Crowdsourcing Visual Analysis

Maneesh Agrawala

CS 294-10: Visualization
Fall 2013

Last Time: Collaborative Visual Analysis

A graveyard of "YouTubes for Data"

swivel

verifiable.com

Track-n-Graph™

icharts™
PUBLIC BETA RELEASE

Chartle.net
simple & interactive charts online

Many Eyes – circa 4/2012

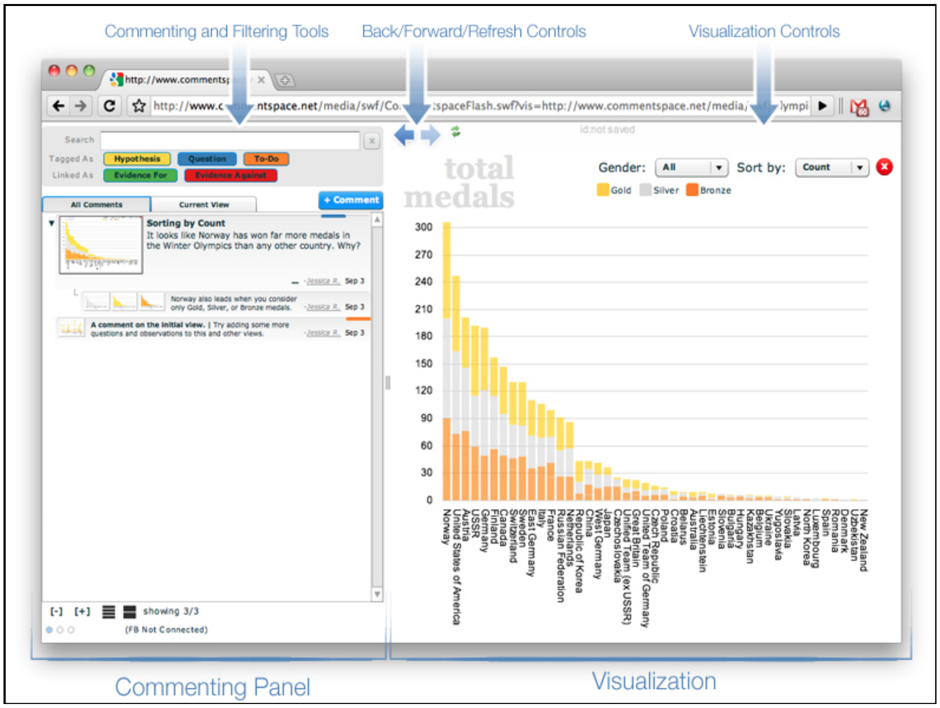
128,478 Visualizations

17,340 Comments

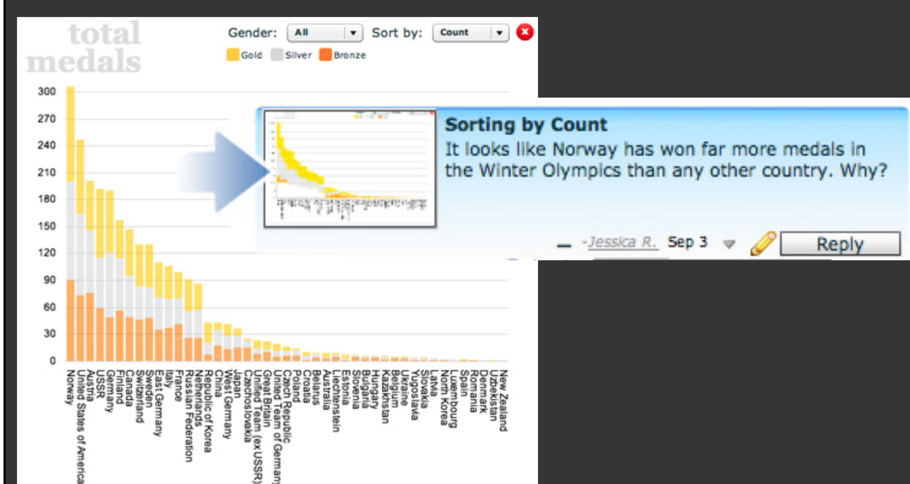
only ~11% of comments provided a plausible hypothesis or explanation for the data in the chart

CommentSpace: Structured Support for Social Data Analysis

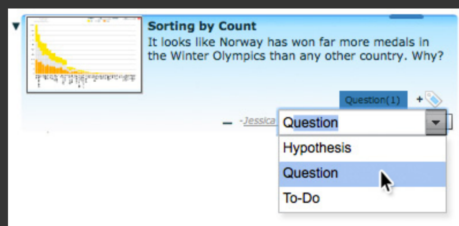
Can we augment social data analysis to support deeper analysis and synthesis?



Tight coupling of comments & views



Tags and links for organization

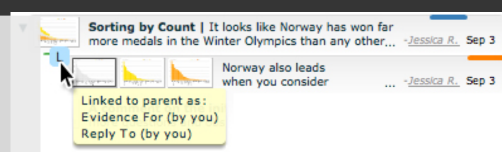


Tags

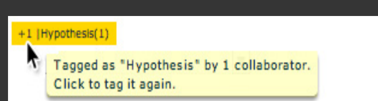
- Hypothesis
- Question
- To-Do

Links

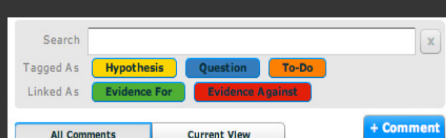
- Evidence For
- Evidence Against
- (Related)



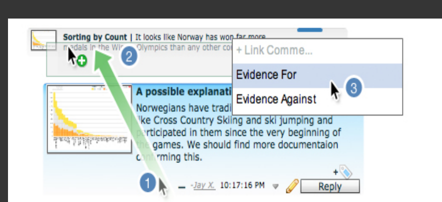
Hypothesis generation/evidence gathering



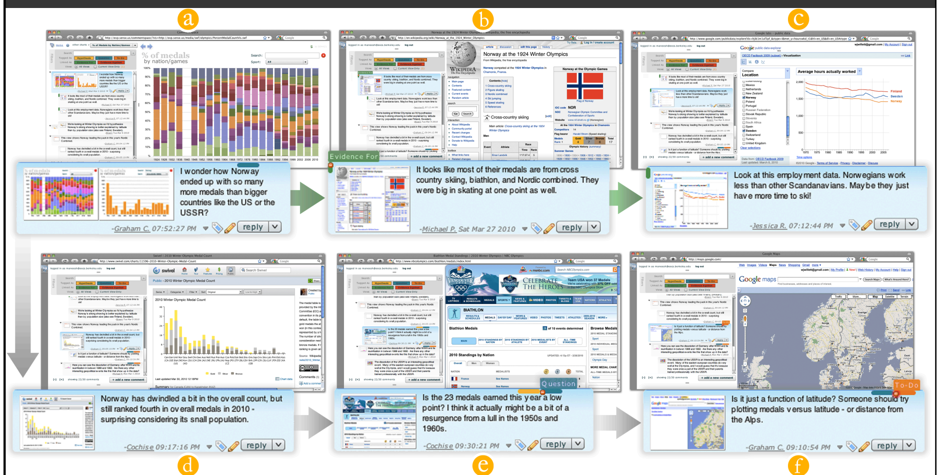
Contributors highlight important items with tags



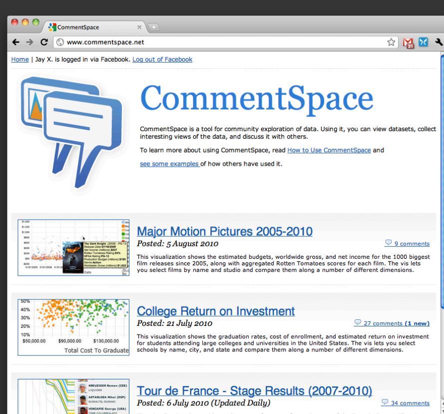
Tags help late-joiners identify and build on important comments



Contributors use links to organize contributions and build narrative



Studies and Deployments



Controlled lab studies to test core analysis subtasks

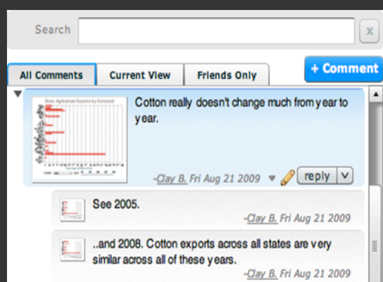
Live deployments (www.commentspace.net)

Study: Use of Tags and Links

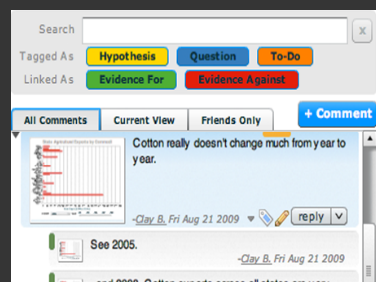
Hypothesis: Tags and links can provide common ground and encourage continued discussion.

A between subjects study (n=24) with 2 conditions.

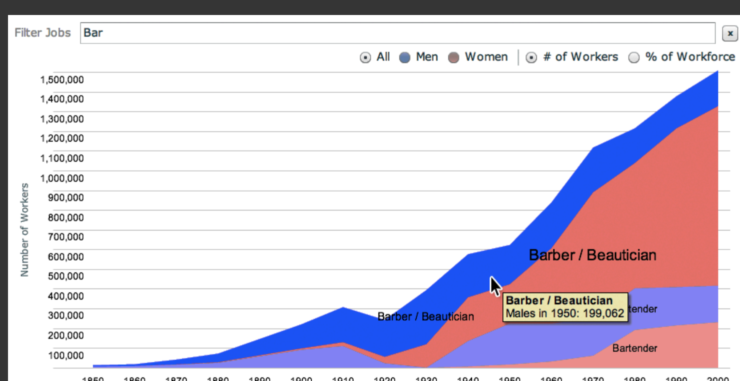
"No-Tag" Condition



"Tag" Condition



Study: Prompt



Hypothesis: Stereotypically male jobs have remained almost entirely male even as women have joined the work force.

Study: Results

Participants who used tags and links classified comments more consistently and accurately than those who didn't

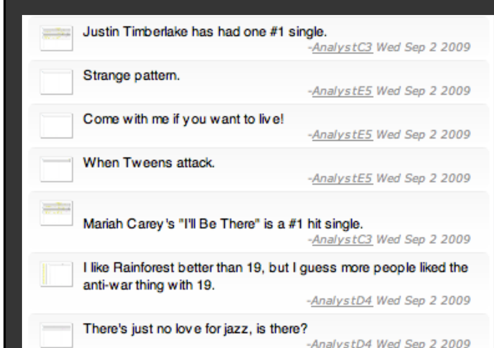
(greater in-group agreement) (greater agreement with experts)

Participants using tags and links generated significantly more replies to existing comments

Tag (Median=7)

No-Tag (Median=2)

But ….



In open-ended tasks, participants still engaged mostly in superficial, exploratory analysis

We saw very little use of tags or links

Can we use paid crowds to perform small pieces of analysis tasks?

Announcements

Final project

Design new visualization method

- Pose problem, Implement creative solution

Deliverables

- Implementation of solution
- 8-12 page paper in format of conference paper submission
- 1 or 2 design discussion presentations

Schedule

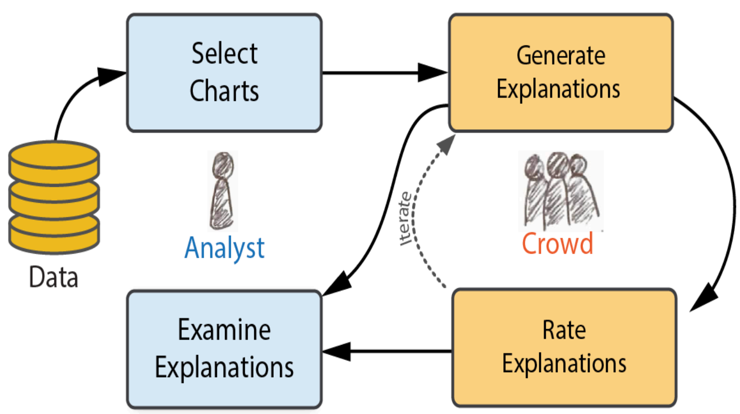
- Project proposal: 10/28
- Project presentation: 11/13-11/20
- Final paper and presentation: 12/2-12/6

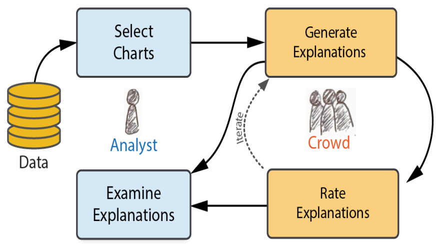
Grading

- Groups of up to 3 people, graded individually
- Clearly report responsibilities of each member

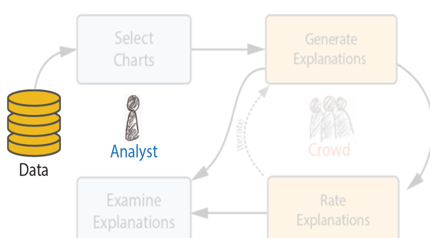
Crowdsourcing Visual Analysis

Can we use paid crowds to perform small pieces of analysis tasks?



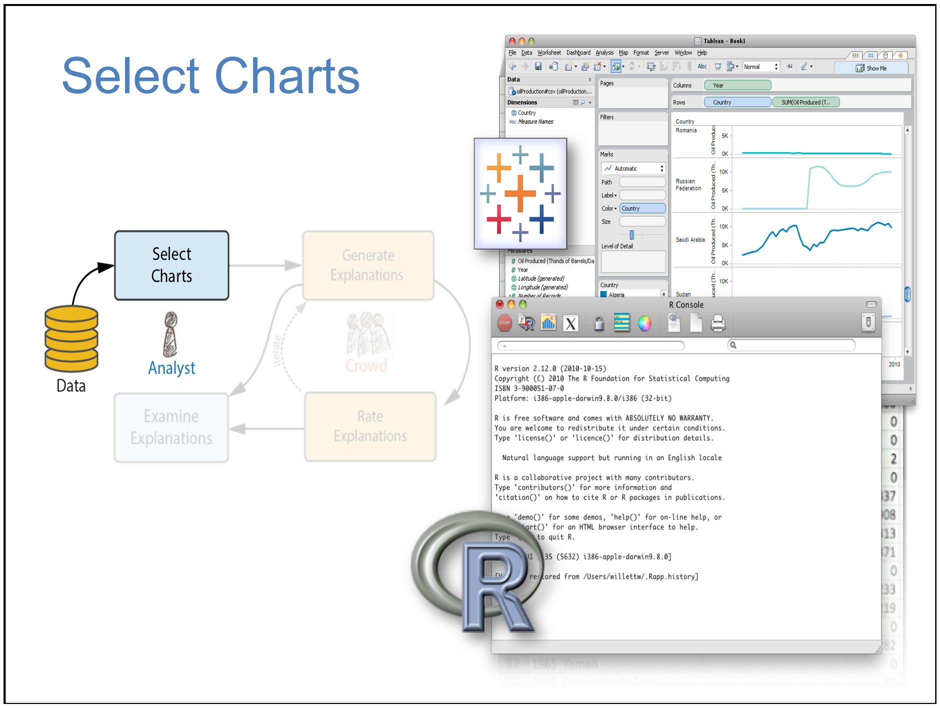


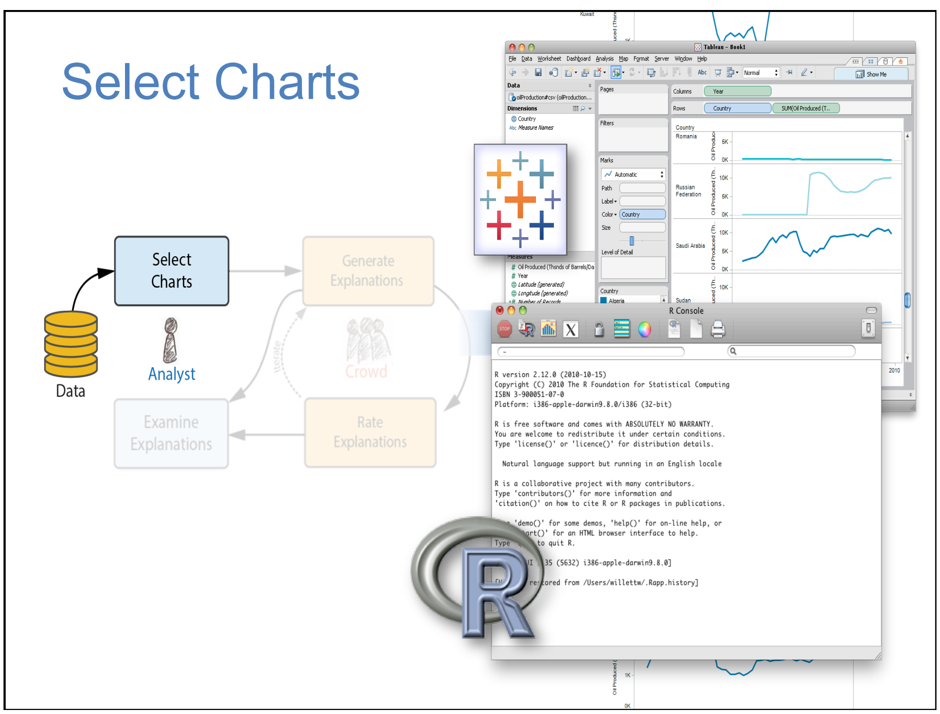
Data



World Oil Production 1965-2010

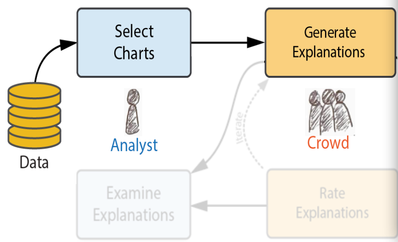
1	Year	Country	Oil Produced (Thsnds. of Barrels/Day)
2	1965	United States	9014
3	1965	Canada	920
4	1965	Mexico	362
5	1965	Argentina	276
6	1965	Brazil	96
7	1965	Colombia	203
8	1965	Ecuador	8
9	1965	Peru	66
10	1965	Trinidad & Tobago	135
11	1965	Venezuela	3503
12	1965	Other S. & Cent. America	47
13	1965	Azerbaijan	0
14	1965	Denmark	0
15	1965	Italy	48
16	1965	Kazakhstan	0
17	1965	Norway	0
18	1965	Romania	266
19	1965	Russian Federation	0
20	1965	Turkmenistan	0
21	1965	United Kingdom	2
22	1965	Uzbekistan	0
23	1965	Other Europe & Eurasia	5337
24	1965	Iran	1908
25	1965	Iraq	1313
26	1965	Kuwait	2371
27	1965	Oman	0
28	1965	Qatar	233
29	1965	Saudi Arabia	2219
30	1965	Syria	0
31	1965	United Arab Emirates	282
32	1965	Yemen	0

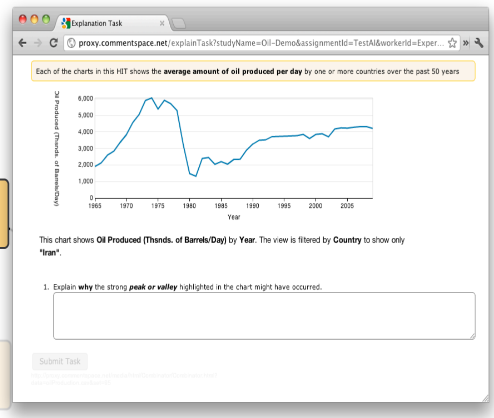

Select Charts
Select
Charts
Generate
Explanations
Analyst
Crowd
Data
Examine
Explanations
Rate
Explanations
R Console

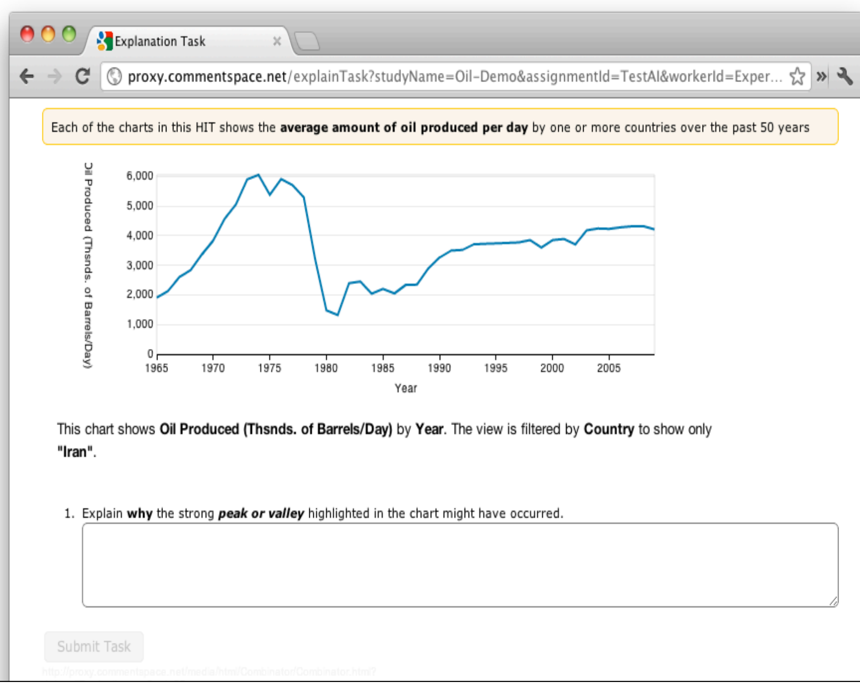

Select Charts
Select
Charts
Generate
Explanations
Analyst
Crowd
Data
Examine
Explanations
Rate
Explanations
R Console

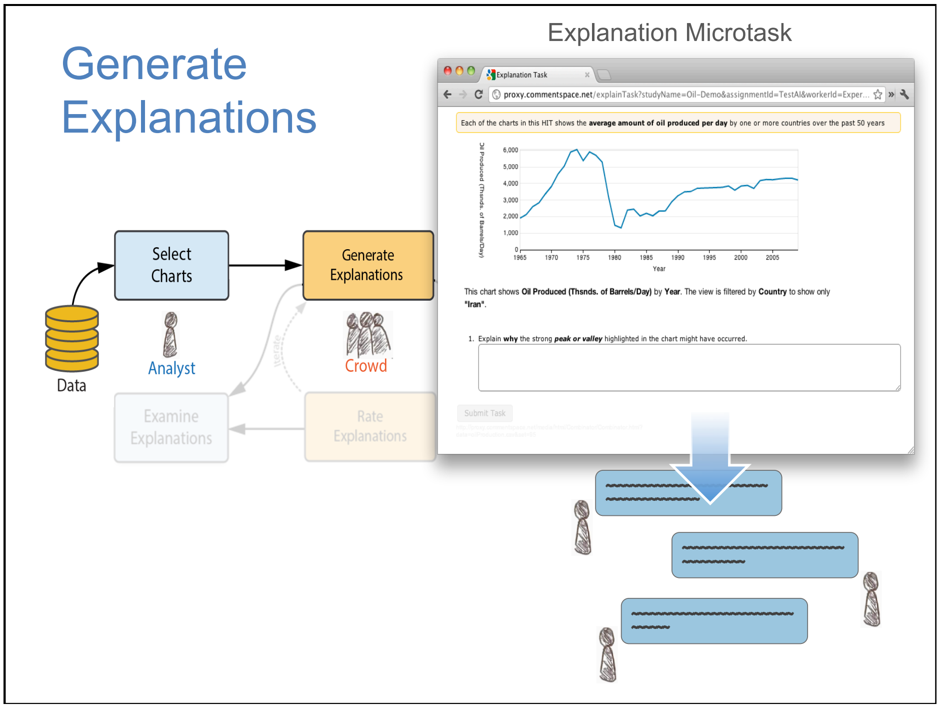
Generate Explanations

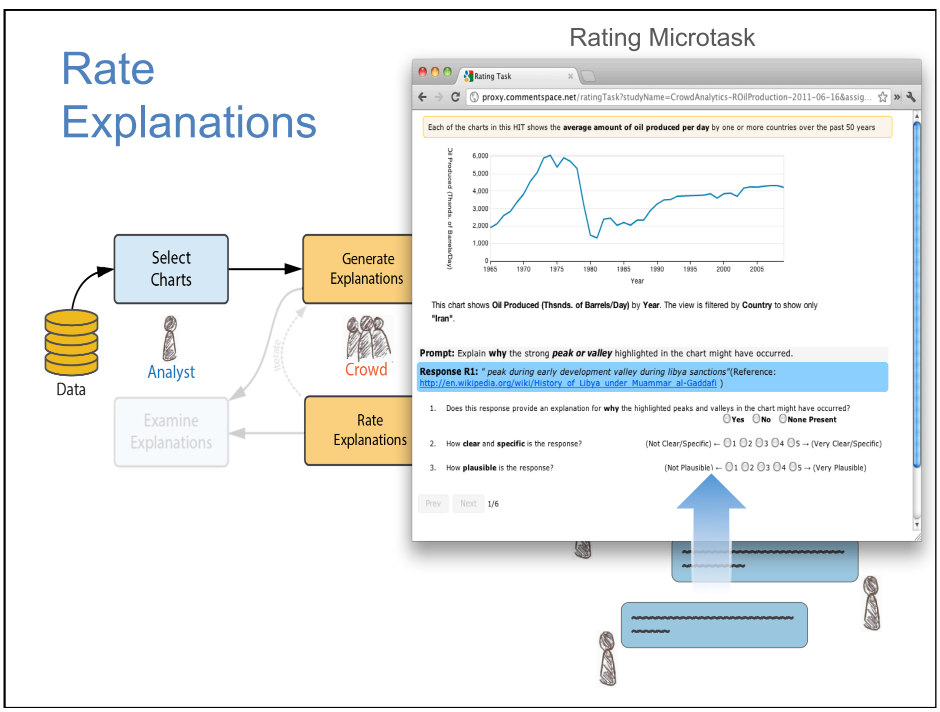
Explanation Microtask

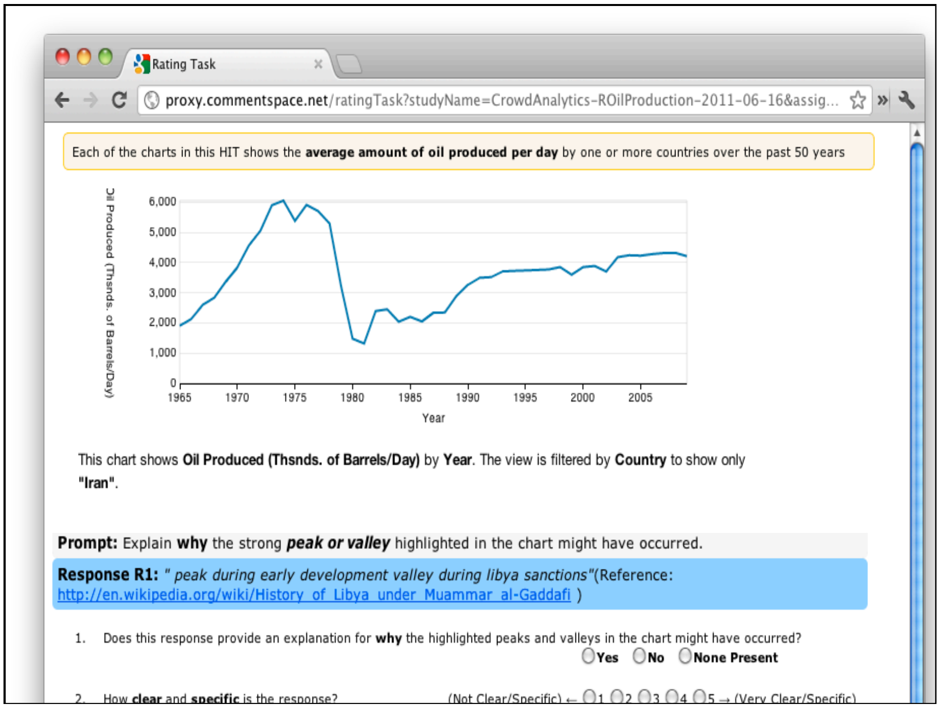




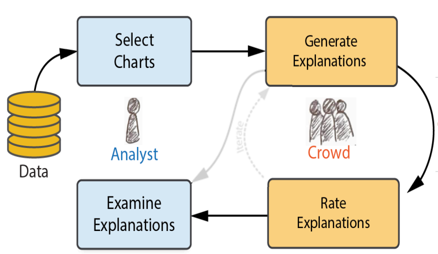


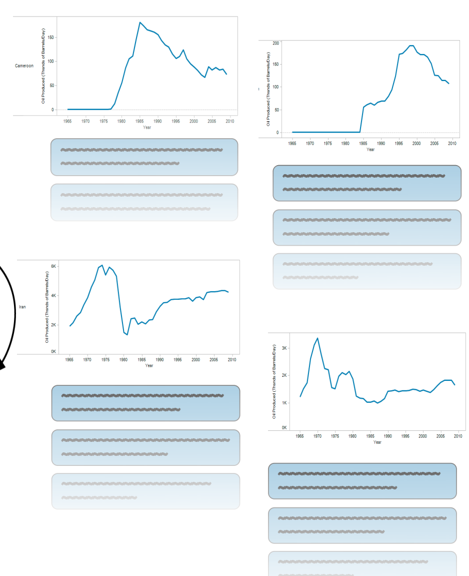

Explanation Microtask
Generate Explanations
Select Charts
Generate Explanations
Data
Analyst
Crowd
Examine Explanations
Rate Explanations
Explanation Task
Each of the charts in this HIT shows the average amount of oil produced per day by one or more countries over the past 50 years
This chart shows Oil Produced (Thsnds. of Barrels/Day) by Year. The view is filtered by Country to show only "Iran".
1. Explain why the strong peak or valley highlighted in the chart might have occurred.
Submit Task


Rating Microtask
Rate Explanations
Select Charts
Generate Explanations
Data
Analyst
Crowd
Examine Explanations
Rate Explanations
Rating Task
Each of the charts in this HIT shows the average amount of oil produced per day by one or more countries over the past 50 years
This chart shows Oil Produced (Thsnds. of Barrels/Day) by Year. The view is filtered by Country to show only "Iran".
Prompt: Explain why the strong peak or valley highlighted in the chart might have occurred.
Response R1: " peak during early development valley during libya sanctions"(Reference: http://en.wikipedia.org/wiki/History_of_Libya_under_Muammar_al-Gaddafi)
1. Does this response provide an explanation for why the highlighted peaks and valleys in the chart might have occurred?
Yes No None Present
2. How clear and specific is the response?
(Not Clear/Specific) ← 1 2 3 4 5 → (Very Clear/Specific)
3. How plausible is the response?
(Not Plausible) ← 1 2 3 4 5 → (Very Plausible)
Prev Next 1/6



Examine Explanations





Problems

Expectations may be unclear to workers.

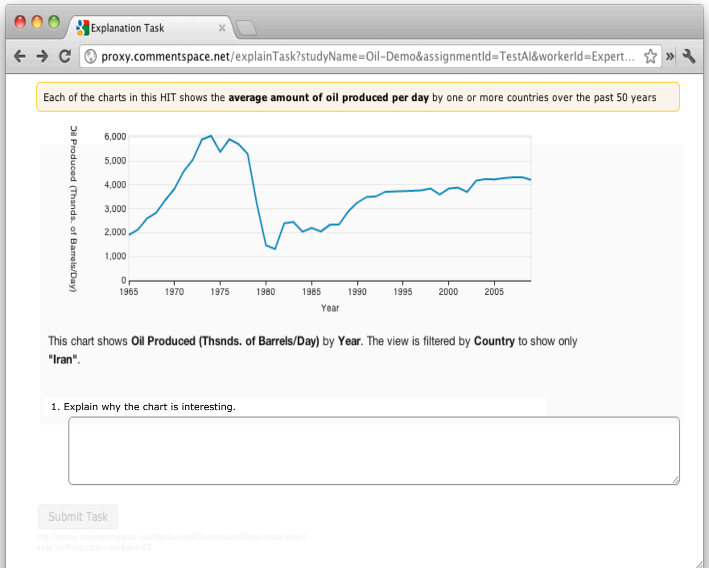
Workers may explain irrelevant features.

Workers' may give speculative explanations.

Workers may not attend to chart details.

Seven Strategies for Crowdsourcing Social Data Analysis

Explanation Microtask



S1. Use feature-oriented prompts.

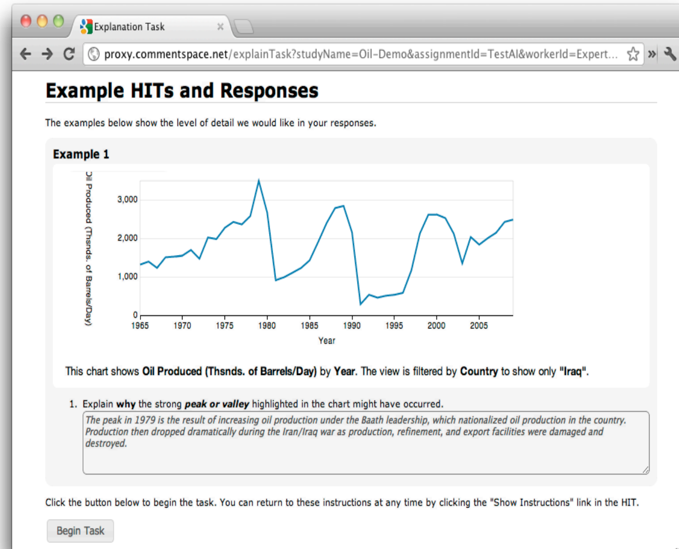
~~"Explain why the chart is interesting."~~

"Explain the **long term trend** in the chart."

"Explain the **peaks and/or valleys** in the chart (if any exist)."

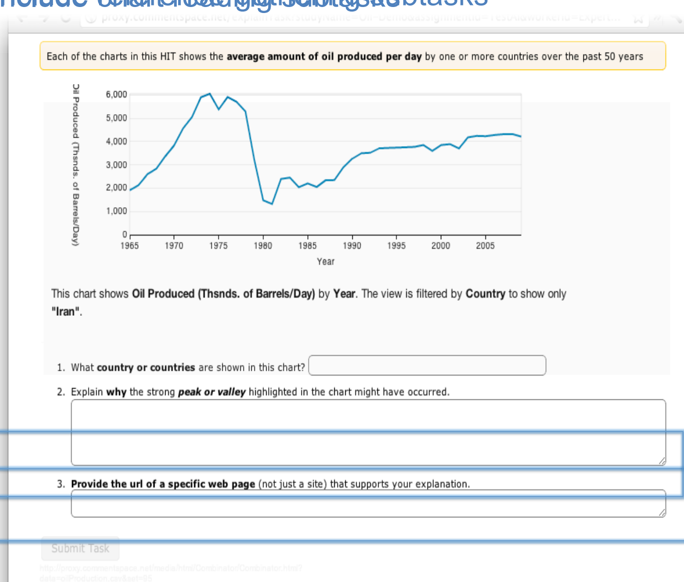
S2. Provide sample explanations

Explanation Microtask



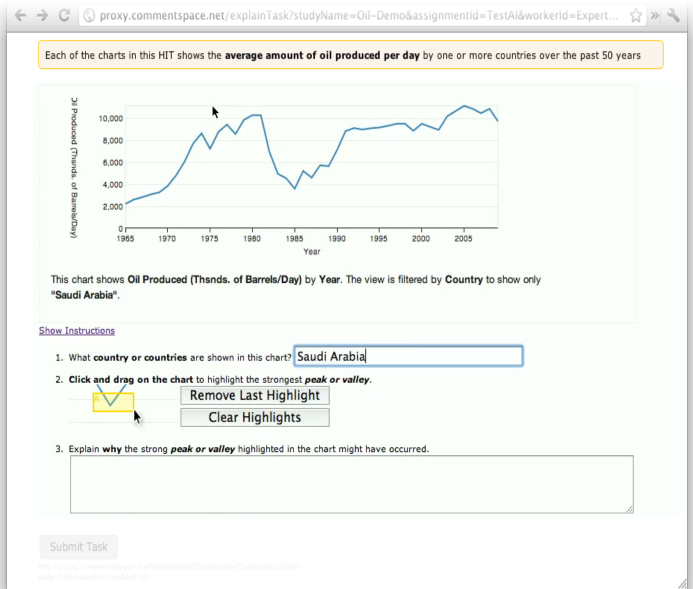
S3. Include chart reading subtasks

S4. Include reference gathering subtasks



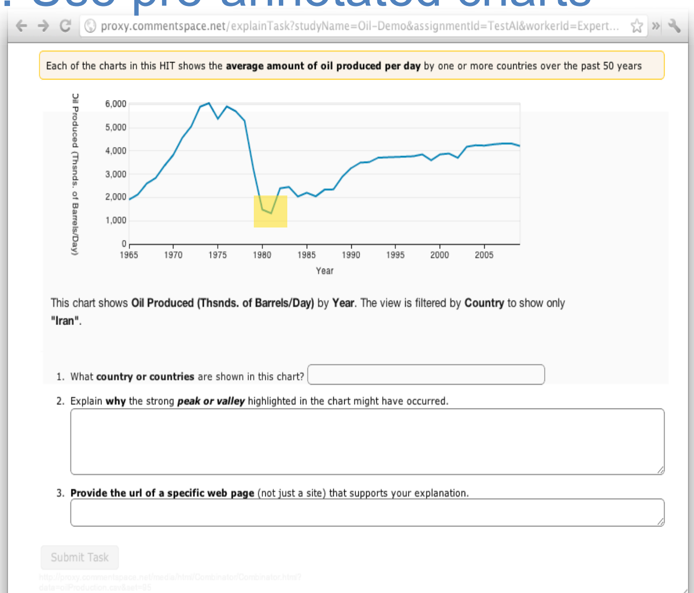
Explanation Microtask

S5. Include annotation subtasks

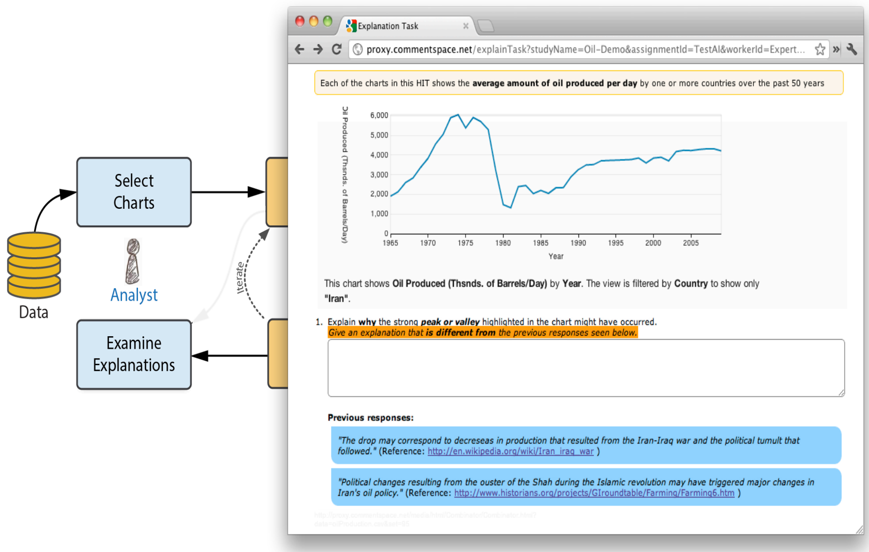


Explanation Microtask

S6. Use pre-annotated charts



S7. Elicit explanations iteratively

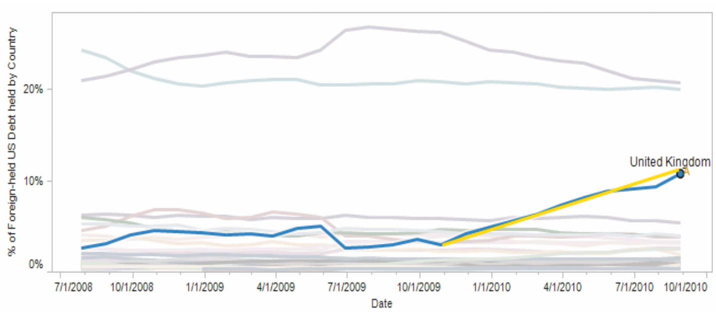


Deployment

Deployment via Mechanical Turk

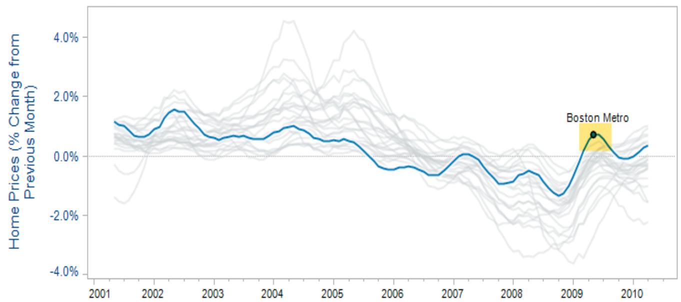
910 Responses for
64 Charts from
16 Datasets

Dataset: Foreign Holders of US Sovereign Debt 2008-2010



"… Speculation is that China is using UK brokers to purchase more US securities."

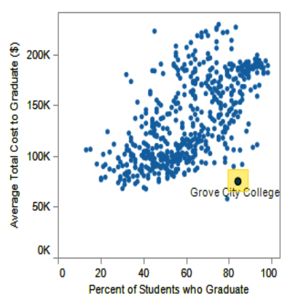
Dataset: US Metro Home Prices (% Change / Month) 2001-2010



"The primary reason for the peak is …
… annual variation where home prices rise in the spring.
… the $8,000 home buyer tax credit.

url: http://articles.boston.com/2011…indices-housing-prices

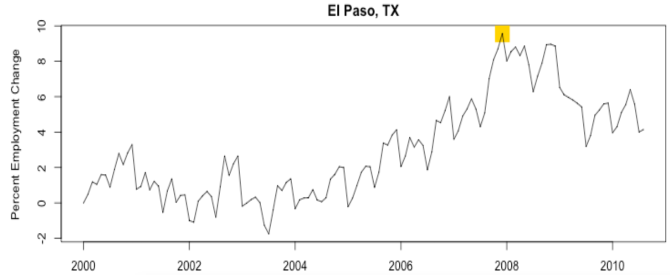
Dataset: US College Graduation Rates vs. Cost To Graduate



"Grove City College is Christian **based and** strives to keep tuition rates affordable **for most students.**
… located in a rural community …
… selective enrollment …"

url: http://stateuniversity.com/…college.html

Dataset: US Metro Percent Employment Change 2000-2010



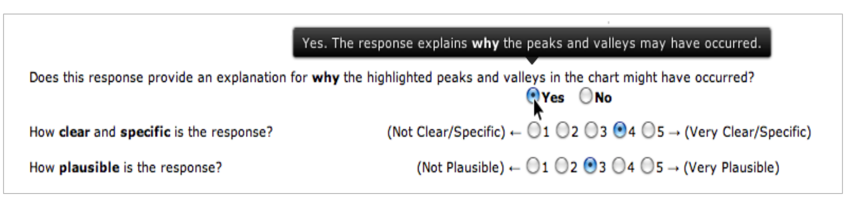
"Four large initiatives ...
... 1) Fort Bliss expansion ...
... 2) Construction at UTEP ...
... 3) new spending on highways ...
... 4) the Medical Center of the Americas "

url: http://newspapertree.com/opinion/3561-the-el-paso-stimulus

Quantifying Response Quality

relevance (0-1) Does the response explain the requested feature?
clarity (1-5) How clear and specific is the explanation?
plausibility (1-5) How likely is the explanation to be valid?

$$quality = relevance * (clarity + plausibility) / 2$$


Yes. The response explains **why** the peaks and valleys may have occurred.

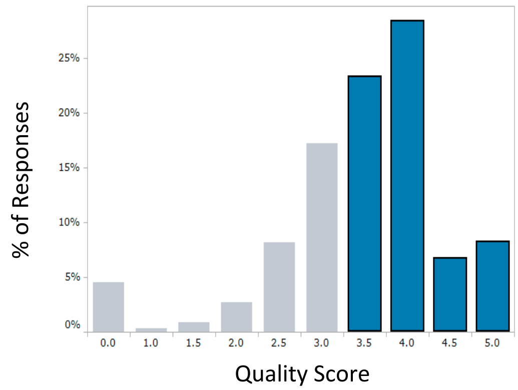
Does this response provide an explanation for **why** the highlighted peaks and valleys in the chart might have occurred?
Yes No

How **clear** and **specific** is the response? (Not Clear/Specific) ← 1 2 3 4 5 → (Very Clear/Specific)

How **plausible** is the response? (Not Plausible) ← 1 2 3 4 5 → (Very Plausible)

Experts scored response quality this way.

Based on our ratings, the majority (63%) of responses were of high quality ($quality \geq 3.5$).

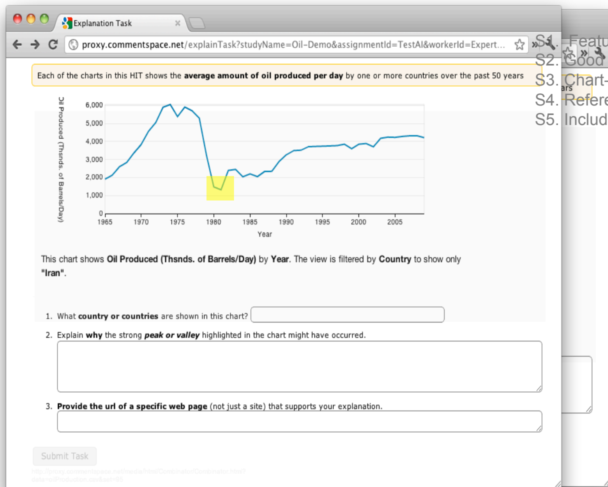


Do our strategies improve explanation quality?

Experiment 1 – Strategies S1-S5

No Strategies vs. Strategies (S1-S5)

US



S1. Feature-oriented prompts.
S2. Good examples.
S3. Chart-reading subtasks.
S4. Reference-gathering subtasks.
S5. Include annotation subtasks.

Experiment 1 - Strategies S1-S5

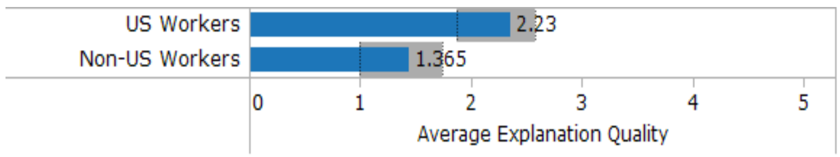
200 Trials, 4 conditions
(2 interfaces x 2 worker pools)

H1: Results from US workers would be of higher quality than results from non-US workers, but

H2: Employing strategies S1-S5 would increase the quality for both groups.

Experiment 1

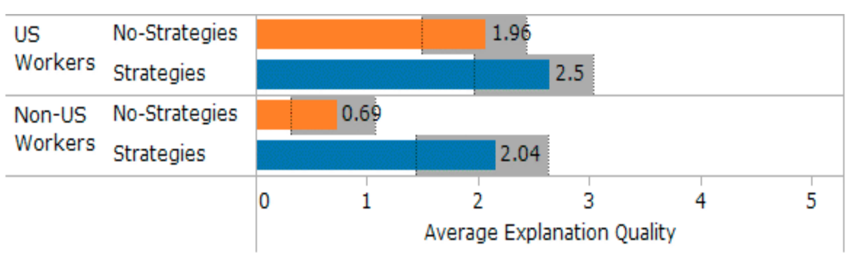
US Workers performed better than Non-US.



Experiment 1

Using S1-S5 increased explanation quality:

196% for Non-US Workers
28% for US Workers



Significant main effects for {

Worker Pool ($F_{1,198} = 12.2$, $p < 0.01$)

Strategies ($F_{1,198} = 14.5$, $p < 0.01$)

Experiments

Reference Gathering (S4)

Produced useful references, but slowed workers, increased attrition, and lowered quality.

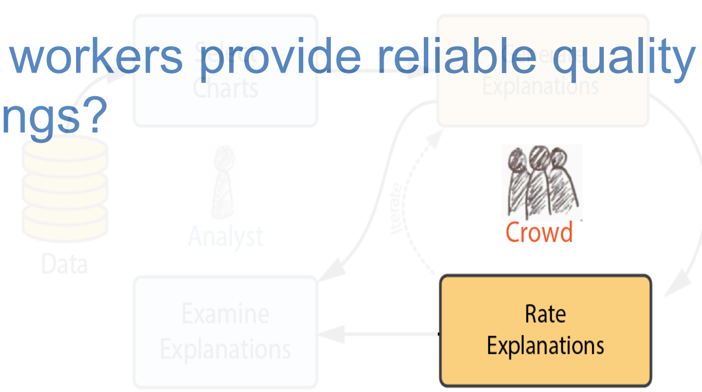
Annotation Subtasks (S5) and Pre-Annotation (S6)

Both improve attention to detail but are useful in different situations.

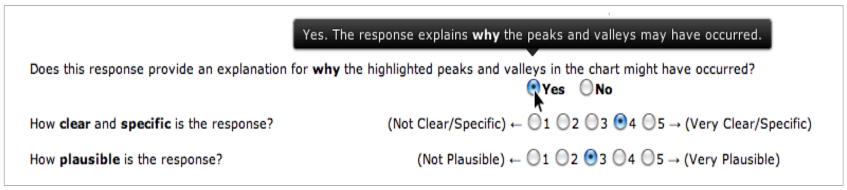
Eliciting Explanations Iteratively (S7)

An Iterative round produced 71% new explanations, increasing diversity.

Do workers provide reliable quality ratings?



Experiment 2 - Rating



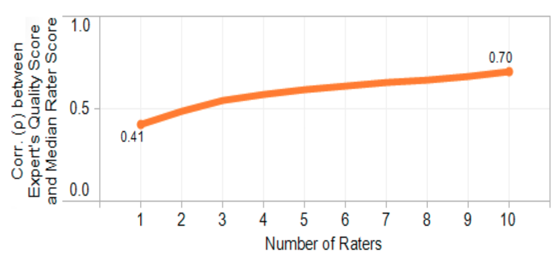
Elicited 1,334 ratings for 192 responses.

Compared against expert quality ratings.

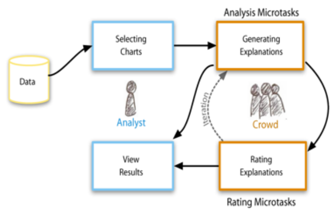
Experiment 2 - Rating

Individual workers' quality scores were moderately correlated with ours.
(ρ = 0.41)

The median quality from multiple workers correlates more strongly correlation with our scores.
(ρ = 0.70 w/ 10 raters)



Takeaways



S1. Use feature-oriented prompts.
S2. Provide good examples.
S3. Include chart-reading subtasks.
S4. Include reference-gathering subtasks.
S5. Include annotation subtasks.
S6. Pre-annotate charts.
S7. Elicit Explanations Iteratively

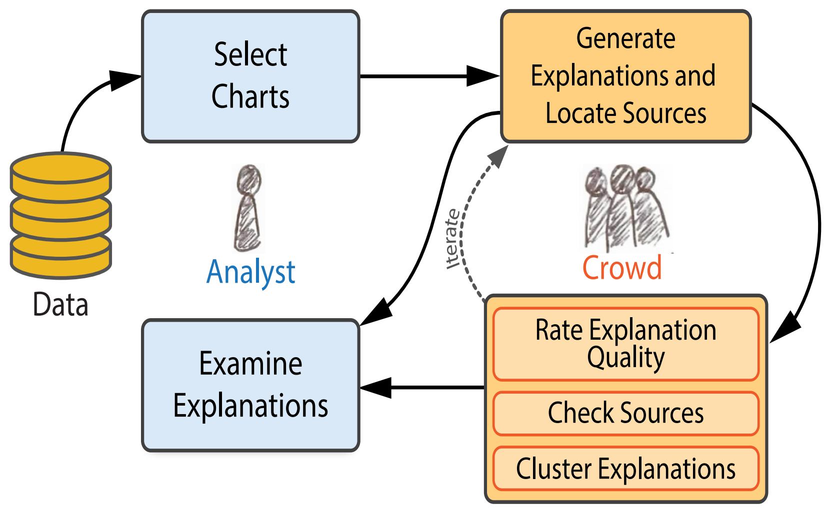
A workflow for generating and assessing explanations for trends and outliers.

Seven strategies for improving crowdsourced explanations.

Generates good explanations (63% were high-quality).

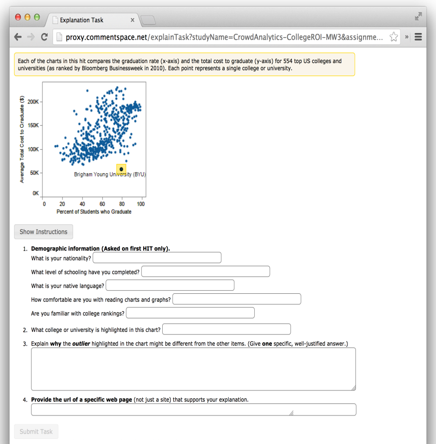
Exposing Provenance and Identifying Redundancy

Adding Additional Crowd Processing



Provenance

Provenance

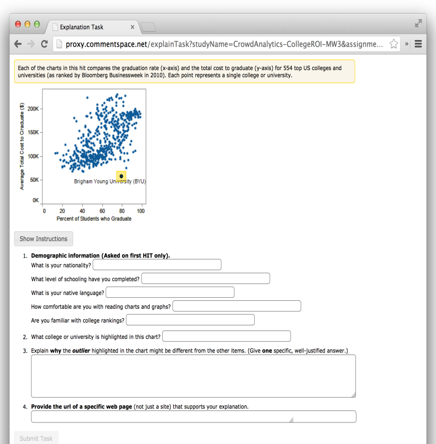


Explanation Task

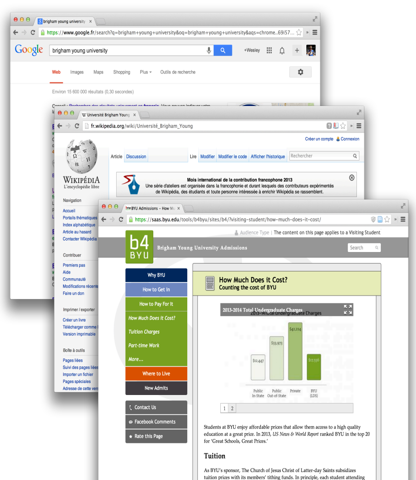
What are our

workers doing?

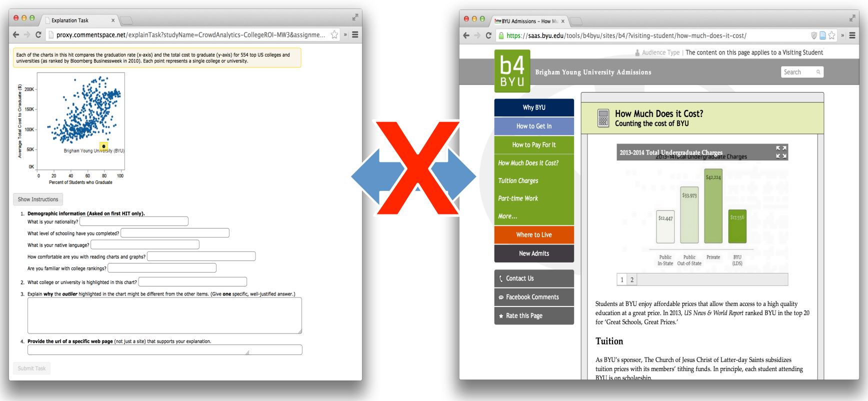
Provenance



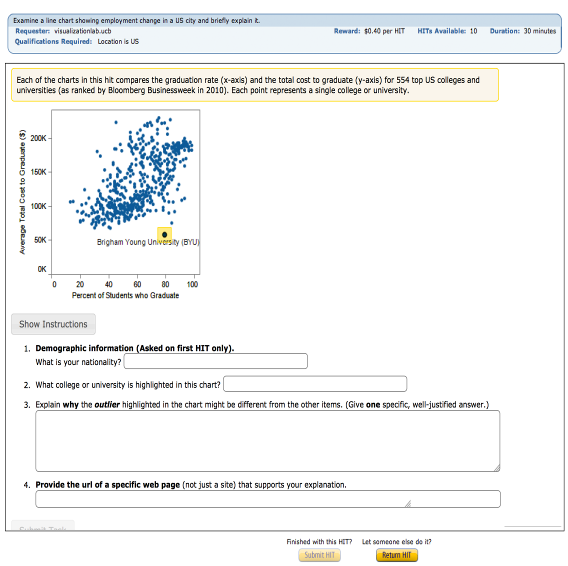
Explanation Task



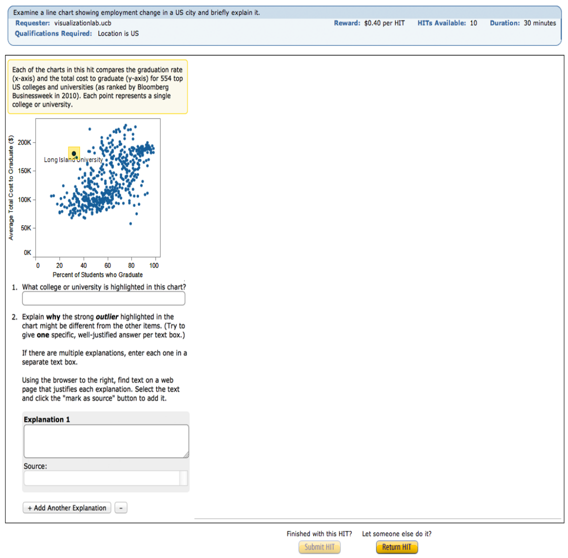
Cross-Domain Access is Limited



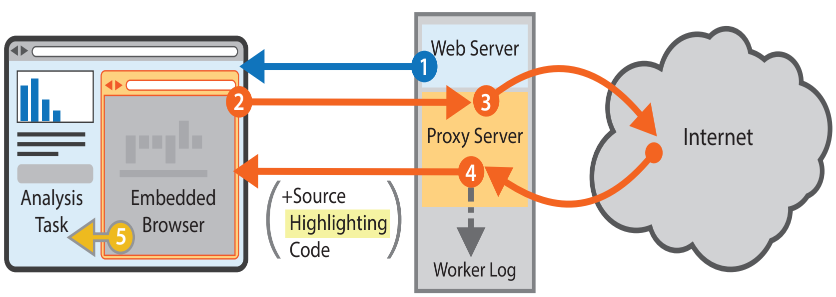
Instrumenting Explanation Tasks



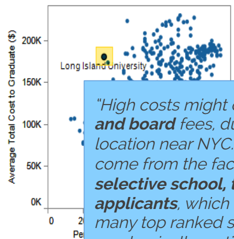
Instrumenting Explanation Tasks



Logging via Proxied Browser



Provenance



"High costs might come from it's **high room and board** fees, due to its geographic location near NYC. Low graduation rates come from the fact that it is ***not a very selective school, taking in over 80% of applicants***, which doesn't allow it take many top ranked students who are more academically motivated."

Paragraph-level citations

#123 Regional Universities (North)

Summary

LIU Post is a private institution that was founded in 1954. It has a total undergraduate enrollment of 8,315, its setting is suburban, and the campus size is 308 acres. It utilizes a semester-based academic calendar. LIU Post's ranking in the 2014 edition of Best Colleges is Regional Universities (North), 123. Its tuition and fees are $34,070 (2013-14).

2014 Quick Stats

720 Northern Boulevard
Brookville, NY 11548-1300
[map]
Phone: (516) 299-2000

2013-2014 Tuition
$34,070 tuition and fees

Students
8,315 enrolled
25% male / 75% female

Admissions
rolling admission
78.8% accepted

More Information

Visitation logs

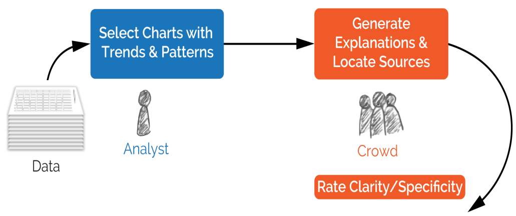
2011-12-11 09:22:04 google.com
2011-12-11 09:22:04 sqr:help
2011-12-11 09:23:08 google.com/search?hl=en&source=hp
2011-12-11 09:23:11 google.com/search?hl=en&q=Long Is
2011-12-11 09:23:13 google.com/search?q=Long Island Ur
2011-12-11 09:23:31 google.com/search?q=Long Island Ur
2011-12-11 09:23:38 google.com/search?q=Long Island Ur
2011-12-11 09:23:43 google.com/search?q=Long Island Ur
2011-12-11 09:23:54 google.com/search?q=Long Island Ur
2011-12-11 09:24:09 colleges.usnews.rankingsandreviews.c

Provenance

Did the **facts** and **inference** come from the **source** or did the **worker** add them?

Paragraph-level citations
Visitation logs
2011-12-11 09:24:09 colleges.usnews.rankingsandreviews.c
More Information

Source-Checking Microtasks

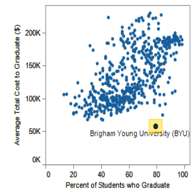


A second group of workers **verifies links** and attributes explanations to the **source** or the **worker**.
(75% accurate in our preliminary tests)

Redundancy

Redundancy

Many explanations provided by workers are **redundant**.



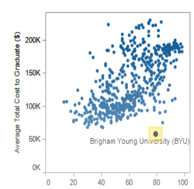
"The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs…"

"The cost of attendance at BYU is subsidized by the LDS church."

"98% of their students are members of LDS and they have lowered tuition…"

Redundancy

Many explanations provided by workers are **redundant**.



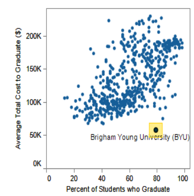
"The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs…"

"The cost of attendance at BYU is subsidized by the LDS church."

− Duplicate results for analysts to examine.

+ Redundancy can signal high support and corroborating sources.

Redundancy



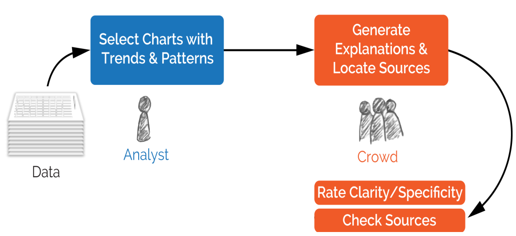
Automated text similarity methods don't deal well with these kinds of content.

"The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs…"

"The cost of attendance at BYU is subsidized by the LDS church."

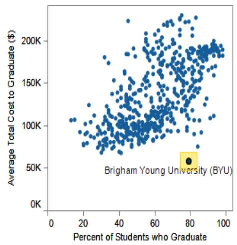
"98% of their students are members of LDS and they have lowered tuition…"

Redundancy



Can we crowdsource redundancy detection?

Clustering via Distributed Comparison

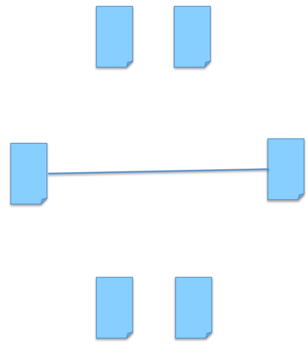


"98% of their students are members of LDS and they have lowered tuition…"

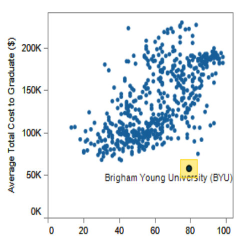
"The cost of attendance at BYU is subsidized by the LDS church."

"…students are mostly members of the church and bound by the honor code…"

"The Church of Jesus Christ of Latter Day



Clustering via Distributed Comparison

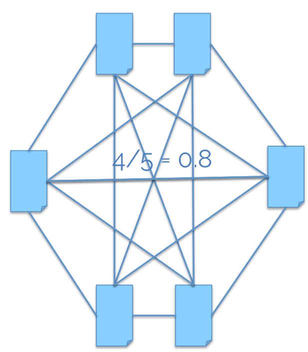


"98% of their students are members of LDS and they have lowered tuition…"

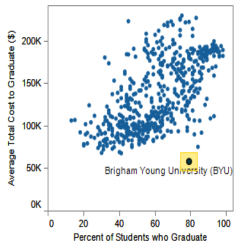
"The cost of attendance at BYU is subsidized by the LDS church."

Do these two responses give the same general explanation for the peaks and valleys in the chart?

- ◯ **Yes. Both responses give the same general explanation.**
- ◯ **No. The responses do not give the same explanation.**



Clustering via Distributed Comparison



"98% of their students are members of LDS and they have lowered tuition..."

"The cost of attendance at BYU is subsidized by the LDS church."

Do these two responses give the same general explanation for the peaks and valleys in the chart?

◯ **Yes. Both responses give the same general explanation.**

◯ **No. The responses do not give the same explanation.**

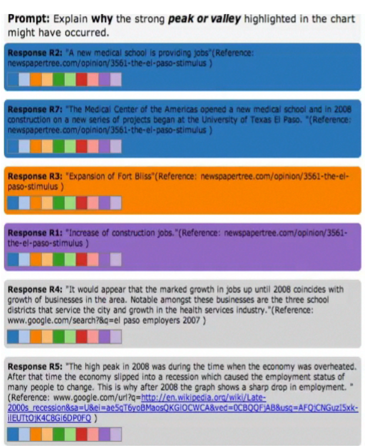
+ **Simple** tasks for workers

− **Scales p rly** 2)

Sensitive to **clustering method**

Workers have **little context**

Clustering via Color-Coding



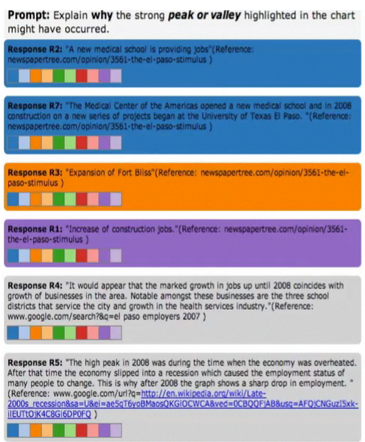
Individual workers cluster the whole set.

+ Workers have **complete context**

Individual workers can cluster badly

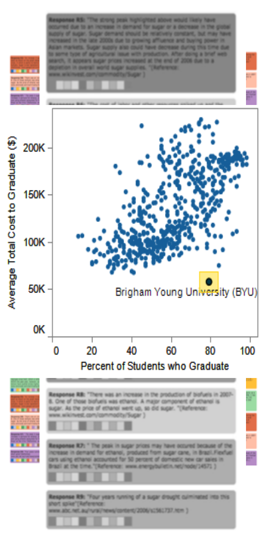
− **Hard to integrate clusterings** from multiple workers

How to Integrate Color-Clusterings?



1. A **single worker's clustering** is preferable to a combination of multiple clusterings.
2. **Clusters reproduced by multiple independent workers** are likely to reflect actual redundancy.
3. Errors tend to be either **noisy** or **easy to catch**.

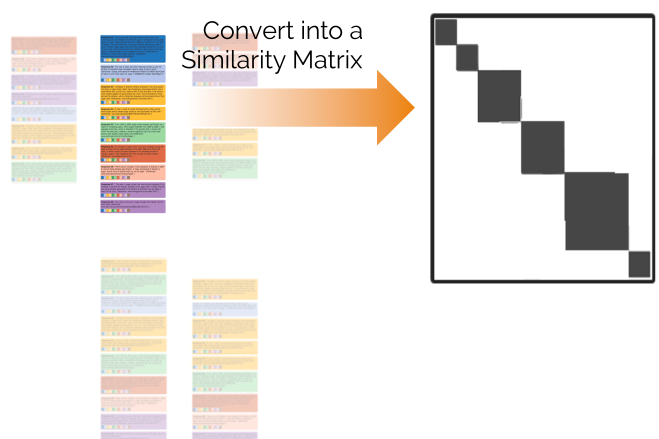
How to Integrate Color-Clusterings?



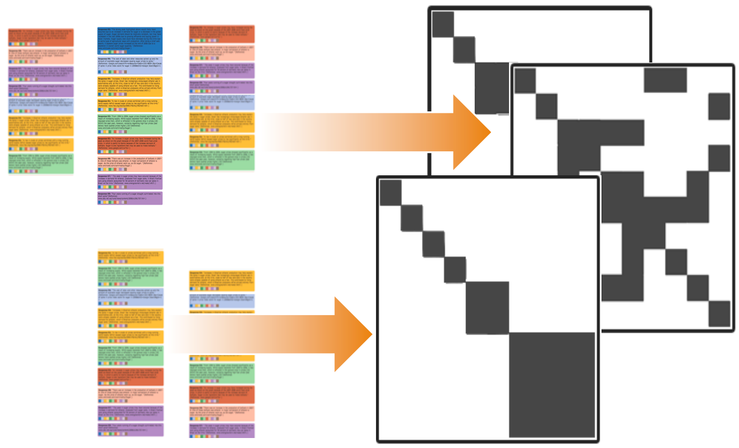
Selecting the **Most-Representative Clustering**

How to Integrate Color-Clusterings?

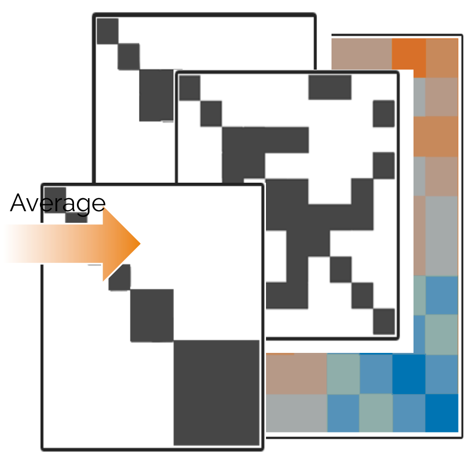
Selecting the Most-Representative Clustering



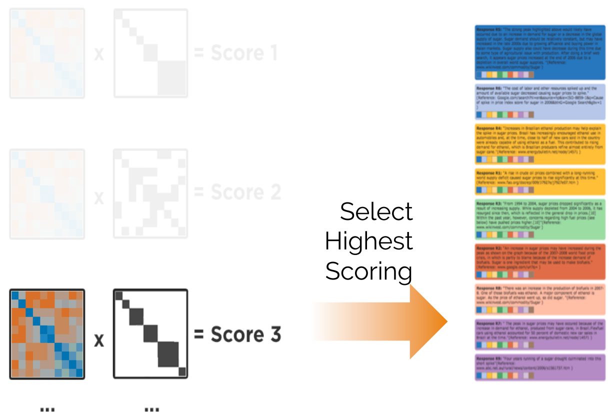
Selecting the Most-Representative Clustering



Selecting the Most-Representative Clustering



Selecting the Most-Representative Clustering



Evaluating Redundancy-Detection

Does color clustering with most-representative selection produce good clusterings?

Our Explanation Dataset

12 charts (4 each from 3 different data sets)

10 workers explained each chart

93 Workers produced 156 explanations (avg=13/chart)

Evaluating Redundancy-Detection

Does color clustering with most-representative selection produce good clusterings?

10 Workers used **color clustering** to group the explanations for each chart. (120 total clusterings)

We used **most-representative selection** to pick the best clustering for each chart. (12 clusterings)

Evaluating Redundancy-Detection

Baseline - Expert clustering (x 3)

To score a clustering, we use the **F-measure** to compute similarity to each expert, then average.

(completely dissimilar) [0 ⟷ 1] (identical)

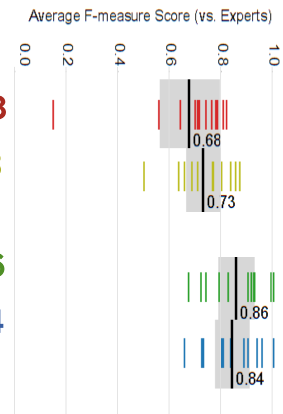
Evaluating Redundancy-Detection

Unclustered Results **F=0.68**

Color Clustering **F=0.73**

Most-Representative Selection **F=0.86**

Experts vs. One Another **F=0.84**



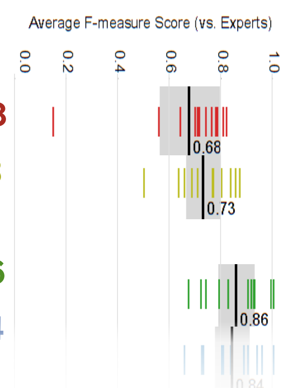
Evaluating Redundancy-Detection

Unclustered Results **F=0.68**

Color Clustering **F=0.73**

Most-Representative Selection **F=0.86**

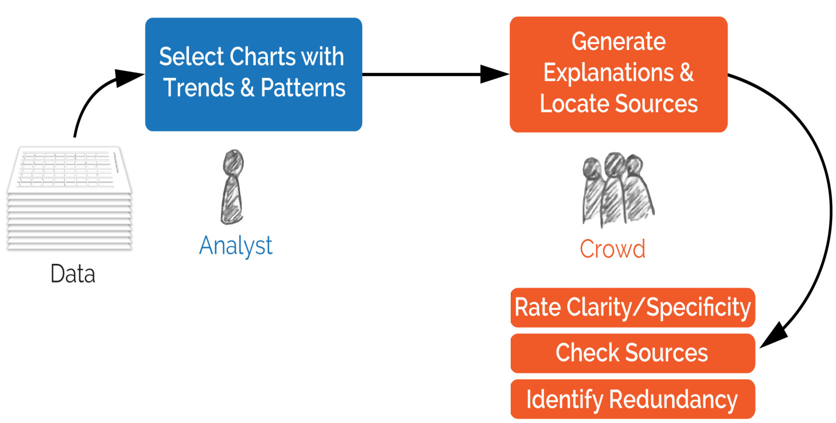
Experts vs. One Another **F=0.84**



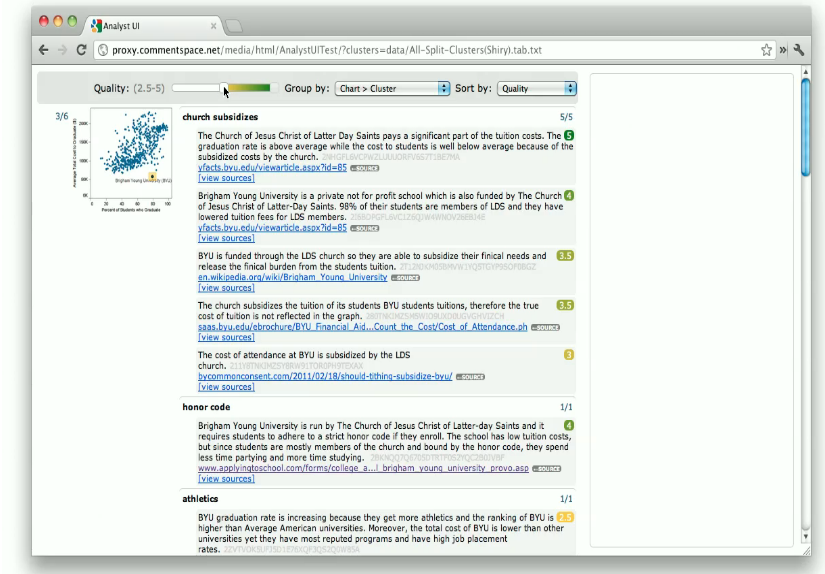
T-tests showed our **most-representative** results were significantly closer to experts than **color clustering** or **unclustered** were. (both $p < 0.01$)

Managing the Crowd's Work

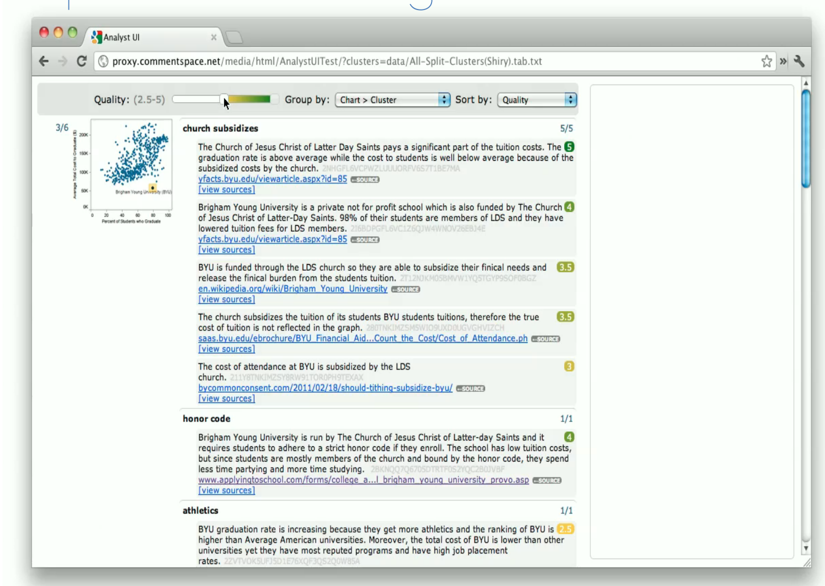
Managing the Crowd's Work



Explanation Management Interface



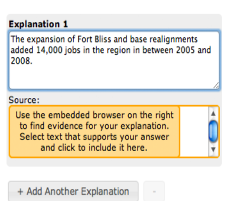
Explanation Management Interface



Open Questions

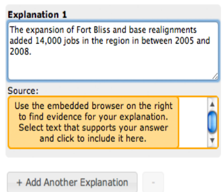
Open Questions

Segmenting Explanations



Open Questions

Segmenting Explanations



Crowd Composition

cheap low-skill crowds
vs.
more knowledgeable trusted ones

Conclusion