NETS 213: CROWDSOURCING AND HUMAN COMPUTATION

Expert Crowds

Professor Chris Callison-Burch

Penn Engineering
UNIVERSITY of PENNSYLVANIA

Recruiting is hard

- MTurk, CrowdFlower, oDesk, or Freelancer gives us access to a lot of people
- But are they useful for specialized skills?

Attracting Contributors via Online Advertising

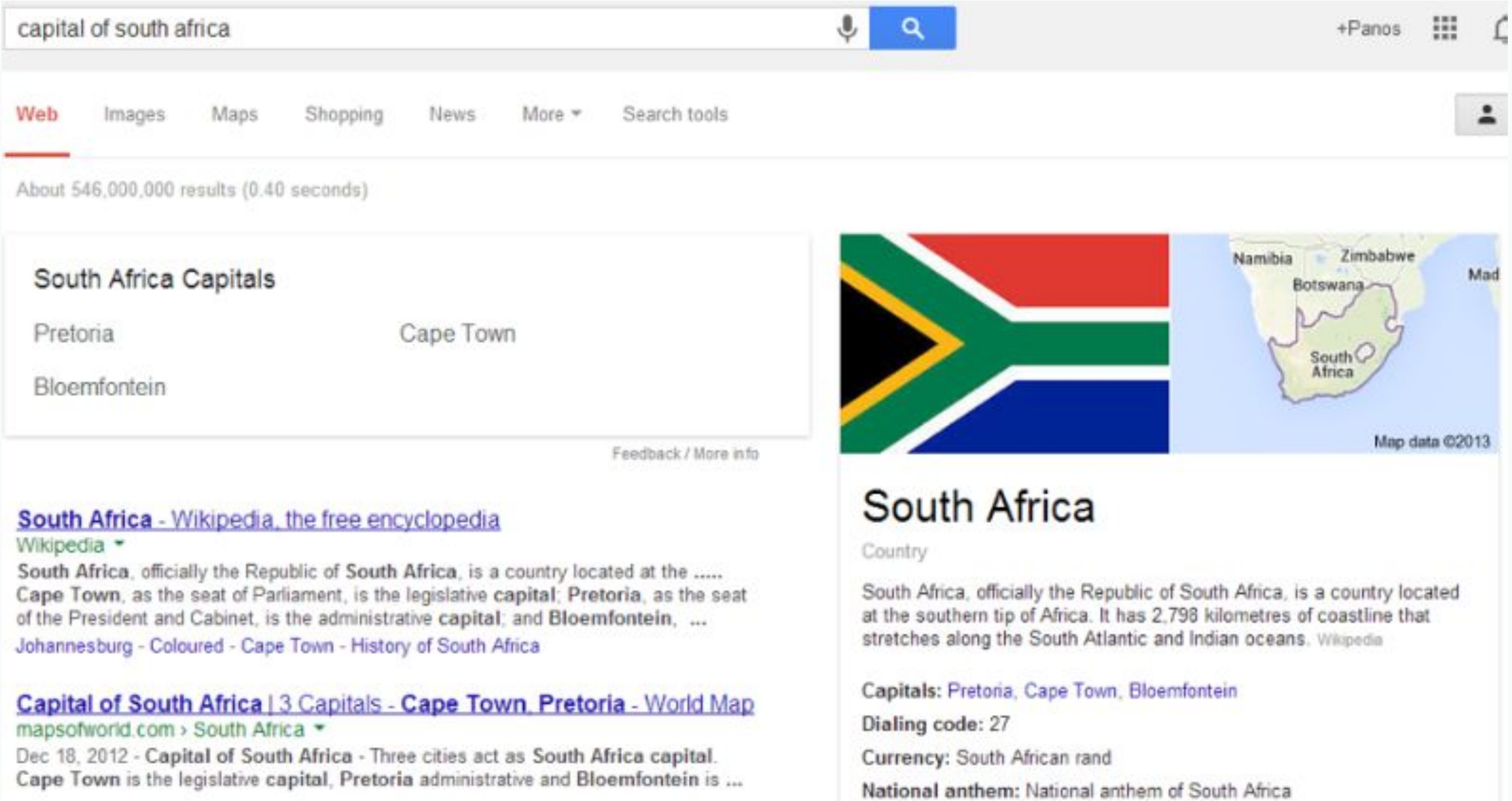
Panos Ipeirotis spent a sabbatical at Google, and they tasked him with finding experts to fill in their Knowledge Graph

“We have a billion users… leverage their knowledge …”

“Let’s create a new crowdsourcing system…”

“Crowdsource in a predictable manner, with knowledgeable users, without introducing monetary rewards”

Knowledge Graph: Things not Strings



capital of south africa

About 546,000,000 results (0.40 seconds)

South Africa Capitals

Pretoria

Cape Town

Bloemfontein

Feedback / More info

South Africa - Wikipedia, the free encyclopedia
Wikipedia
South Africa, officially the Republic of **South Africa**, is a country located at the ..... **Cape Town**, as the seat of Parliament, is the legislative **capital**; **Pretoria**, as the seat of the President and Cabinet, is the administrative **capital**; and **Bloemfontein**, ...
Johannesburg - Coloured - Cape Town - History of South Africa

Capital of South Africa | 3 Capitals - **Cape Town**, **Pretoria** - World Map
mapsofworld.com › South Africa
Dec 18, 2012 - **Capital of South Africa** - Three cities act as **South Africa capital**. **Cape Town** is the legislative **capital**, **Pretoria** administrative and **Bloemfontein** is ...

South Africa

Country

South Africa, officially the Republic of South Africa, is a country located at the southern tip of Africa. It has 2,798 kilometres of coastline that stretches along the South Atlantic and Indian oceans. Wikipedia

Capitals: Pretoria, Cape Town, Bloemfontein

Dialing code: 27

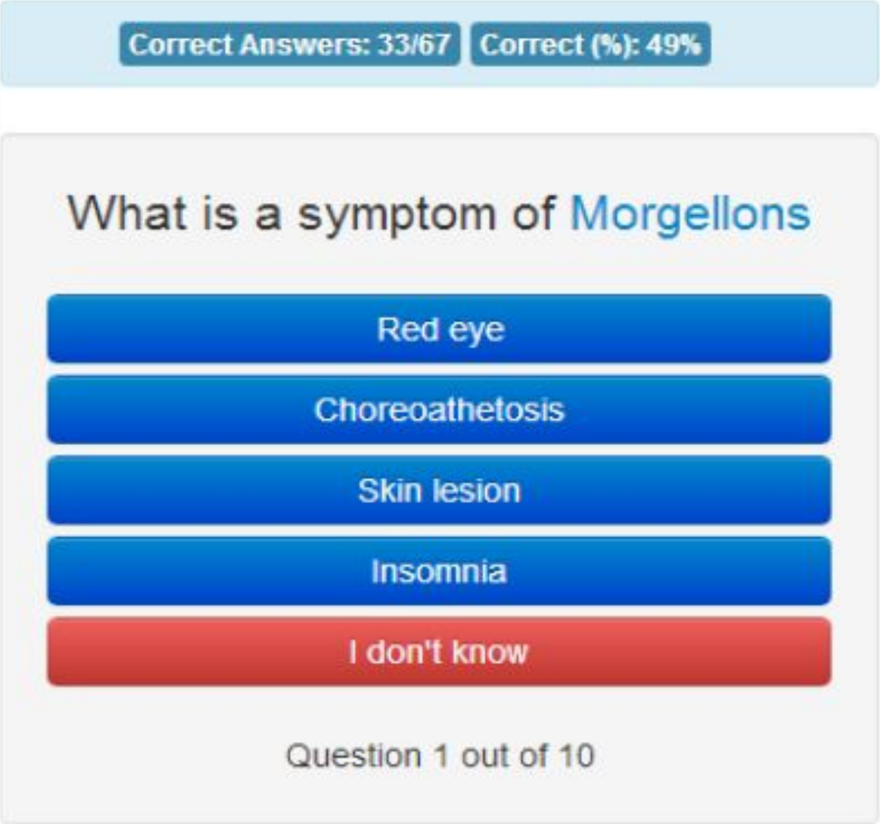
Currency: South African rand

National anthem: National anthem of South Africa

Still incomplete...

- “Symptom of strep throat”
- “Side effects of treximet”
- “Who is Cristiano Ronaldo dating
- “When is Jay Z playing in New York”
- “What is the customer service number for Google”
- ...

Quizz



Correct Answers: 33/67 Correct (%): 49%

What is a symptom of Morgellons

Question 1 out of 10

Calibration vs. Collection

- Calibration questions (known answer): Evaluating user competence on topic at hand
- Collection questions (unknown answer): Asking questions for things we do not know
- Trust more answers coming from competent users

Tradeoff

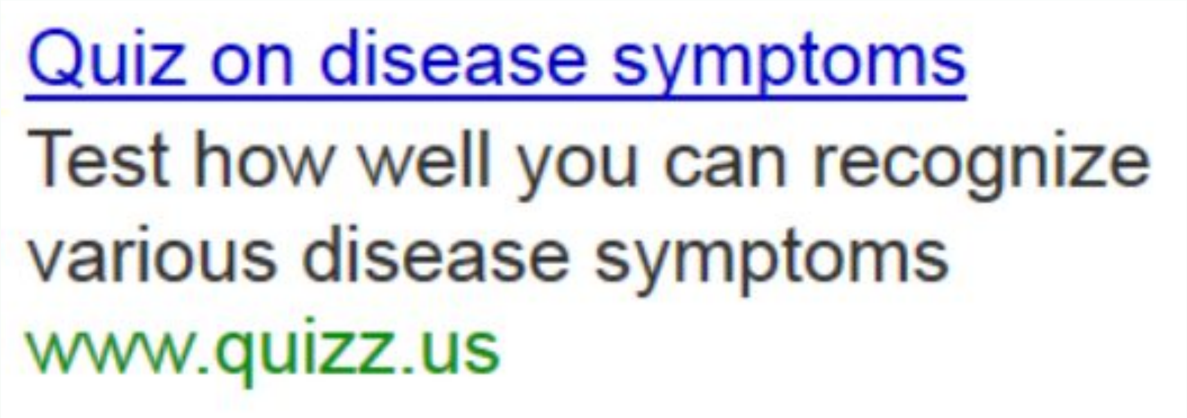
Learn more about user quality vs. getting answers

(technical solution: use a Markov Decision Process)

Challenges

- Why would anyone come and play this game?
- Why would knowledgeable users come?
- Wouldn’t it be simpler to just pay?

Attracting Visitors: Ad Campaigns



Quiz on disease symptoms

Test how well you can recognize various disease symptoms

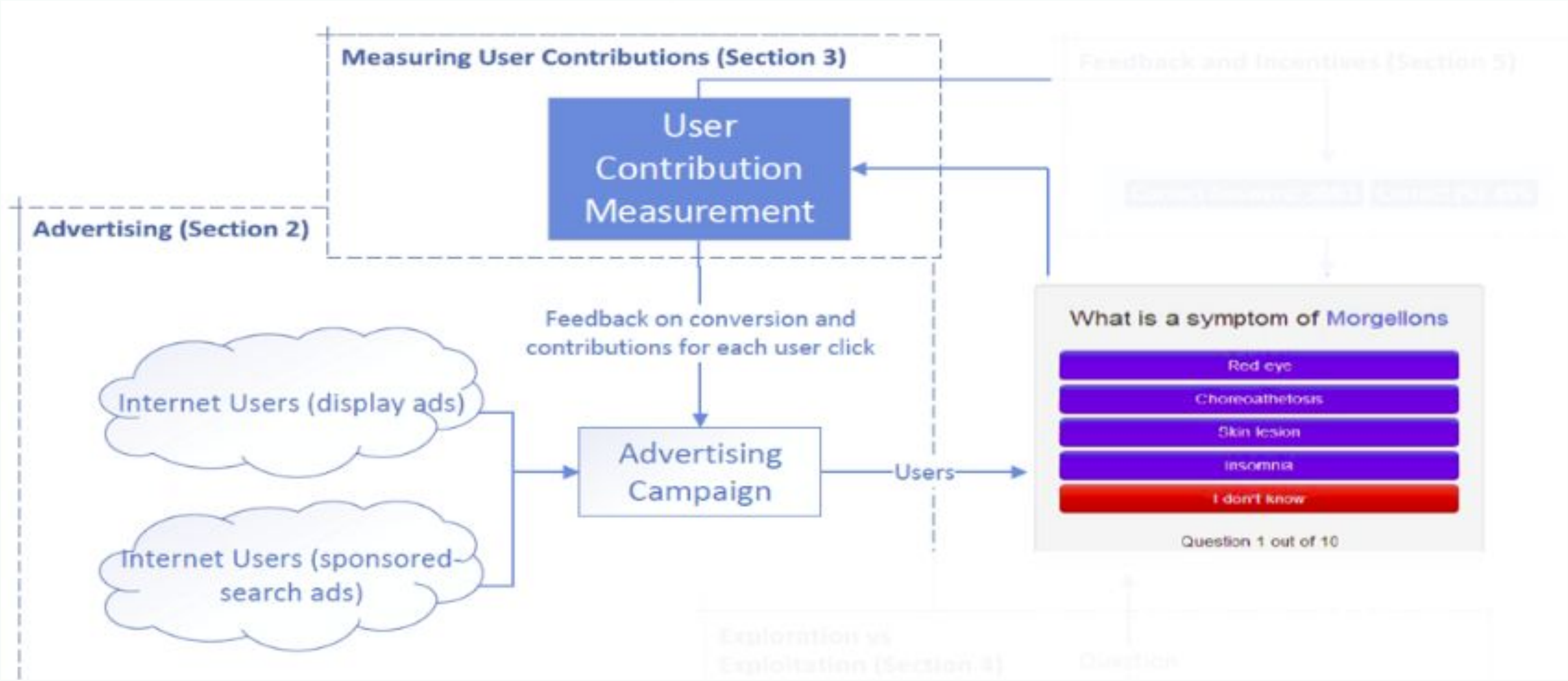
www.quizz.us

Running Ad Campaigns: Objectives

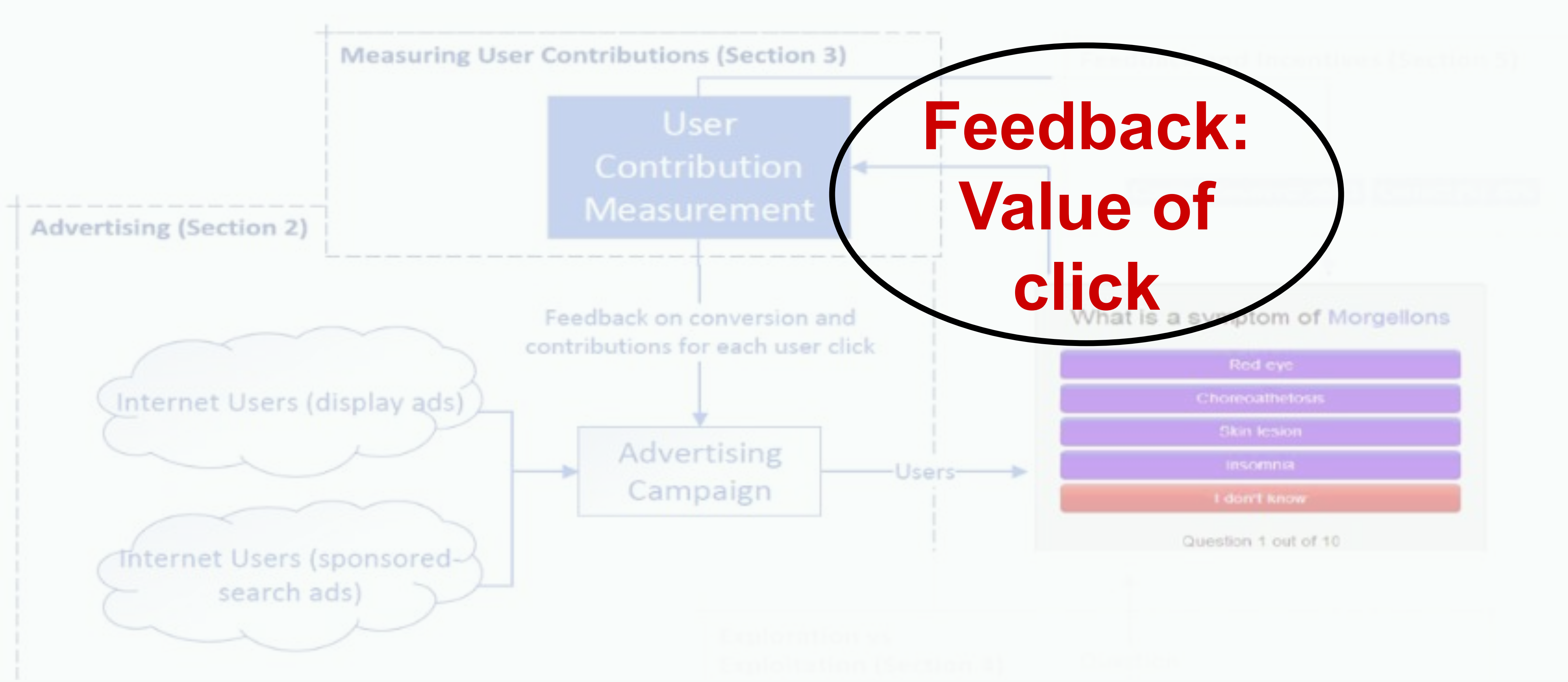
- We want to attract good users, not just clicks
- We do not want to think hard about keyword selection, appropriate ad text, etc.
- We want automation across thousands of topics

  (from treatment side effects to celebrity dating)

Solution: Treat Quizz as eCommerce Site


Measuring User Contributions (Section 3)
User Contribution Measurement
Advertising (Section 2)
Feedback on conversion and contributions for each user click
Internet Users (display ads)
Internet Users (sponsored-search ads)
Advertising Campaign
Users
Feedback and Incentives (Section 5)
What is a symptom of Morgellons
Red eye
Choreoathetosis
Skin lesion
insomnia
I don't know
Question 1 out of 10
Exploration vs Exploitation (Section 4)
Question

Solution: Treat Quizz as eCommerce Site


Measuring User Contributions (Section 3)
User Contribution Measurement
Feedback and Incentives (Section 5)
Advertising (Section 2)
Feedback on conversion and contributions for each user click
Internet Users (display ads)
Internet Users (sponsored-search ads)
Advertising Campaign
Users
What is a symptom of Morgellons
Red eye
Choreoathetosis
Skin lesion
Insomnia
I don't know
Question 1 out of 10
Exploration vs. Exploitation (Section 4)
Question

Feedback: Value of click

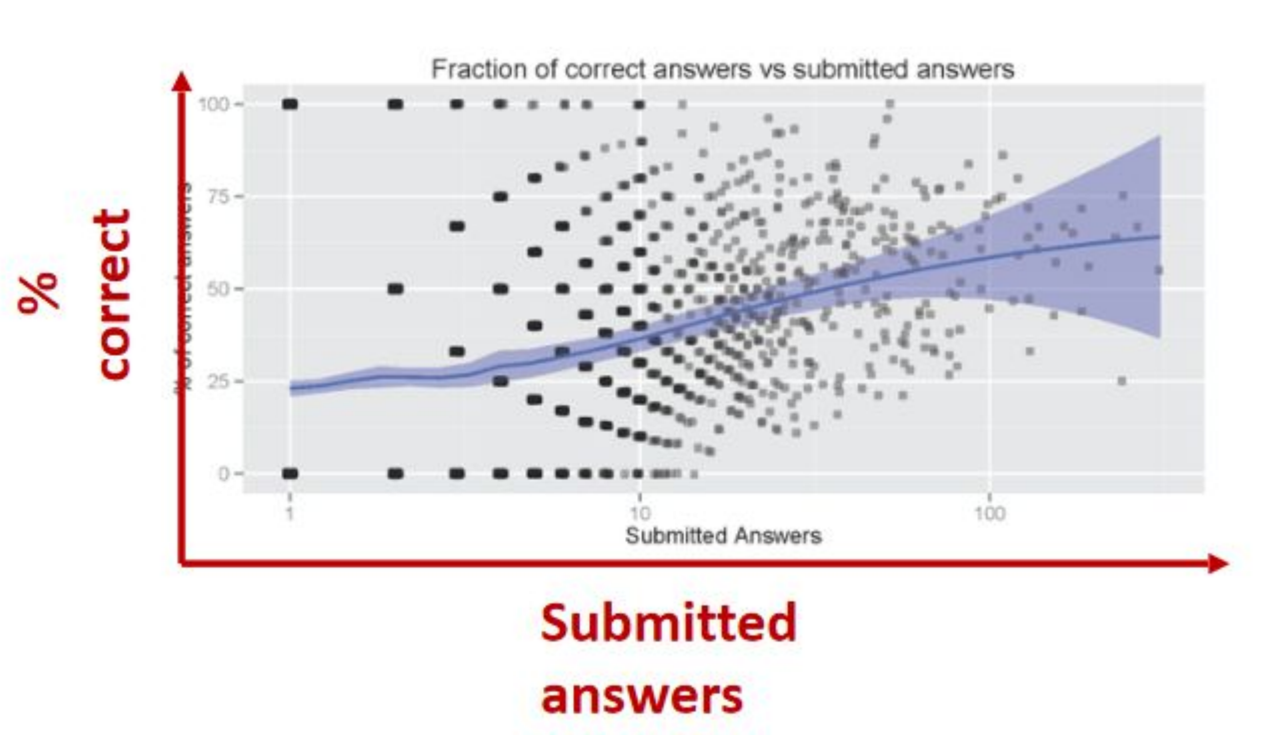
Example of Targeting: Medical Quizzes

- Medical topics They the best performing quizzes…

- Users coming from sites such as Mayo Clinic, WebMD

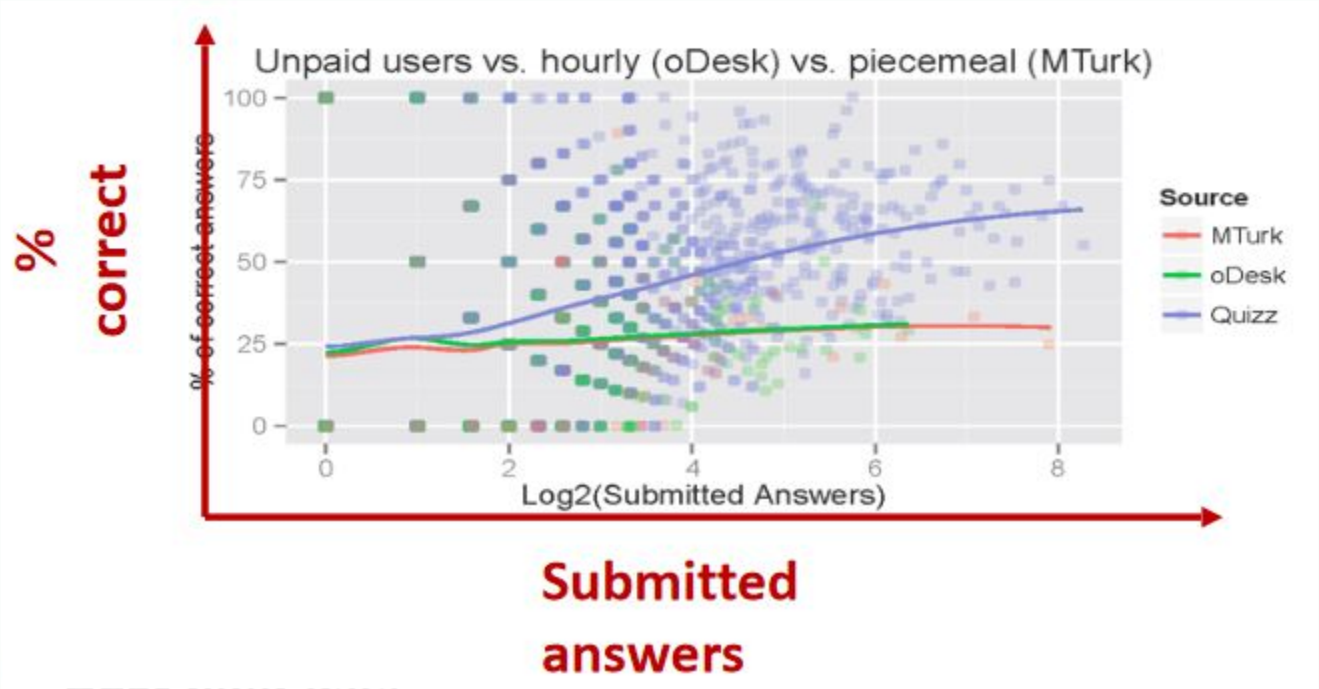
- Likely “prosumers” (proactive consumers, not professionals

Self-selection and participation



- Low performing users naturally drop out
- With paid users, monetary incentives keep them

Comparison with paid crowdsourcing



- **Best** paid user
  - 68% quality, 40 answers (~1.5 minutes per question)
  - Quality-equivalency: 13 answers @ 99% accuracy, 23 answers @ 90% accuracy
  - 5 cents/question, or $3/hr to match advertising cost of unpaid users

- Knowledgeable users are much faster and more efficient

Targeted Advertising

- New way to run crowdsourcing, targeting with ads
- Engages unpaid users, avoids problems with extrinsic rewards
- Provides access to expert users, not available labor platforms
- Experts not always professionals (e.g., Mayo Clinic users)

Online Labor Markets

- Help employers and employees connect

- Face a similar challenge

- How do they assess worker skills?

mobileworks
oDesk
+
Elance
freelancer.com
amazon mechanical turk
Artificial Artificial Intelligence

Skill Testing

- Skill certification through testing
- Workers take online tests
- Display score on profile
- Tests licensed from companies ExpertRating
- Domain-experts paid to create questions
- Static question banks

ExpertRating Categories

- Airlines and Aviation
- Building & Construction
- Career guidance
- Clothing and Fashion
- Engineering
- English language skills
- Finance & Accounting
- Food and hospitality
- Foreign language skills
- Graphic design
- Healthcare
- IT & Computer skills
- Law
- Management
- Media
- Medical transcription and billing
- Office temp skills
- Sales and Marketing

Problems

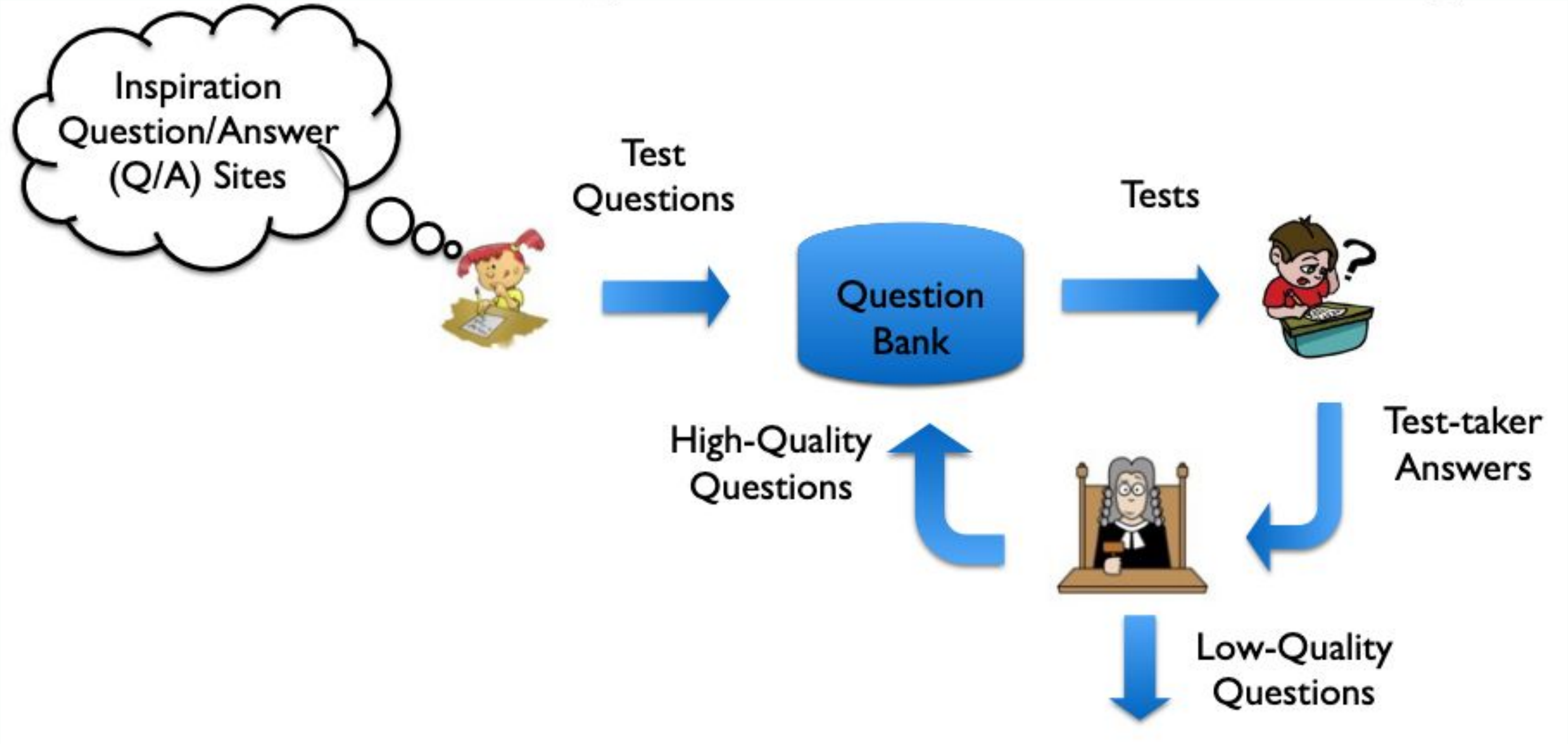
- Static Question Banks
  - Questions become outdated
  - Cheating
- Lack of evaluation
  - Questionable long-term performance predictors
  - Questions may have errors or ambiguities

STEP: A Scalable Testing and Evaluation Platform

Christoforaki and Ipeirotis (2014)

- Continuously generate new questions
  - Make tests more cheating proof
  - Keep questions up-to-date
- Evaluate question quality
  - Identify errors or ambiguities
  - Use real-market performance data for evaluation

STEP system summary



Stack Overflow

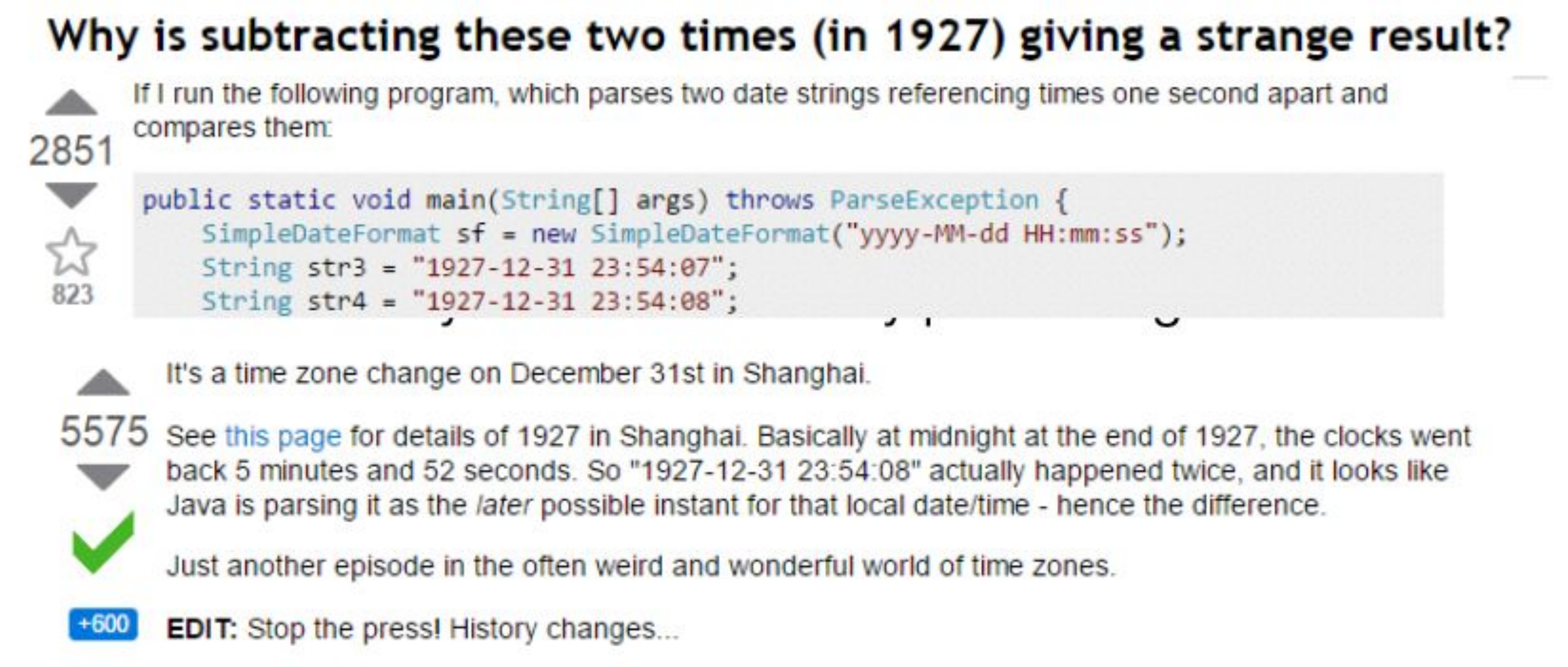
“A Q/A site for professional and enthusiast programmers”

- 3 million subscribed users
- 8 million questions
- 35K tags
- 91% at least one answer

Topic	Questions	%
Java	737,563	8.9
Javascript	723,150	8.7
C#	714,774	8.6
PHP	658,827	8.0
Android	585,017	7.1
Jquery	545,775	6.6
Python	355,093	4.6
HTML	352,146	4.2
C++	325,667	3.9
mysql	280,946	3.4

Stack Overflow Challenges

- Volume of questions
  - Large base of candidate questions for tests



Why is subtracting these two times (in 1927) giving a strange result?

2851

823

If I run the following program, which parses two date strings referencing times one second apart and compares them:

```
public static void main(String[] args) throws ParseException {
    SimpleDateFormat sf = new SimpleDateFormat("yyyy-MM-dd HH:mm:ss");
    String str3 = "1927-12-31 23:54:07";
    String str4 = "1927-12-31 23:54:08";
```

5575

It's a time zone change on December 31st in Shanghai.

See this page for details of 1927 in Shanghai. Basically at midnight at the end of 1927, the clocks went back 5 minutes and 52 seconds. So "1927-12-31 23:54:08" actually happened twice, and it looks like Java is parsing it as the *later* possible instant for that local date/time - hence the difference.

Just another episode in the often weird and wonderful world of time zones.

+600

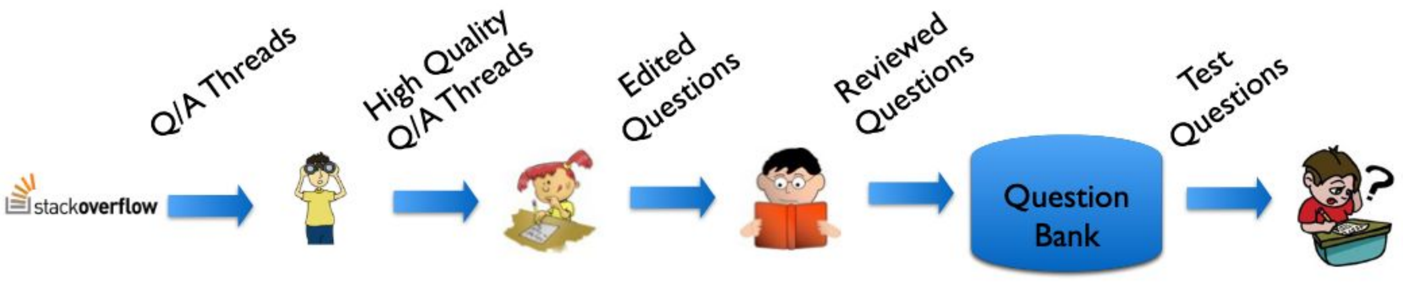
EDIT: Stop the press! History changes...

Question Spotter

- Identifies promising Q/A threads 👍👎
- Train classifier with obtained labels: ~90% precision

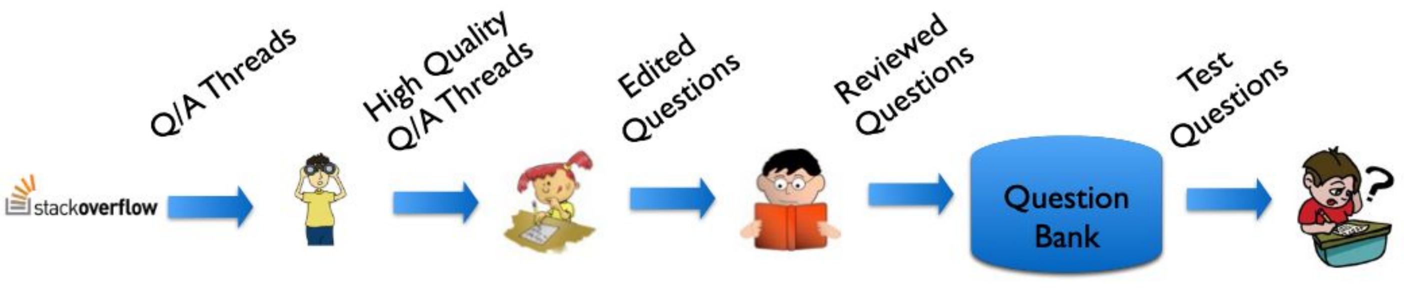
Features

- ○ Question text length
- ○ Answer count
- ○ Answer score entropy
- ○ Popularity distribution of tags
- ○ Question popularity score
- ○ Weekly view count
- ○ Max answer author reputation



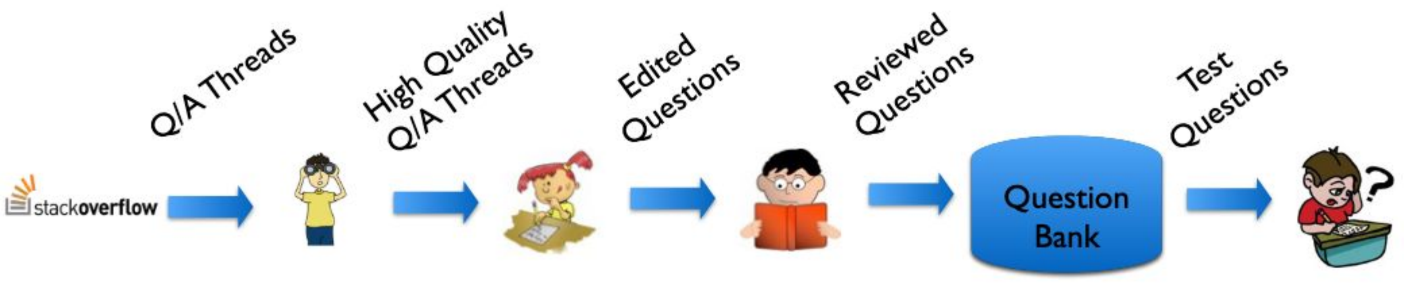
Question Editor

- Humans with expertise in topic at hand
- Visit and read promising Q/A thread
- Reformulate into multiple choice test-question
- Discard questions not considered appropriate



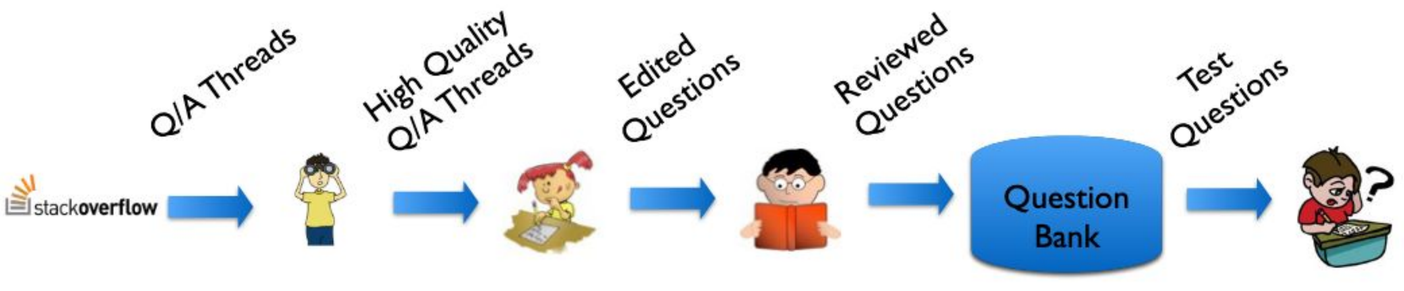
Question Reviewer

- Have a good handle of English Language,
- Check for spelling, grammar
- Check for compliance with test standards
  - Vocabulary usage
  - Question text length
  - Answer count
  - Answer text length
- Reviewers do not need to be topic experts

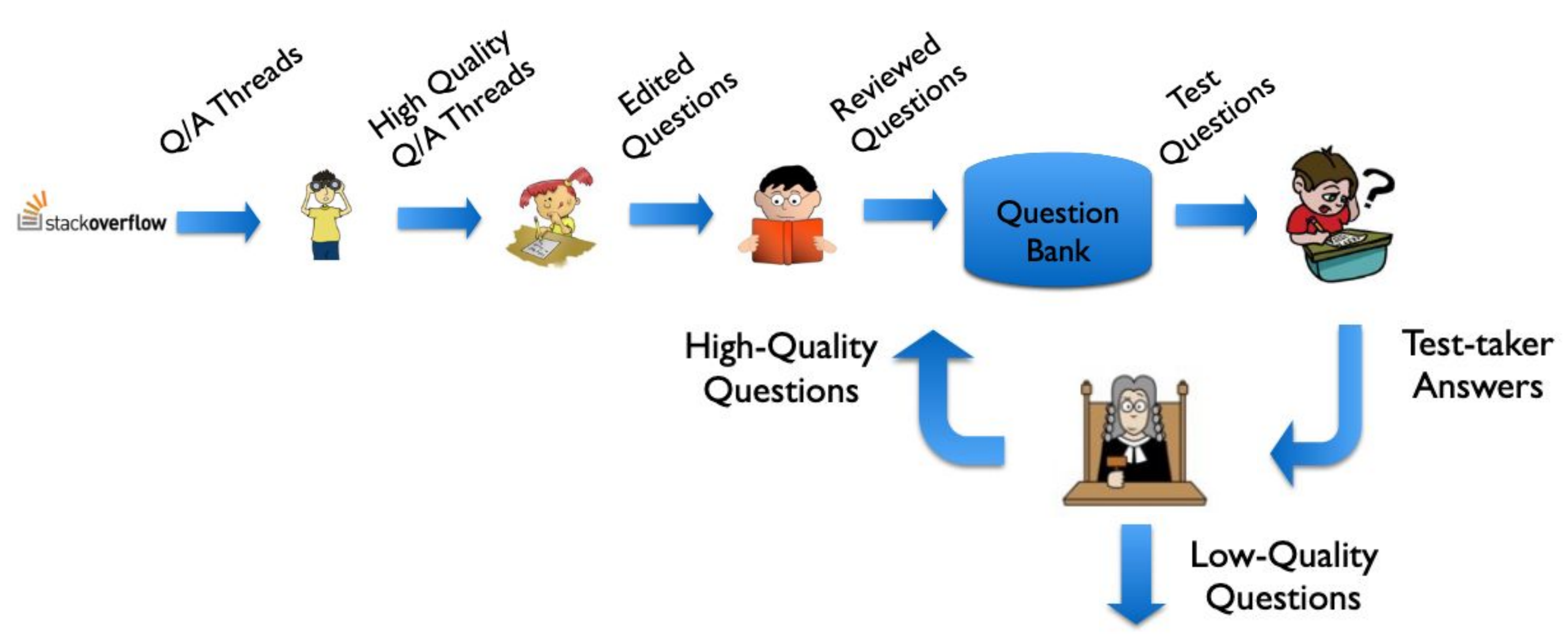


Question Bank

- Experimental Question Bank
  - Stores newly created questions
  - Not used for test-taker evaluation
  - Gather answers waiting for evaluation
- Production Question Bank
  - Are used for the test-taker evaluation



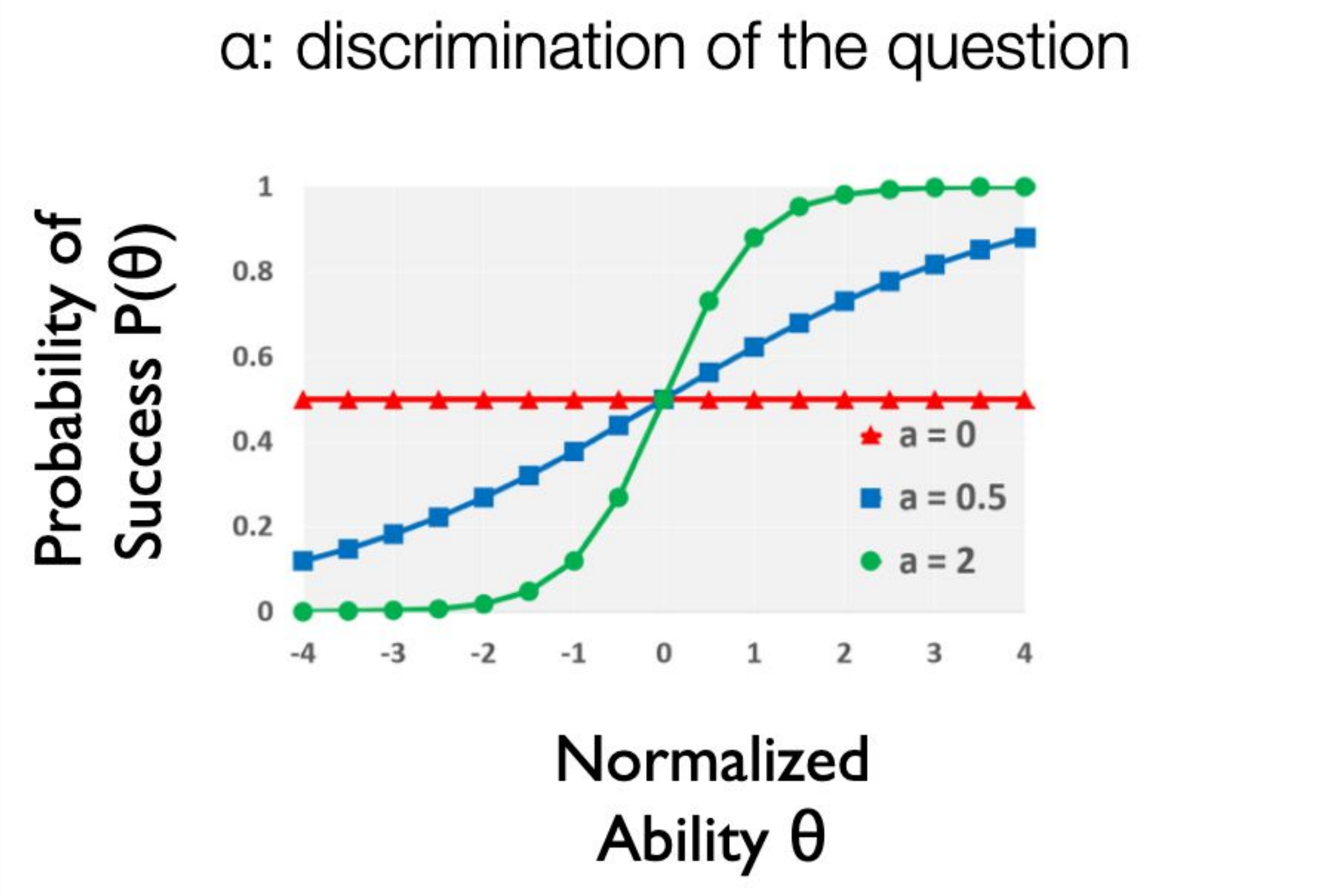
System Overview



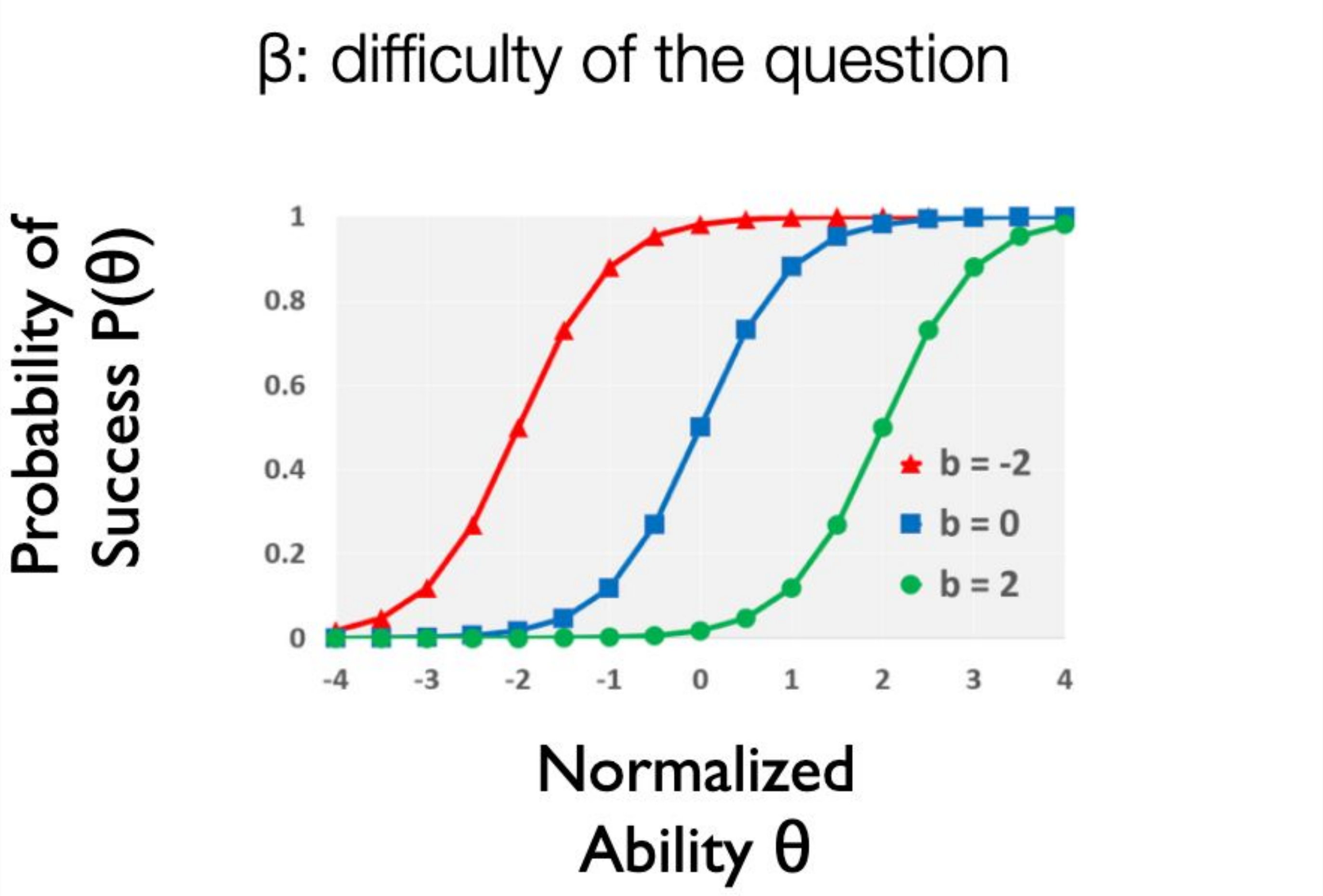
Item Response Theory

- Test takers have a single ability parameter θ
- Questions are modeled by **Item Characteristic Curve**:
  - α: discrimination of the question
  - β: difficulty of the question

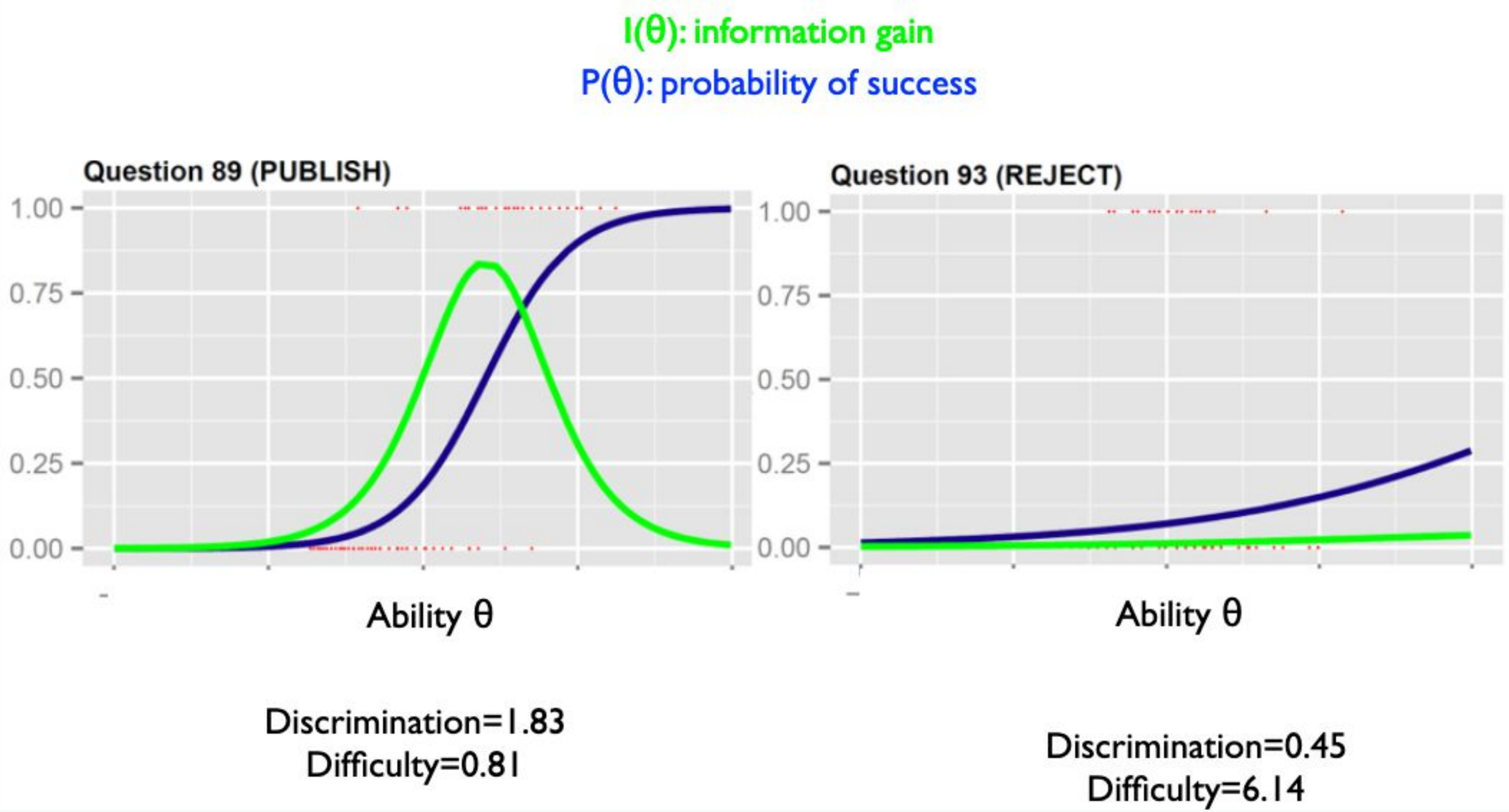
Item Response Theory



Item Response Theory



Question Evaluation Theory



Discrimination=1.83
Difficulty=0.81

Discrimination=0.45
Difficulty=6.14

Ability measures

Endogenous measures

- $\theta(u)$: Test score of candidate u
- Fit the function using logistic regression
- Derive discrimination and difficulty values for each question

Ability measures

Exogenous measures

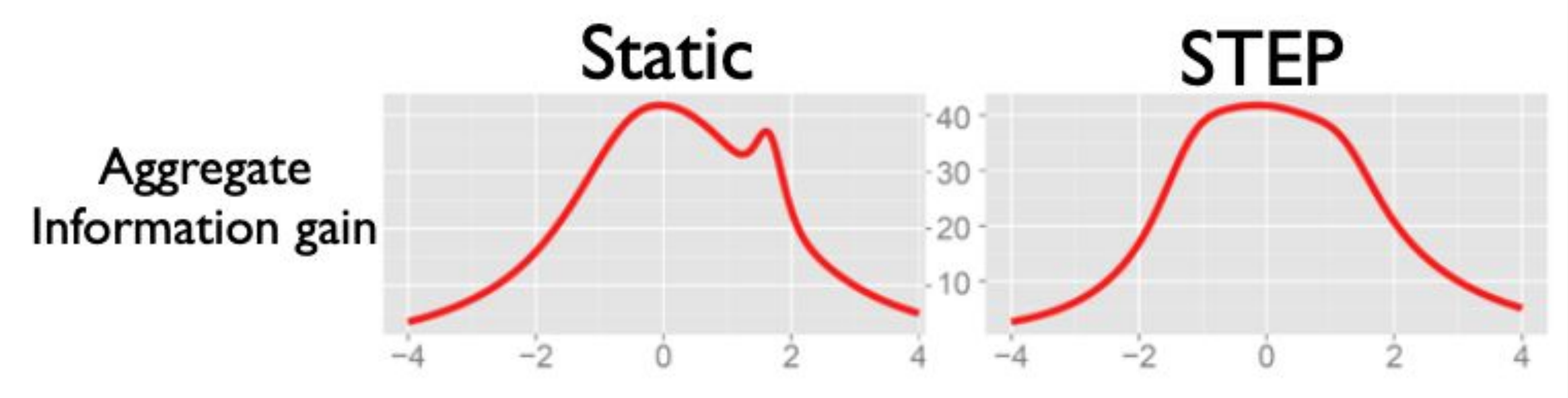
- $\theta(u)$: Hourly wage of candidate u after taking the test
- Use wage data from ODesk
- More robust to cheating
- Evaluates importance of skills in the marketplace

STEP cost

- Using oDesk data
- Question cost
  - Static question bank licensing: $10 per question
  - STEP: $4 per question
  - Create question “from scratch” (IKM data): $25 per question

STEP performance

- Question quality (Java test example)
- Static Question Bank: 87% acceptance rate
- STEP generated questions: 89% acceptance rate



STEP

- System that continuously generates new questions
- Makes tests more cheating-proof
- Assesses test quality with real-market performance data
- Identify potential errors or ambiguities
- Is of equal or higher quality with existing tests
- Cheaper to generate questions than licensing

What would the ability to find and engage experts allow you to do?