 50% wrong classifications

Results – Spammer Detection_1/2



New rule for spammer detection **<u>without ground truth</u>**

Performer correctness on **current** majority. Spammer if > 50% wrong classifications