

---

**CSE 519: Data Science**  
**Steven Skiena**  
**Stony Brook University**

---

Lecture 24: Human-centric Data  
Science

---

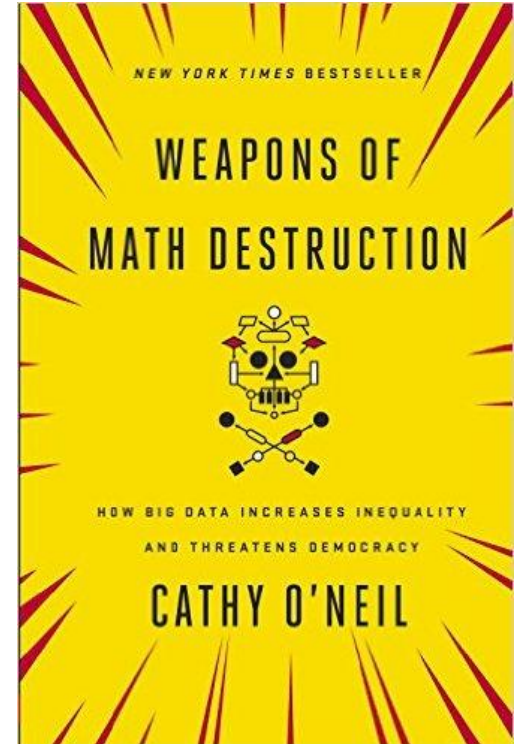
# Weapons of Math Destruction

---

It is important to understand the potential harm data-driven models can cause.

Correlation is not causation, but models so-trained can trigger actions and feedback mechanisms, resulting in self-fulfilling prophecies.

It is important for us as data scientists to think about societal issues in a constructive way.



# Properties of a Good Model

---

- It uses relevant data, not just data that happens to be available.
  - It is transparent, making clear why it is making its decisions.
  - There is a clear measure of success, and an embedded feedback mechanism to evaluate and learn from it.
-

# Teacher Ratings: A Terrible Model

---

School systems are firing teachers based on the test-scores of their students:

- **Relevant or available data?**
  - **Transparent?** Enough to see the statistical significance issues from small samples.
  - **Feedback/evaluation?** Are you firing bad teachers or those with bad students?
-

# Societal / Ethical Issues in Big Data

---

- Integrity in communications and modeling
  - Transparency and data ownership
  - Model-driven bias and filters
  - Maintaining the security of large data sets
  - Maintaining privacy in aggregated data
-

# Integrity in Communications and Modeling

---

Fight the temptation to inflate your results, because you know the model and audience:

- Present p-values, not just correlations.
  - Don't cherry-pick to present the best results.
  - Use visualization to reveal, not conceal.
  - Make clear the limitations of your models to non-technical audiences.
-

# Transparency and Ownership

---

- Does your organization follow data retention/use policies?
  - To what extent do users own the data they have generated?
  - Can data errors propagate? Are there mechanism for correction, like credit data?
  - Is data provenance maintained?
-

# Model-Driven Bias and Filters

---

Learning algorithms pick up biases from biased training data:

- Will your search engine show better job opportunities to men than women?
  - Are predatory ads shown to poor people?
  - Do news filters reinforce political polarization?
-



# Maintaining Security of Big Data

---

There are ethical responsibilities to encrypt and delete data to avoid security breaches:

- Making 100 million people change their password costs 190 man-years of effort.
  - Released data on addresses, ID numbers, and accounts persist for many years/forever.
  - Can anyone get/keep life insurance?
-

# Maintaining Privacy in Aggregate

---

People are often identifiable even if their names, address, and ID number are deleted.

- The AOL search engine data release.
- Destinations/tipping habits of celebrities, by correlating GPS locations of trips with paparazzi photos.

**Data scientists must strive to be responsible.**

---

# Aggregating Expertise

---

No person/program has all the answers.

Social media and other new technologies have made it easier to collect and aggregate opinions on a massive scale.

But how can we separate the wisdom of crowds from the cry of the rabble?

---

# Francis Galton's Ox

---

At a livestock fair in 1906, villagers were invited to guess the weight of a ox.

Galton observed that none of the almost 800 observers guessed the correct weight (1,178 pounds).

Yet the average guess was amazingly close: 1,179 pounds!

---

# Wisdom of Crowds: Penny Demo

---

How many pennies do I have in this jar?



[Watch the video!](#)

# Wisdom of Crowds: Rules of Demo

---

How many pennies do I have in this jar?

Ten of you write your opinions on cards

Ten more tell me them one by one.

Who wants to bet?

---

# Independent Guesses

---

537, 556, 600, 636, 1200, 1250, 2350, 3000,  
5000, 11000, 15000

Median: 1250, Mean: 3739, Actual: 1879

The median is closer than any other guess.

---

# Conditioned Guesses

---

A second group of students made guesses after seeing the first group's answers:

750, 750, 1000, 1000, 1000, 1250, 1400, 1770, 1800, 3500, 4000, 5000

Median: 1325, Mean: 1935, Actual: 1879

Showing other people's guesses substantially conditioned the distribution (no outliers)

---



# Financial Stakes

---

Allowing people to bet on the outcome yielded two wagers:

1500, 2000 **Actual: 1879**

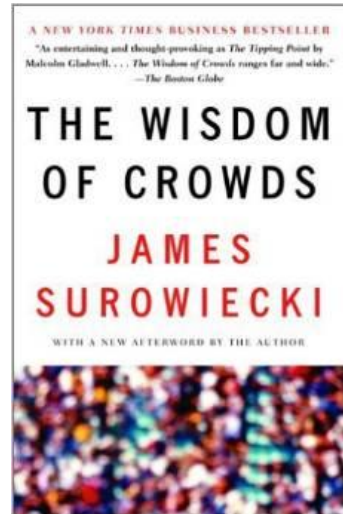
People willing to bet their own money on an event are by definition more confident in their selection.

---

# When is the Crowd Wise?

---

- When the opinions are independent: this avoids groupthink.
- When the crowd consists of people with diverse knowledge / methods.
- When the problem is in a domain where the crowd does not need specialized knowledge.
- Opinions can be fairly aggregated.



# Diversities of Opinion

---

Crowds only add information when there is disagreement.

A committee with perfectly correlated experts contributes nothing more than any one of them.

Be an incomparable element on the partial order of life.

---

# Mechanisms for Aggregation

---

On numerical aggregation, the mean or median work if errors are symmetrically distributed.

But which is more robust: mean or median?

On classification problems, voting is the basic aggregation mechanism.

But should all votes be treated equally?

---

# The Condorcet Jury Theorem

---

If the probability of each voter being correct is  $p > 0.5$ , the probability of a majority of voters being correct  $P(n)$  is  $> p$ .

$$P_n = \sum_{h=(n+1)/2}^n \left[ \frac{n!}{h!(n-h)!} \right] p^h (1-p)^{n-h}$$

For  $p=0.51$ , a jury of 101 members is right 57%,  
 $P(1001)=0.73$  and  $P(10001)=0.9999$ .

As  $n$  gets large enough  $P(n)$  approaches 1.

---

# Arrow's Impossibility Theorem

---

No election system for summing permutations of preferences as votes satisfies:

- Each voter is allowed any ranking they want.
  - The relative order of any subset of candidates is independent of all others.
  - If all voters like  $a > b$ ,  $a$  wins ahead of  $b$ .
  - No single voter can dictate rankings.
-

# Crowdsourcing Services

---

Services like Amazon Turk and Crowdfunder provide the opportunity to hire large numbers of people for small amounts of piece work.

In life, you generally get what you pay for.

Getting people to do your bidding requires both incentives and clear instructions.

---

Your Account

HITS

Qualifications

442,955 HITS  
 available now

All HITS | **HITS Available To You** | HITS Assigned To You

Find  containing

that pay at least \$

for which you are qualified

require Master Qualification



**All HITS**

1-10 of 3194 Results

Sort by:

[Show all details](#) | [Hide all details](#)

1 2 3 4 5 > [Next](#) >> [Last](#)

Is this Category Valid for the Groupon Deal?		<a href="#">View a HIT in this group</a>	
<b>Requester:</b> <a href="#">Relevance Groupon</a>	<b>HIT Expiration Date:</b> Jul 9, 2014 (6 days 4 hours)	<b>Reward:</b> \$0.04	
	<b>Time Allotted:</b> 5 minutes	<b>HITS Available:</b> 78236	
Mark all the cells in the 10 images		<a href="#">View a HIT in this group</a>	
<b>Requester:</b> <a href="#">Cell Counting</a>	<b>HIT Expiration Date:</b> Jul 6, 2014 (3 days 9 hours)	<b>Reward:</b> \$0.01	
	<b>Time Allotted:</b> 25 minutes	<b>HITS Available:</b> 33091	
Extract purchased items from a shopping receipt		<a href="#">View a HIT in this group</a>	
<b>Requester:</b> <a href="#">Jon Brelig</a>	<b>HIT Expiration Date:</b> Jul 10, 2014 (6 days 23 hours)	<b>Reward:</b> \$0.08	
	<b>Time Allotted:</b> 2 hours	<b>HITS Available:</b> 26691	
Geo Result Relevance-Sun Jun 08 08:20:10 PDT 2014		<a href="#">View a HIT in this group</a>	
<b>Requester:</b> <a href="#">Amazon Requester Inc.</a>	<b>HIT Expiration Date:</b> Jul 9, 2014 (5 days 14 hours)	<b>Reward:</b> \$0.00	
	<b>Time Allotted:</b> 60 minutes	<b>HITS Available:</b> 23052	
Geo Result Relevance-Fri Jun 27 10:50:25 PDT 2014		<a href="#">View a HIT in this group</a>	
<b>Requester:</b> <a href="#">Amazon Requester Inc.</a>	<b>HIT Expiration Date:</b> Jul 27, 2014 (3 weeks 3 days)	<b>Reward:</b> \$0.00	
	<b>Time Allotted:</b> 60 minutes	<b>HITS Available:</b> 16493	
Inv_B_2		<a href="#">View a HIT in this group</a>	
<b>Requester:</b> <a href="#">rohzi0d</a>	<b>HIT Expiration Date:</b> Jul 17, 2014 (2 weeks)	<b>Reward:</b> \$0.00	
	<b>Time Allotted:</b> 48 minutes	<b>HITS Available:</b> 11727	



# Good Uses of Turkers

---

- Measuring aspects of human perception (e.g. what name would you call this color?)
  - Language understanding tasks (read cans for the blind).
  - Obtaining training data for machine learning classifiers.
  - Creative efforts (blog posts/reviews or 10,000 sheep drawings, see [www.sheepmarket.com](http://www.sheepmarket.com))
  - Economic/psychological experiments
  - Reviewing transcripts for accuracy
-

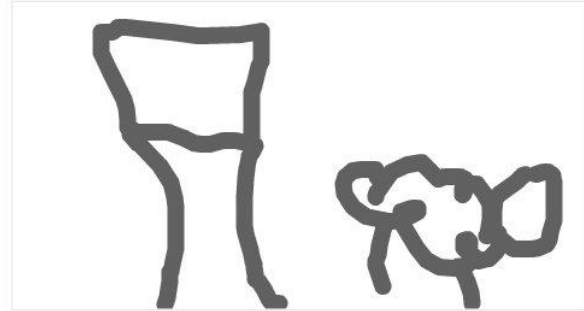
# Sheepmarket.com

---

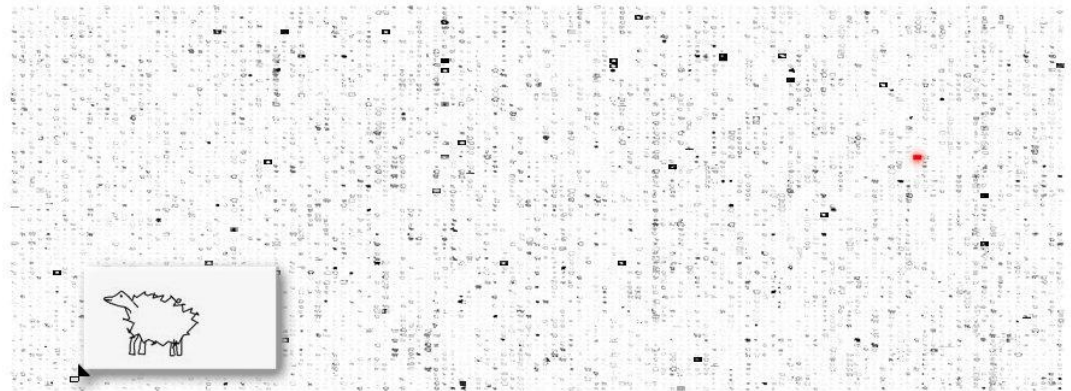
What creative endeavors can **you** think of that people will do for \$0.02 each?

THE SHEEP MARKET

10,000 sheep created  
by online workers.  
More...



638 / 10000



Aaron Koblin © 2006

# Bad Uses of Turkers

---

- Any task which requires advanced training.
  - Any task you cannot specify clearly.
  - Any task where you cannot perform some type of validation as to whether they are doing a good job or not.
  - Any illegal task, or one too inhuman to subject people to (IRB ethical standards)
-

# Applications for Turkers? (Projects)

---

- Miss Universe?
  - Movie gross?
  - Baby weight?
  - Art auction price?
  - Snow on Christmas?
  - Super Bowl / College Champion?
  - Ghoul Pool?
  - Future Gold / Oil Price?
-

# Gamification

---

Make things fun so that people work for free!

Motivational techniques include:

- leader boards
- points
- badges
- leader boards
- progress bars



# Games with a Purpose (GWAP)

---

- CAPTCHAs for OCR
- ESP game for labeling image components
- Psychological/IQ testing in games/apps
- FoldIt game for predicting protein structures

Keys to success include:

- Being playable enough to become popular
  - Abstracting technicality as scoring functions
-