Bigger Than Any One
Solving Large Scale Data Problems with People and Machines

Tyler Bell: @twbell
"You just want to survive long enough to have a long run, to prove that your product has value. And in the short term, human solutions require much less work. Worry about scaling when you need to"
And now for something completely Factual
Discussing the Human Role as:

- Turkers (Sloggers)
- Experts (Thinkers)
- Contributors (Players)
Turkers (Sloggers)
A Good Body of Literature Exists:

- **CrowdDB** "harnessing Human Computation" to address the Closed World Assumptions of SQL

- **Crowdsourced Databases**: "Query Processing with People"

- **Deco**: Declarative Crowdsourcing "Human Computation"

- **Mob Data Sourcing**: a good state-of-the-art survey
Our Business Is People.
The Reasons Why Amazon Mechanical Turk No Longer Accepts International Turkers

As many of you may have heard, Amazon is not approving international Mturk accounts anymore. In 2012 Amazon put a hold on accepting new worker accounts for a couple of months. When they started accepting new workers again, it was only for workers in the United States.

There are three reasons for this:

**Fraud** - Amazon was having a difficult time verifying these accounts and a lot of fraud was occurring from outside the United States. Multiple accounts, sweatshops, money transfers and other Amazon TOS violations were occurring on these international accounts. Amazon would close the worker account and then they would simply open a new one. If you want to see some of this fraud, take a look at this forum- Romanian Mturk Forum. Over 300 pages long and 6000 posts. Almost every page of this forum has discussions on how to break Amazon’s TOS. It is no longer active forum because of the halts on new international accounts.

**Complaints** - Requester complaints about the quality of work is the second reason. I personally know quite a few workers outside of the US that are excellent turkers and speak and write English better than many Americans. Because of the constant stream of complaints to Amazon about the poor quality of work from international IP addresses, something had to be done. Since fraudulent workers outside of the US were not required to verify their accounts, they could just make a new account and get back to work under a different user ID.

**Blocks** - Blocks are a way Amazon to monitor worker accounts, although Amazon does not tell requesters this. When a worker account starts accumulating blocks from requesters, the workers account is suspended and reviewed by someone at Mturk. A requester could block a US worker and over time if that worker is not performing well, they are removed completely from the worker pool and cannot make a new account. With the international accounts, a worker would simply change IP address and make a new account and set up a click script to automatically cheat HITs until blocked.
"We’ve built a real-time human computation engine to help us identify search queries as soon as they’re trending, send these queries to real humans to be judged, and then incorporate the human annotations into our back-end models."

"we use a small custom pool of Mechanical Turk judges to ensure high quality."

"Clockwork Raven steps in to do what algorithms cannot: it sends your data analysis tasks to real people and gets fast, cheap and accurate results."
Gold Data → Search for Candidate Matches → Resolve Candidates → Compare Attributes → Likely Matches → Unsure → Likely Mismatches → Send to MTurk for Evaluation → Likely Matches → Likely Mismatches → Compute Accuracy Score → Accuracy Score
Factual Moderation Queue

~60% automatically accepted or rejected based on preset rules/filters

~30% can be parsed, processed by attribute, and validated by MTurkers

~10% handled by in-house moderators or 'Trusted Turkers'
Experts (Thinkers)
The Data Science Debate 2012

Peter Skomoroch (LinkedIn), Michael Driscoll (Metamarkets), DJ Patil (Greylock Partners), Toby Segaran (Google), Pete Warden (Jetpac), Amy Heineike (Quid)

"In data science, domain expertise is more important than machine learning skill"
"We've discovered that creative-data scientists can solve problems in every field better than experts in those fields can."

"The expertise you need is strategy expertise in answering these questions."

Interview with Jeremy Howard, Chief Scientist, Kaggle: [http://slate.me/UM0wQ7](http://slate.me/UM0wQ7)
"Data struggles with context. Human decisions are not discrete events. They are embedded in sequences and contexts. The human brain has evolved to account for this reality [...] Data analysis is pretty bad at narrative and emergent thinking, and it cannot match the explanatory suppleness of even a mediocre novel."

"Data inherently has all of the foibles of being human, [it] is not a magic force in society; it’s an extension of us."

NYT Article 24 Feb 2013: [http://nyti.ms/15cMV80](http://nyti.ms/15cMV80)
David Brooks NYT Op-Ed 18 Feb 2013: [http://nyti.ms/12EQmpQ](http://nyti.ms/12EQmpQ)
"People from different places and different backgrounds tend to produce different sorts of information. And so we risk ignoring a lot of important nuance if relying on big data as a social/economic/political mirror."

"Big data is undoubtedly useful for addressing and overcoming many important issues face by society. But we need to ensure that we aren't seduced by the promises of big data to render theory unnecessary."
"Without [cleansing and harmonisation] they’re not data markets; they’re data jumble sales."


Contributors (Players)
"Humans don’t tend to think in floating point pairs."

"It’s just dumb that a 100mil+ people carry GPS device in their pockets and we have to buy expensive proprietary data to find out about the shape of where we live."

"It is equal parts sad-making and hate-making that we're all still stuck suffering the lack of a comprehensive and open dataset for places."

"Seriously, if you’re presenting people geographic data, ask them sometime if you’re getting it right."

Aaron Cope: http://www.aaronland.info/weblog/2013/02/03/reality/#youarehere
Kellan Elliott-McCrea: http://laughingmeme.org/2013/02/03/are-you-here-a-feature-on-the-side/
Precision Market Insights offers exciting new measurement solutions for:

**Media Owners**

End the measurement gap with innovative insights. Validate the value of your inventory to increase occupancy rates and best match your assets to advertisers' campaigns. Now you can identify out-of-home traffic patterns and demographics, and even recognize the interests of your target audience to optimize current campaigns and improve planning. You can also measure campaign response via traffic changes to advertised locations or increased mobile site engagement. End the measurement gap with cutting-edge analytics and powerful insights.

More Benefits for Media Owners →

**Sponsors**

Make sponsorship investments confidently with unprecedented insights. What if you could identify what a venue audience was doing before, during or after a big event in order to up-sell and cross-sell to their interests? What if you could connect audiences to your brand and measure whether or not your sponsorship was driving increased traffic to a selected location or mobile site? Now you can use demographics, traffic patterns and even online browsing trends to confidently grow your business with unprecedented marketing metrics.

More Benefits for Sponsors →

**Advertisers**

Optimize out-of-home campaigns with new insights. Maximize media and validate the reach and impact of your campaigns. Based on innovative insights into audience interests, mobile browsing habits, travel patterns, detailed demographics and more, you can easily measure the effectiveness of outdoor ad units and the value of your vendor contracts. Optimize out-of-home and end the measurement gap with cutting-edge analytics from Precision Market Insights.

More Benefits for Advertisers →

**Retail**

Gain definitive insight into your retail-location dynamics. Know which consumer groups pass by a brick-and-mortar location by daypart and how effectively your outdoor campaign is driving traffic to your locations or mobile sites. Now you can better understand engagement trends. You can even use advanced metrics to determine the optimal locations for your retail stores. Our solutions help you improve decision making based on customers' shopping habits, interests, travel patterns and mobile browsing trends—so you can grow your business with confidence and a stronger MROI.

More Benefits for Retailers →

**Venue Owners**

What if you knew where your audience was going before, during and after a big event? What if you could know the customer profiles of not just your ticket purchasers, but the demographics, geographics and mobile Web-browsing interests of your actual attendees as well? Now you can take event analytics further than ever. Enhance your audience experience and drive sales with unprecedented marketing metrics that will keep your fans and sponsors coming back for more.

More Benefits for Venue Owners →
"So far InfoArmy has paid $146K to researchers (averaging $23/report), and some researchers have made over $5,000 individually. We are very happy that many hard working researchers have made money. However, without a clear path to revenue this model is unsustainable."

- Single report owners
- High prices
- Low quality
- Unpopularity of updates
- Report Abandonment
- Sold 44 reports total
Apple Maps' Report a Problem

40 million users

65,000 Waze users made 500 million edits.

Users fixed 70% of system-detected map problems and 100% of all user-reported map problems.

Almost all user-reported map problems were taken care of within a week.

Users added more than 50,000 gas stations on the map in the first month.

Users spend 7 hours a month on the platform.

Factual provides both 'Flag' and 'Write' interfaces

(this is good)
With No Means to Follow-Up

Thank you for your submission.

The response from your submission was:

```
{
  "factual_id": "7b579990-1e18-012f-0643-0030487f54d8",
  "new_entity": false,
  "commit_id": "H4sIAAAAAAAAAb3Nyb2xvcmxF0Y08erj1fUqgAAwwEzlsInAYuoHxy92cuQUHX73G8n2SzW0SgcLIXcF0FKq0AEk2uux0TJKkXtay67L0nK02Wy9ZdsNd_qc5Xv-L8IdEzf18FrdPPR-wBndZ-hYdMNSPUGhpBpV-11eJ0kAAAAA"
}
```

Additional details from your submission:

```
{
  "timestamp": "2013-02-27T16:24:50+00:00",
  "view_alias": "places",
  "factual_id": "7b579990-1e18-012f-0643-0030487f54d8",
  "user": "Tyler",
  "reference": "",
  "values": {
    "status": "0"
  }
}
```

(this is bad. But we're on it)
Now: extend this thinking to organizations
How Waze’s crowd-sourced data helped FEMA deliver the gas after Sandy

by Ryan Kim  NOV. 6, 2012 - 11:31 AM PST

3 Comments  tweet this

SUMMARY: With gas shortages rampant following Hurricane Sandy, FEMA and the White House turned to crowd-sourced navigation app Waze to gather data on where to send gasoline fueling trucks. The episode showed how mobile crowd-sourced data and tools like Waze can be helpful in a crisis.

photo: Twitter users @mattvas
Play ball with the competitors: "Think instead about collaborating with all the other players and getting the network effects of a larger audience and a bigger pool of offers. [...] The individual companies behind the consortium still compete to acquire subscribers, but they now collaborate with a common platform for monetizing those subscribers through marketing."

Bringing mobile commerce to life.

Weve has been formed to solve a critical problem for every business wanting to get involved in mobile commerce. Today, any organisation trying to engage with an increasingly mobile customer base is forced to engage with a bewildering mix of operators, services, platforms, channels and technologies – thus duplicating effort, increasing costs and making it impossible to get accurate and consistent insight into customer behaviour.
"Seeds [...] are one of the original information storage devices, it’s almost hard to understand why libraries haven’t always included seeds"
Data In game theory and economic theory, a zero–sum game is a mathematical representation of a situation in which a participant's gain (or loss) of utility is exactly balanced by the losses (or gains) of the utility of the other participant(s). If the total gains of the participants are added up, and the total losses are subtracted, they will sum to zero. Thus cutting a cake, where taking a larger piece reduces the amount of cake available for others, is a zero–sum game if all participants value each unit of cake equally (see marginal utility). In contrast, non-zero–sum describes a situation in which the interacting parties' aggregate gains and losses are either less than or more than zero. A zero–sum game is also called a strictly competitive game while non-zero–sum games can be either competitive or non-competitive. Zero–sum games are most often solved with the minimax theorem which is closely related to linear programming duality or with Nash equilibrium.
Tyler Bell
Factual
@twbell
tyler@factual.com