Two Questions

1. Do you enjoy listening to K-Pop?
2. What percent of people in this room do you think enjoy listening to K-Pop?

There is no incentive to misreport what you truly believe to be your answers as well as others' answers. You receive higher payoff if you submit answers that are more surprisingly common than collectively predicted.

Bayesian Truth Serum

A Bayesian Truth Serum for Subjective Data (Prelec 2004)

An Algorithm That Finds Truth Even If Most People Are Wrong (Prelec & Seung 2010)

A Bayesian Truth Serum For Subjective Data

Prelec 2004

Motivation

- **Subjective data**
  - No practical truth or omniscient grader
  - No ultimate outcome that can be observed
- Examples: behavior/intention/opinion
  - Environmental risk analysis, voting behavior surveys, product/service feedback
- Why might respondents not be truthful?
  - Social acceptability of answer

Goal

How can we **incentivize agents to report truthfully** when there is no defined truth or outcome?

> *When "**objective truth is intrinsically or practically unknowable**"?*

Related Work

- Methods that privilege the consensus answer
  - Simple majority voting
  - Delphi method
- Peer prediction

Related Work

- Methods that privilege the consensus answer
  - Simple majority voting
  - Delphi method
- Peer prediction
  - Assumes that the mechanism designer knows the prior!

With BTS, we will eliminate the assumption that we know the prior

BTS, Informally

1. Each respondent must provide:
   a. Personal opinion
   b. Estimated distribution of opinions in population
2. Reward responses that are "**surprisingly common**"

Intuition

- Why does this reward truthfulness?
- **Bayesian updating argument**: Individuals with a certain opinion report a higher frequency of that opinion in the population
  - **Why**? "Informative sample of one"
  - **Corollary**: One expects that the population will underestimate true frequency of one’s own opinion
  - **Therefore**: One's truthful opinion has the best chance of being "surprisingly common"

Formal Model

r-indexed respondents
m opinions/responses (in multiple choice question)
$t^r = (t^r_1 \dots t^r_m)$ indicator variable for truthful opinion
$\omega = (\omega_1 \dots \omega_m)$ distribution of frequencies over population
$x^r = (x^r_1 \dots x^r_m)$ indicator variable for personal answer
$y^r = (y^r_1 \dots y^r_m)$ predicted distribution of frequencies

Scoring Rule

For each *k*, calculate frequency of endorsement and geometric mean of predicted frequencies:

$$\overline{x}_k = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} x_k^r$$

$$\log \overline{y}_k = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \log y_k^r$$

Scoring Rule

$$u^r = \sum_k x_k^r \log \frac{\overline{x}_k}{\overline{y}_k} + \alpha \sum_k \overline{x}_k \log \frac{y_k^r}{\overline{x}_k}$$

Scoring Rule

$$u^r = \sum_k x_k^r \log \frac{\overline{x}_k}{\overline{y}_k} + \alpha \sum_k \overline{x}_k \log \frac{y_k^r}{\overline{x}_k}$$

information score

Scoring Rule

$$u^r = \sum_k x_k^r \log \frac{\overline{x}_k}{\overline{y}_k} + \alpha \sum_k \overline{x}_k \log \frac{y_k^r}{\overline{x}_k}$$

information score + prediction score

Assumptions

- Large or countably infinite *n**
- Rational Bayesians
- A1: Common prior**
- A2: Exchangeable prior / conditional independence
- A3: Stochastic relevance

Scoring rule & theorems are for **countably infinite case*

***The mechanism designer won't need to know beforehand what the prior distribution is!*

"Impersonally informative"

- A2: Exchangeable prior

$$p(t^r, \ldots, t^s) = p(t^{\pi(r)}, \ldots, t^{\pi(s)})$$

$$t^r = t^s \implies p(\omega|t^r) = p(\omega|t^s)$$

- Same opinion implies same posterior belief
- Conditional independence

"Impersonally informative"

- A3: Stochastic relevance
  - A form of dependence
  - Reverse implication of previous assumption: different opinions imply different posteriors, or equivalently,

$$p(\omega|t^r) = p(\omega|t^s) \implies t^r = t^s$$

"Impersonally informative"

- Together:

$$t^r = t^s \iff p(\omega|t^r) = p(\omega|t^s)$$

- "Respondents believe that others sharing their opinion will draw the same inference about population frequencies"
- Why is this important?

Results

- T1: Collective truthtelling is a strict BNE for any alpha > 0.
- T2: Expected equilibrium information scores in any BNE are
  - (a) nonnegative,
  - (b) at a weak maximum for all respondents in truth-telling equilbrium.
- T3: Zero-sum game when alpha = 1.

In practice?

"In actual applications of the method, one **would not teach respondents the mathematics of scoring** or explain the notion of equilibrium. Rather, one would like to be able to tell them that truthful answers will maximize their expected scores, and that in arriving at their personal true answer they are free to ignore what other respondents might say."

In practice?

“There is no incentive to misreport what you truly believe to be your answers as well as others’ answers. You will have a higher probability of winning a lottery (bonus payment) if you submit answers that are more surprisingly common than collectively predicted."

Designing Incentives for Inexpert Human Raters, Shaw et al. 2011

In practice?

“There is no incentive to misreport what you truly believe to be your answers as well as others’ answers. You will have a higher probability of winning a lottery (bonus payment) if you submit answers that are more surprisingly common than collectively predicted."

"confusion and cognitive demand"

Designing Incentives for Inexpert Human Raters, Shaw et al. 2011

In practice?

- Creating Truthtelling Incentives with the Bayesian Truth Serum [DW '08]
  - Claiming awareness of "foils" reduced when scoring with BTS
  - Description: "BTS scoring rewards you for answering honestly. Even though there is no way for anyone to know if your answers are truthful — they're your personal opinions and beliefs — your score will be higher on average if you tell the truth."

Advantages & Limitations

- Limitations
  - Doesn't work for small *n*
  - Cumbersome for large *m*
  - When might certain assumptions not hold?

- Advantages
  - No incentive to bias answers towards the expected group consensus answer
  - Not easy to circumvent by collective collusion
  - Can be applied to previously unasked questions: **we don’t need to know the prior**

Intermission

Mini-experiment results!

An algorithm that finds truth even if most people are wrong

Prelec and Seung 2010

Goal

BTS - Incentivize truthfulness

This paper - Find the truth

Challenge: When using BTS, everyone reports their belief, but not everyone is right. How to aggregate the truth?

Metaknowledge

How much an individual knows about their peers' responses

Metaknowledge is effective as truth diagnostic when information is unevenly distributed

BTS treats all respondents equally, regardless of the metaknowledge they display

Key insight

Weight each respondent's response by the metaknowledge that respondent displays.

- Metaknowledge is measured using BTS

Model

Each respondent is asked to endorse the most likely answer, and provide an predicted probability distribution over all possible answers

- We have a single question with *m* answers, indexed by *k*.
- We have *n* respondents, indexed by *r*.
- x_k^r indicates whether *r* has endorsed *k*
- $y = (y_1^r, ..., y_m^r)$ is *r*'s prediction of distribution of answers

Step 1

Calculate the average $\bar{x}_k$ of the endorsements and the geometric mean $\bar{y}_k$ of the predictions:

$$\bar{x}_k = \frac{1}{n}\sum_{r=1}^{n} x_k^r, \quad \log \bar{y}_k = \frac{1}{n}\sum_{r=1}^{n} \log y_k^r$$

Step 2

Calculate the BTS score of each individual *r*:

$$u^r = \sum_{k=1}^{m} x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_{k=1}^{m} \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}$$

Step 3

For each answer *k*, calculate the average BTS score $\overline{u}_k$ of all individuals endorsing answer *k*:

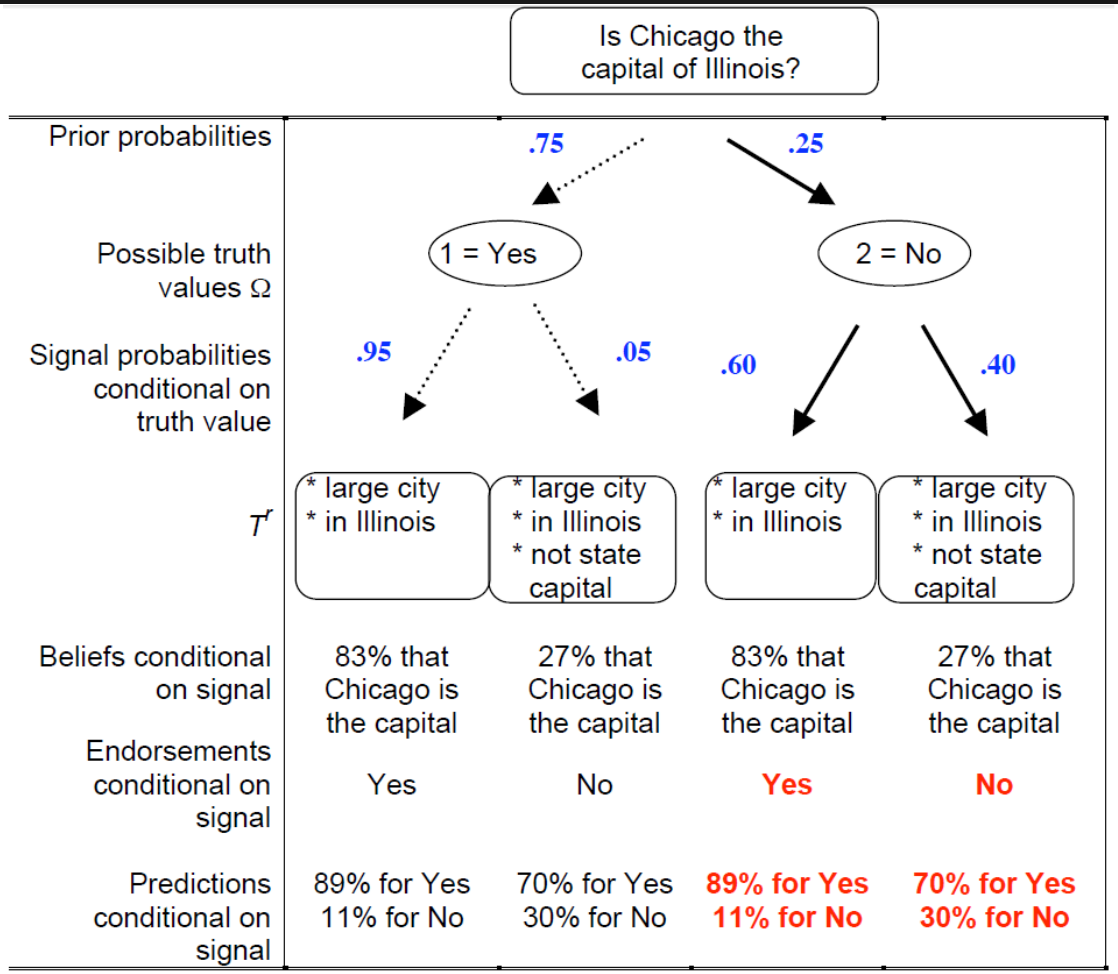
$$\overline{u}_k = \frac{1}{n\ \overline{x}_k} \sum_{r=1}^{n} x_k^r u^r$$

Step 4

Select the answer *k* that maximizes $\overline{u}_k$.

In other words, choose the answer whose endorsers display the most metaknowledge on average.

Example


Is Chicago the capital of Illinois?
Prior probabilities
.75
.25
Possible truth values Ω
1 = Yes
2 = No
Signal probabilities conditional on truth value
.95
.05
.60
.40
T'
* large city * in Illinois
* large city * in Illinois * not state capital
* large city * in Illinois
* large city * in Illinois * not state capital
Beliefs conditional on signal
83% that Chicago is the capital
27% that Chicago is the capital
83% that Chicago is the capital
27% that Chicago is the capital
Endorsements conditional on signal
Yes
No
Yes
No
Predictions conditional on signal
89% for Yes 11% for No
70% for Yes 30% for No
89% for Yes 11% for No
70% for Yes 30% for No
BTS score
+0.06
–1.05
–0.57
+0.86

Truth and belief

- Truth $\Omega = i$, drawn from probability distribution $P(\Omega = k)$
- Respondent *r* receives signal T^r, drawn from $S_{kj} = P(T^r = k \mid \Omega = j)$
- Belief matrix $B_{jk} = P(\Omega = j \mid T^r = k)$
- Metaknowledge matrix $M_{jk} = P(T^s = j \mid T^r = k)$

Assumptions

- Common prior known to all respondents (but not to us).
- $P(\Omega = k \mid T^r = k) > P(\Omega = j \mid T^r = k)$ for all $j <> k$
- $P(\Omega = i \mid T^r = i) > P(\Omega = i \mid T^r = j)$ for all $j <> i$
  - Truth Sensitivity

Connecting to the model

- Our common prior has nonsensical events
  - What is the probability that Chicago is the capital of Illinois, given that Chicago is the capital of Illinois?
- But we don't compute every combination
- "*r* endorses *k*" interpreted as $T^r = k$
- "*r* predicts *y*" interpreted as noisy report of column in metaknowledge matrix
- Thus we have full metaknowledge matrix and the single column in the signal matrix for the true answer.

Key result

$$\lim_{n \to \infty} \overline{u}_k = \log \Pr\left[\Omega = i \mid T^r = k\right] + \mathrm{C}$$

- Given a signal, we can order the conditional probability of each outcome
- This + Truth Sensitivity = algorithm for maximizing likelihood of correctness

Proof

$$\lim_{n\to\infty} \bar{x}_j = \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n} x_j^s = S_{ji} \qquad \text{(a)}$$

Endorsement rate -> signal probability

$$\lim_{n\to\infty} \frac{1}{n\bar{x}_j} \sum_{s=1}^{n} x_j^s \log y_k^s = \log M_{kj} \qquad \text{(b)}$$

Log prediction rate -> metadata

$$\bar{u}_j = \lim_{n\to\infty} \frac{1}{n\bar{x}_j} \sum_{s=1}^{n} x_j^s u^s, \qquad \text{(c)}$$

Average BTS score for j-endorsers

$$u^r = \sum_{k=1}^{m} x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} - \sum_{k=1}^{m} \bar{x}_k \log \frac{\bar{x}_k}{y_k^r} \qquad \text{(e)}$$

BTS score for a single respondent

$$\lim_{n\to\infty} \bar{u}_j = \log \Pr\left[\Omega = i \mid T^r = j\right] + \mathrm{C}, \qquad \text{(d)}$$

Take limit and average

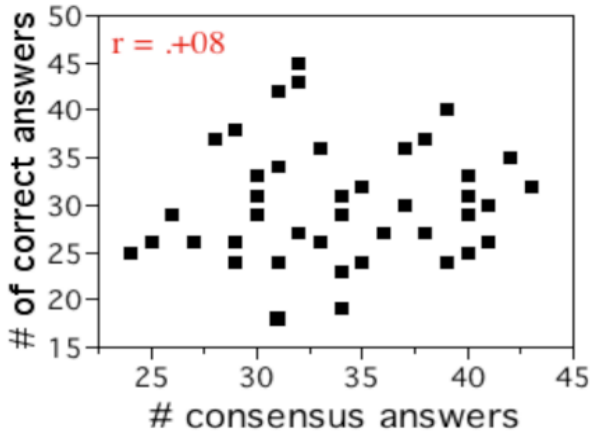
In Practice

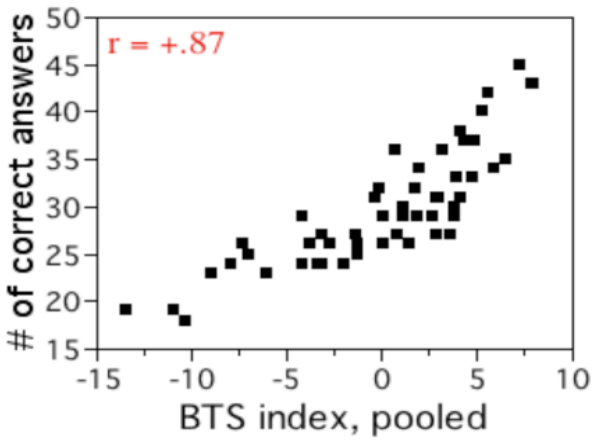
- "Is X the capital of Y?"
- Predictions made by respondents with correct answers were on average more accurate
- BTS vs majority voting: reduces # mistakes from 19 to 9 and from 12 to 6

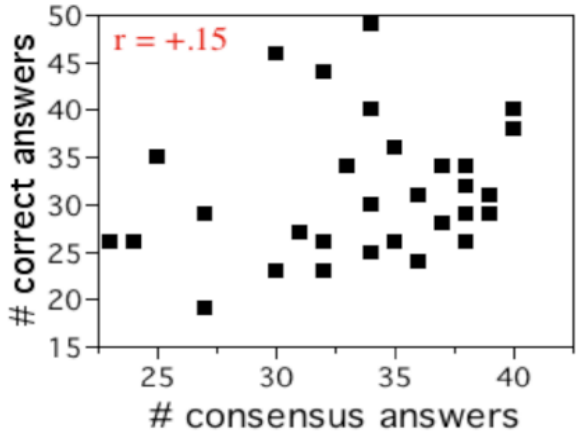
Finding experts

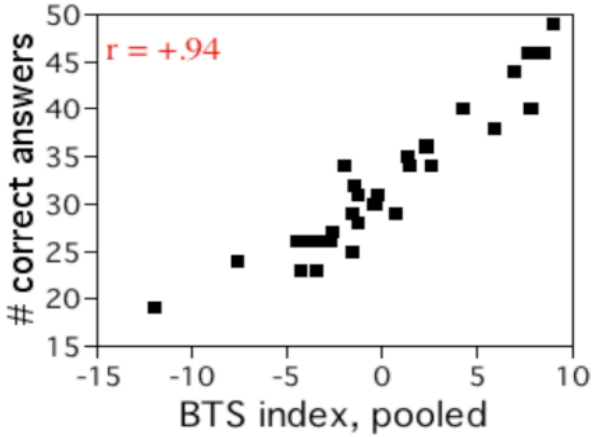
- If knowledge correlates among multiple questions, can identify experts
  - Individual Index - BTS score
  - Pooled Index - Average BTS of endorsed answers
- Conventional wisdom - how often one votes with majority

Finding experts



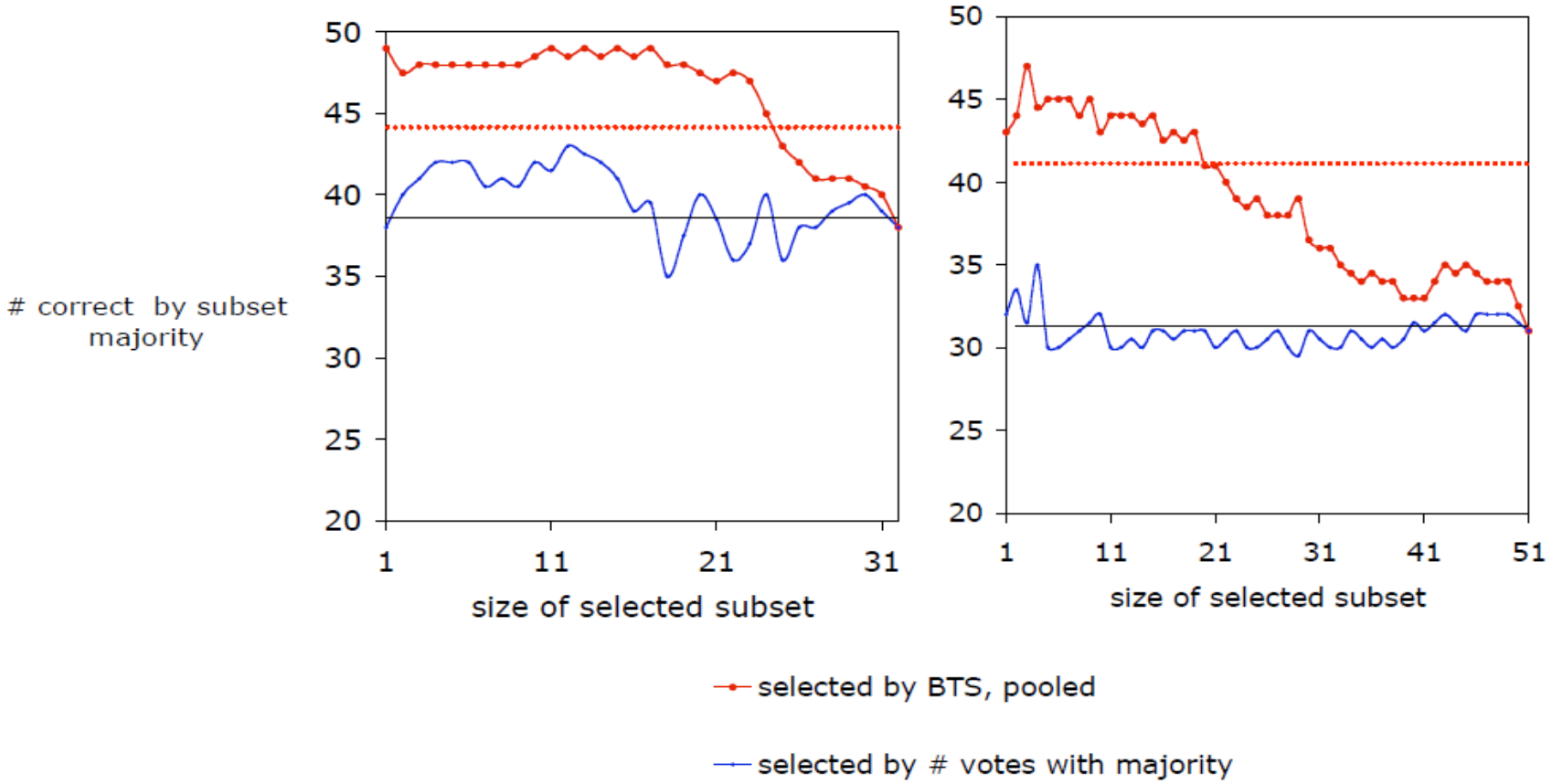






Hybrid Approach

Use majority voting for BTS-identified experts



Final Thoughts(?)

"enforces a meritocratic outcome by an open democratic process"