

# DATA ANALYSIS AT SCALE

PETRA ISENBURG (slides by WESLEY WILLETT)

VISUAL ANALYTICS

# DATA ANALYSIS AT SCALE

CHALLENGES

ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

PARALLELIZING HUMAN INTELLIGENCE

# CHALLENGES FOR ANALYZING LARGE DATA SETS

**SIZE** **SPEED**

**ATTENTION**

# SIZE

KILOBYTES OF DATA

MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

PETABYTES OF DATA

...

# SIZE

KILOBYTES OF DATA

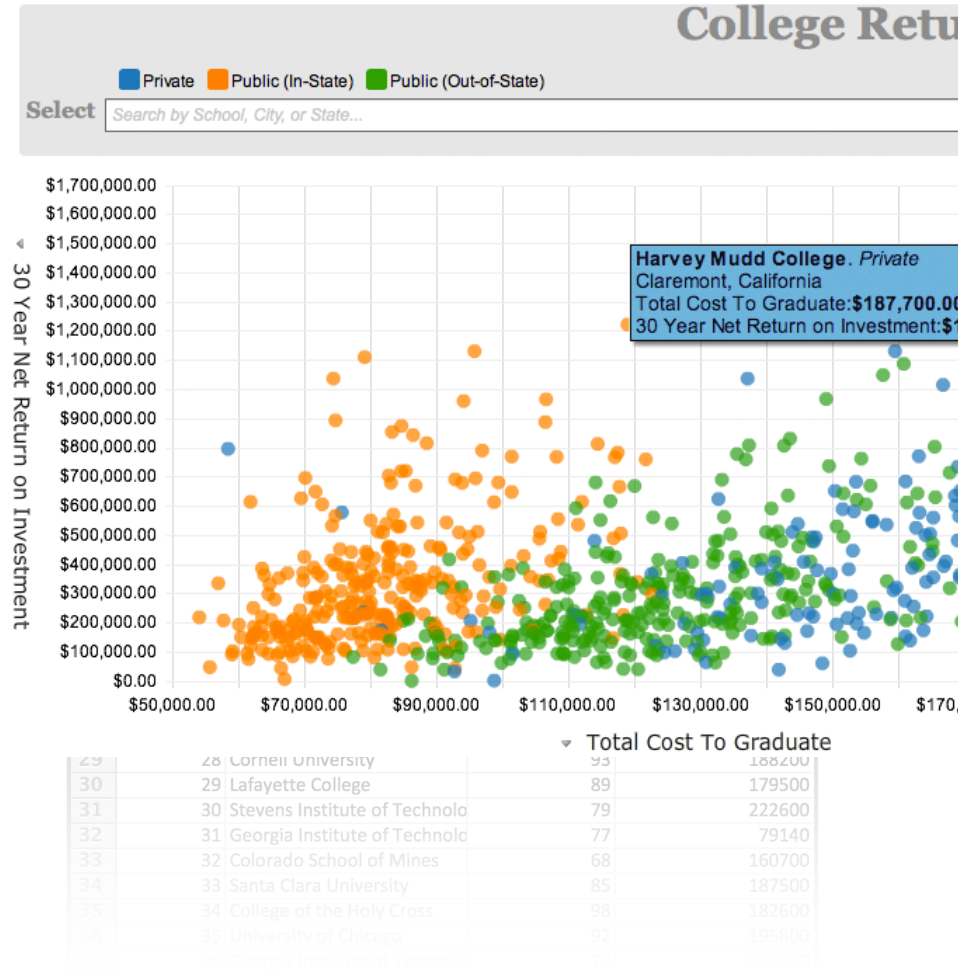
MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

PETABYTES OF DATA

...



# SIZE

KILOBYTES OF DATA

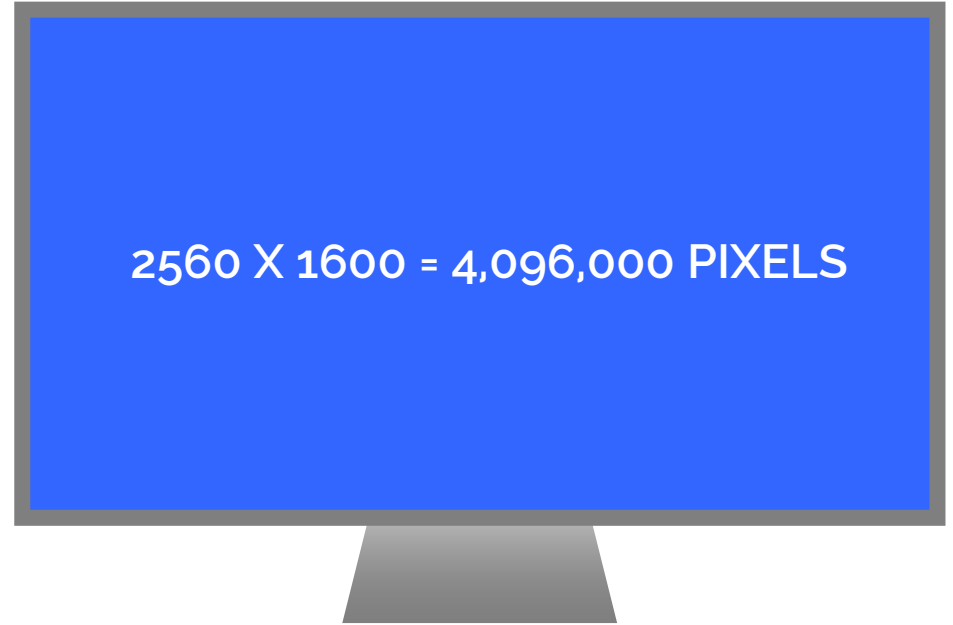
MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

PETABYTES OF DATA

...



**EVEN A MEGABYTE IS MORE BITS OF DATA  
THAN THERE ARE PIXELS ON A SCREEN!**

# SIZE

KILOBYTES OF DATA

MEGABYTES OF DATA

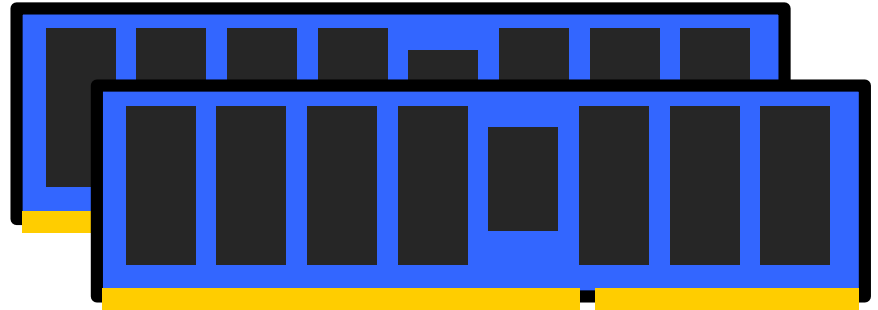
**GIGABYTES OF DATA**

TERABYTES OF DATA

PETABYTES OF DATA

...

**MORE DATA THAN CAN FIT IN MEMORY**



# SIZE

KILOBYTES OF DATA

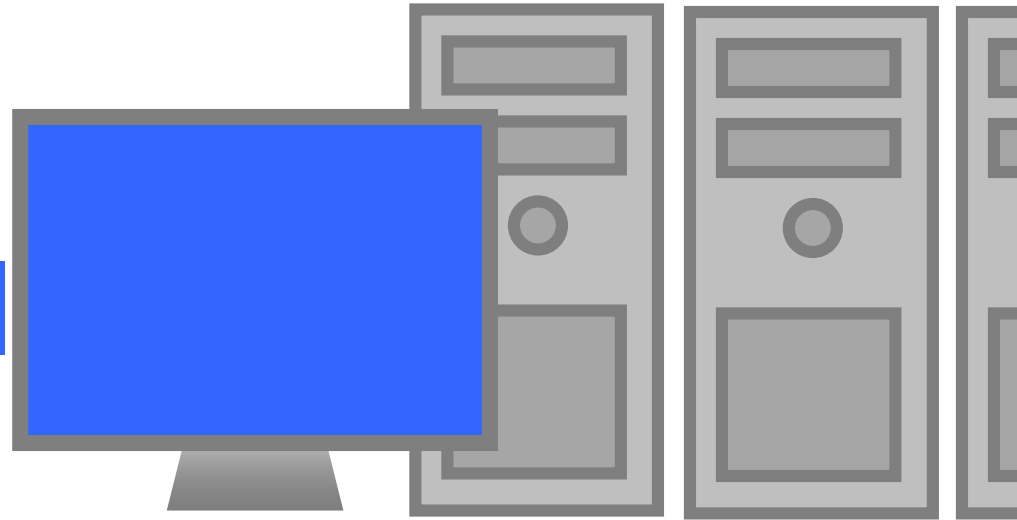
MEGABYTES OF DATA

GIGABYTES OF DATA

**TERABYTES OF DATA**

PETABYTES OF DATA

...



**MORE DATA THAN CAN FIT ON ONE MACHINE!**



# SIZE

KILOBYTES OF DATA

MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

**PETABYTES OF DATA**

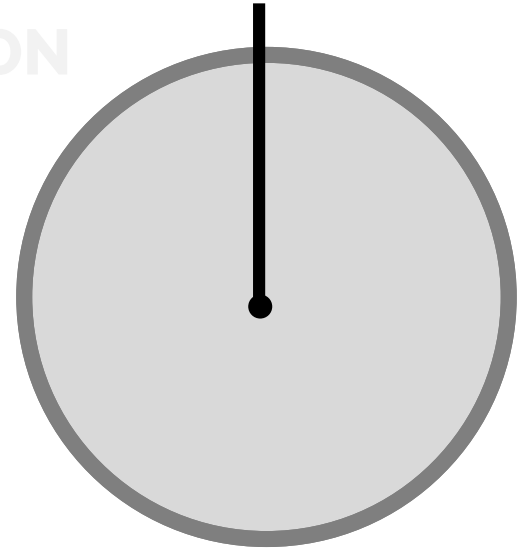


**MANY BIG DATA-DRIVEN  
QUESTIONS TODAY**



# SPEED

~0.1 SECOND	DIRECT MANIPULATION
~1 SECOND	INTERACTIVE
~10 SECONDS	QUERY / RESPONSE
MINUTES	...
HOURS	BATCH PROCESSING (VERY SLOW)



# SPEED

**~0.1 SECOND**

**DIRECT MANIPULATION**

~1 SECOND

INTERACTIVE

~10 SECONDS

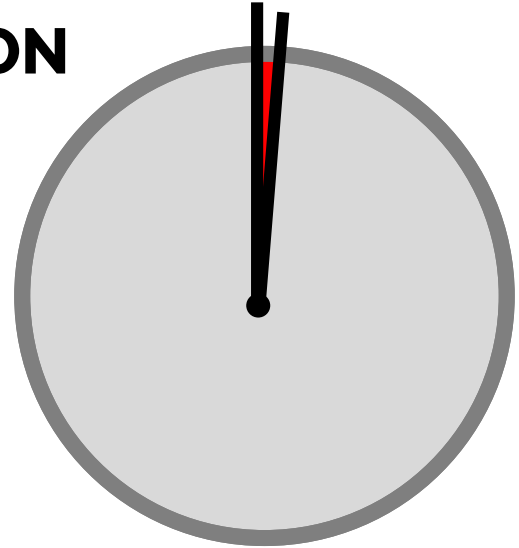
QUERY / RESPONSE

MINUTES

...

HOURS

BATCH PROCESSING  
(VERY SLOW)



# SPEED

**~0.1 SECOND**

DIRECT MANIPULATION

**~1 SECOND**

**INTERACTIVE**

~10 SECONDS

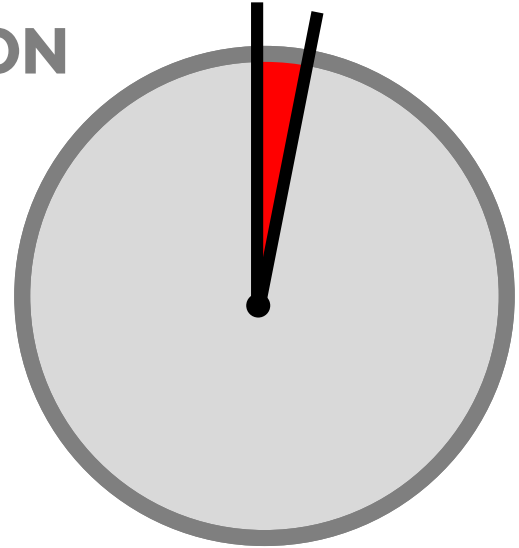
QUERY / RESPONSE

MINUTES

...

HOURS

BATCH PROCESSING  
(VERY SLOW)



# SPEED

**~0.1 SECOND**

DIRECT MANIPULATION

**~1 SECOND**

INTERACTIVE

**~10 SECONDS**

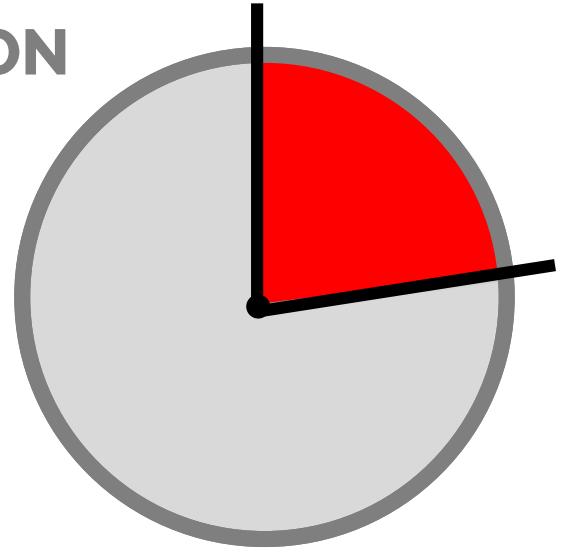
**QUERY / RESPONSE**

MINUTES

...

HOURS

BATCH PROCESSING  
(VERY SLOW)



# SPEED

**~0.1 SECOND**

DIRECT MANIPULATION

**~1 SECOND**

INTERACTIVE

**~10 SECONDS**

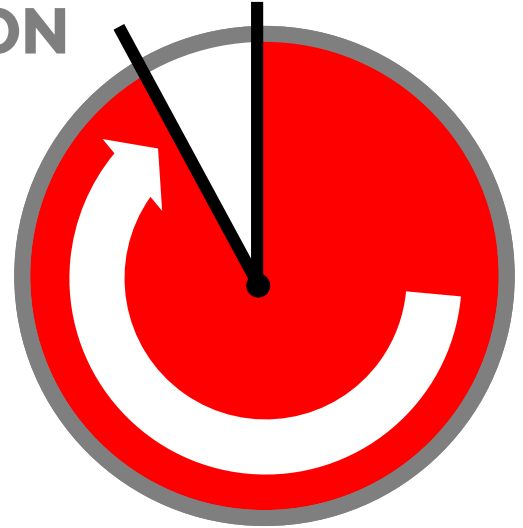
QUERY / RESPONSE

**MINUTES**

...

**HOURS**

**BATCH PROCESSING  
(VERY SLOW)**



# ATTENTION

EVERY PERSON ONLY HAS A FINITE  
NUMBER OF WORKING HOURS

**5-8 PERSON-HOURS PER DAY**

**1,489 PERSON-HOURS PER YEAR (FRANCE)**

(1,388 GERMANY 2,163 IN S. KOREA 1,788 IN USA) [IOECD STATS](#)

HOW LONG CAN YOU AFFORD TO SPEND FINDING EXAMPLES,  
PROCESSING A DATASET, OR ANSWERING A QUESTION?

# ATTENTION

AN INDIVIDUAL ANALYST IS UNLIKELY  
TO BE ABLE TO SEE DATA FROM  
MANY PERSPECTIVES

**“MANY EYES FIND MORE BUGS”**



# DATA ANALYSIS AT SCALE

CHALLENGES

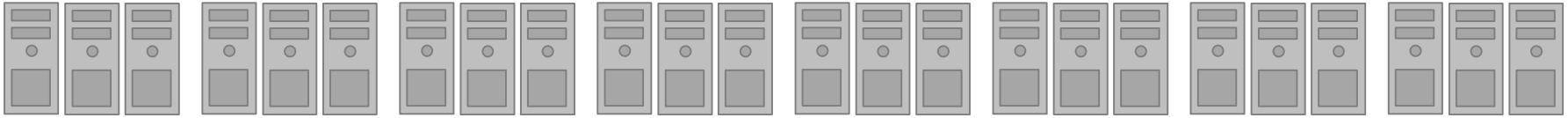
ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

PARALLELIZING HUMAN INTELLIGENCE

# ANALYSIS & CLUSTER COMPUTING

BIG DATASETS ARE LIKELY TO BE  
SPREAD OUT ACROSS A CLUSTER (OR  
CLUSTERS)



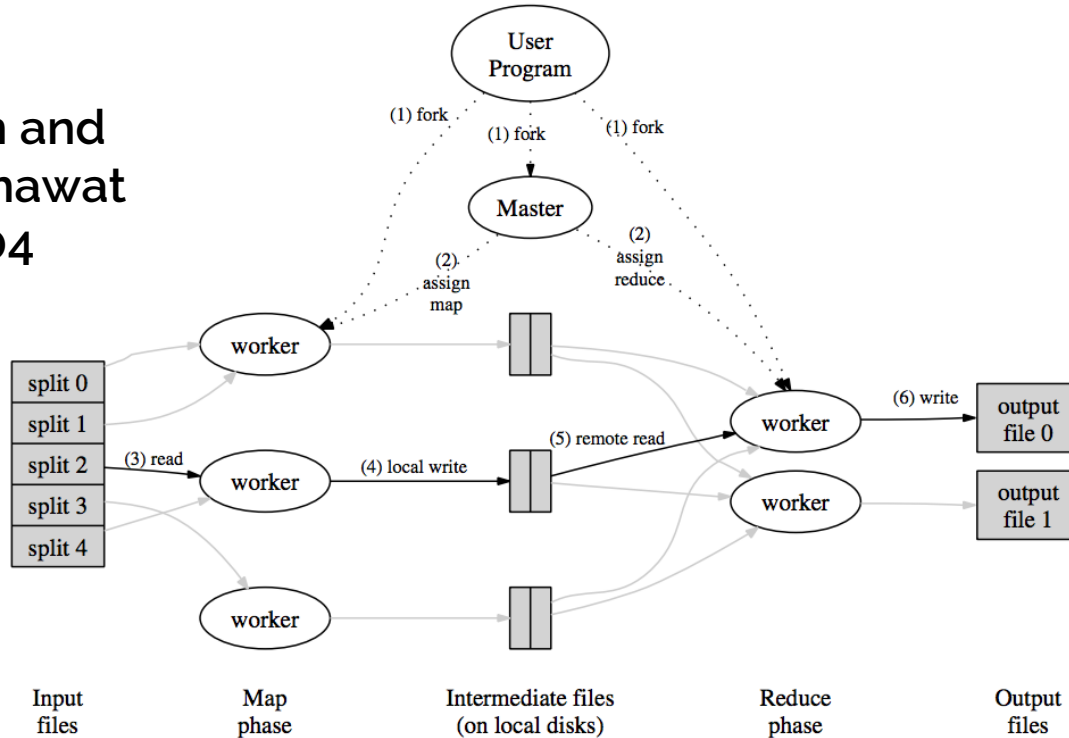
ANALYSIS REQUIRES  
DISTRIBUTED DATA PROCESSING

# **HOW CAN WE PERFORM ANALYSIS ACROSS A CLUSTER?**

How can we split work across machines?

# MAP-REDUCE

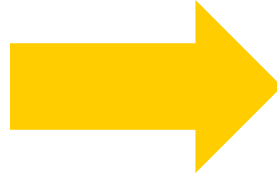
Jeffrey Dean and  
Sanjay Ghemawat  
(Google) 2004



# A SIMPLE EXAMPLE

HOW TO COUNT NUMBER OF TIMES WORDS OCCUR IN A DOCUMENT?  
(IF THAT DOCUMENT IS SPREAD ACROSS MANY MACHINES)

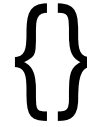
“I am Sam  
I am Sam  
Sam I am  
Do you like  
Green eggs and ham?”



I: 3  
am: 3  
Sam: 3  
do: 1  
you: 1  
like: 1  
...

# JUST A HASH TABLE

“I am Sam  
I am Sam  
Sam I am  
Do you like  
Green eggs and ham?”



# JUST A HASH TABLE

“I am Sam  
I am Sam  
Sam I am  
Do you like  
Green eggs and ham?”

{I:1}

# JUST A HASH TABLE

“I am Sam  
I am Sam  
Sam I am  
Do you like  
Green eggs and ham?”

{I:1,  
am:1}



# JUST A HASH TABLE

“I am Sam  
I am Sam  
Sam I am  
Do you like  
Green eggs and ham?”

{I:1,  
am:1,  
Sam:1}

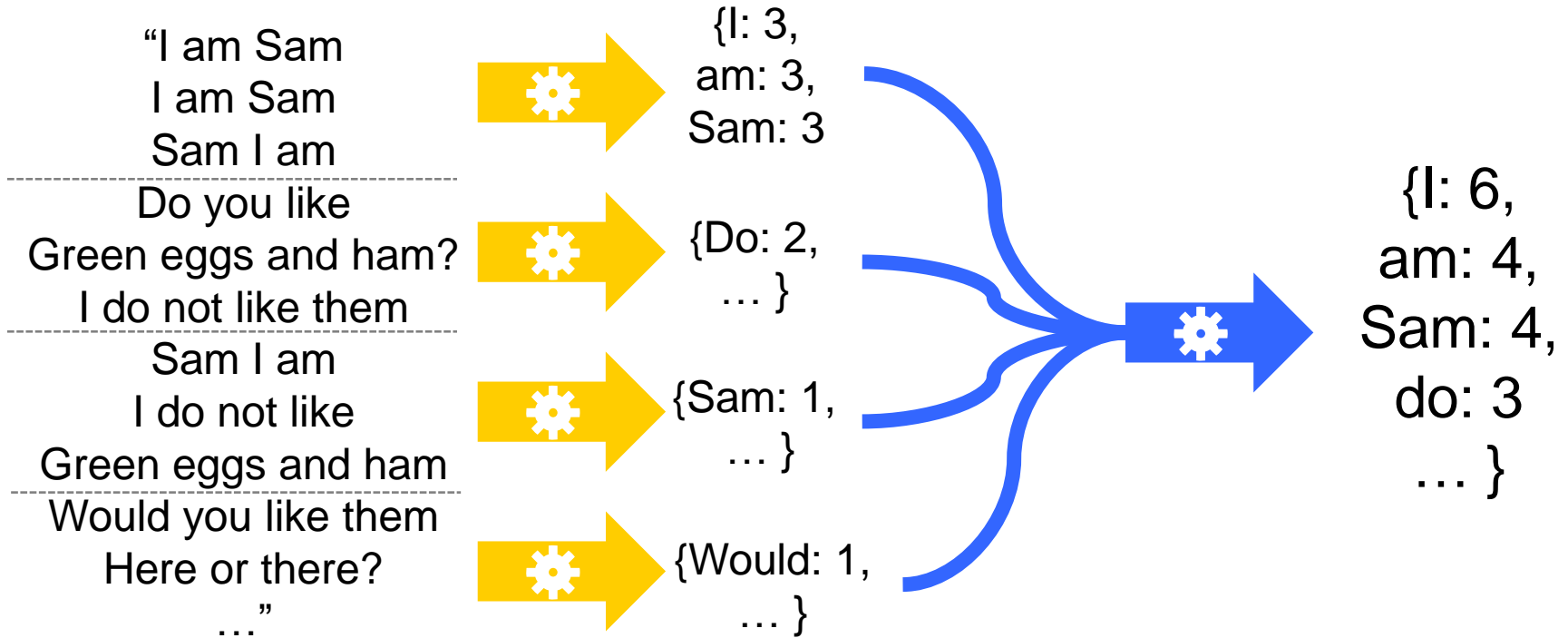
# JUST A HASH TABLE

“I am Sam  
I am Sam  
Sam I am  
Do you like  
Green eggs and ham?”

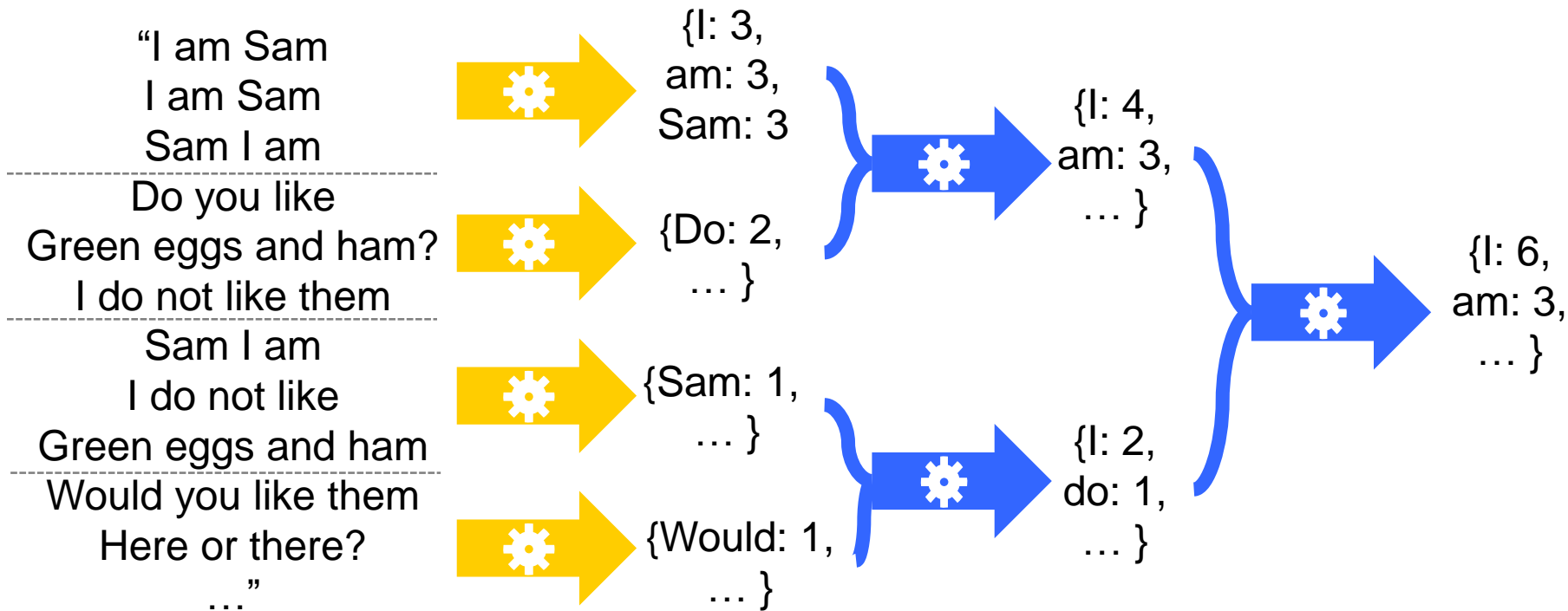
{I:2,  
am:1,  
Sam:1}

**BUT YOU SAID THE  
DOCUMENT IS REALLY BIG?**

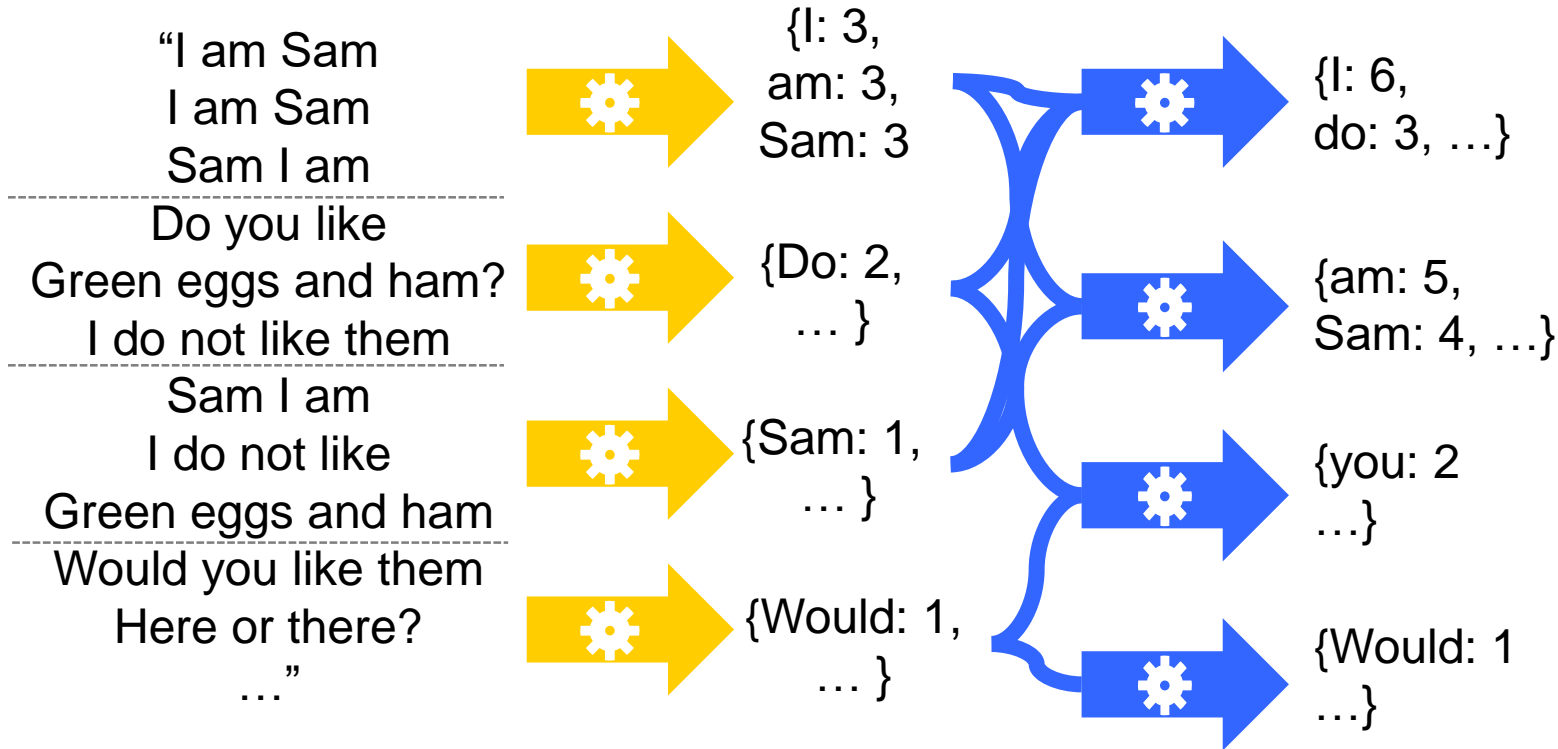
# COMPUTE IN PARALLEL



# COMPUTE IN PARALLEL

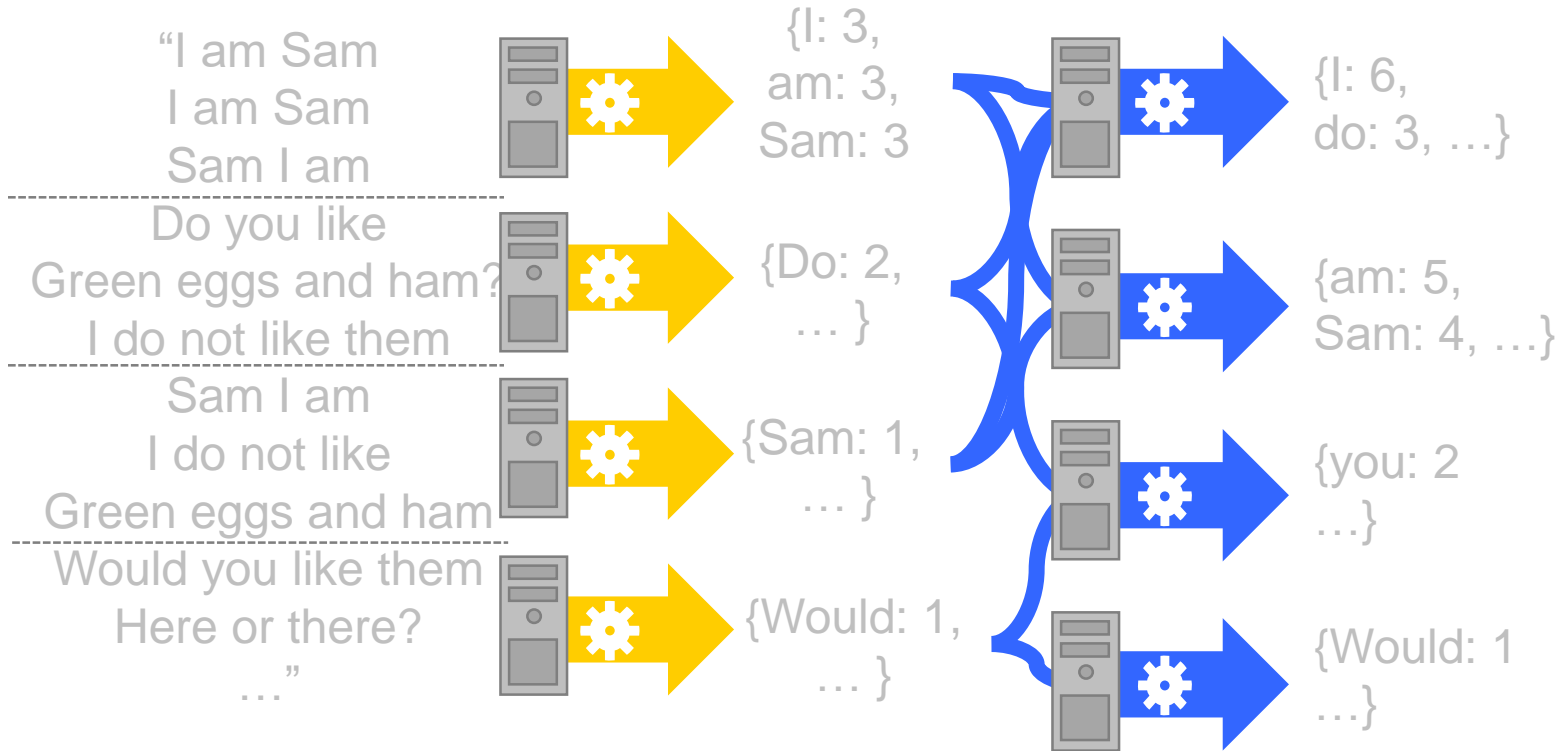


# COMPUTE IN PARALLEL



# MAP

# REDUCE



# MAP-REDUCE

SPLIT DATA & SEND TO MULTIPLE  
MACHINES (IF NOT ALREADY THERE)

**MAP**

FILTER, SORT, AND PROCESS  
DATA LOCALLY

**REDUCE**

CONSOLIDATE AND  
SUMMARIZE



# MAP-REDUCE

CAN BE SHORT, SELF-CONTAINED FUNCTIONS

(HERE AS PYTHON-ESQUE PSEUDO CODE)



MAP

```
function Map(Document document):  
    for each Word w in document:  
        EmitIntermediate(w, 1)
```



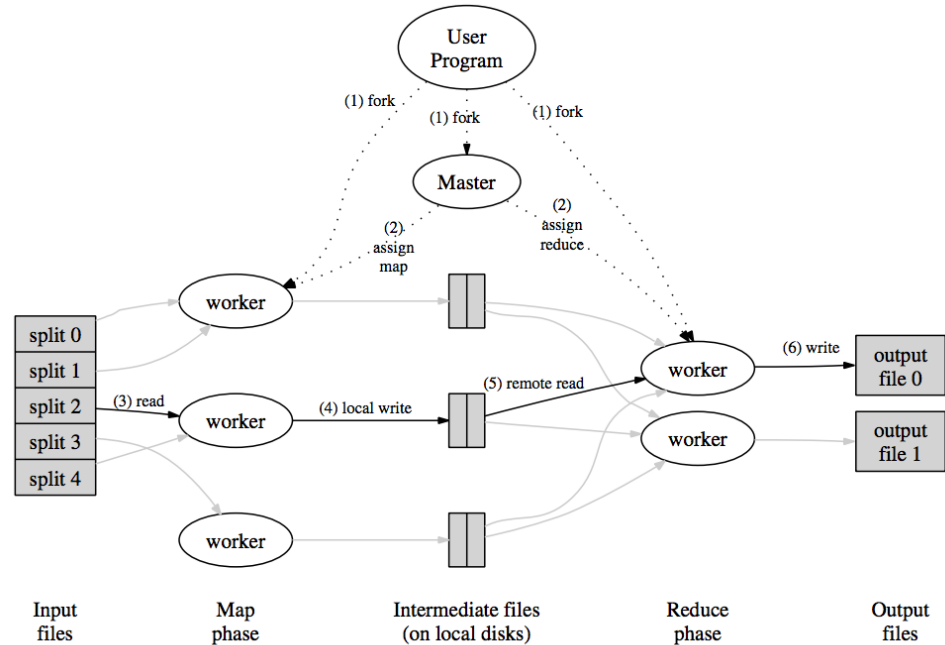
REDUCE

```
function Reduce(Word w, Iterator intermediates):  
    int count= 0  
    for each int value in intermediates:  
        count += value  
    Emit(w, count)
```

# MAP-REDUCE

BIG INSIGHT ISN'T  
MAP / REDUCE METHODS,  
BUT THEIR **SIMPLICITY**  
AND THE **ARCHITECTURE**  
**AROUND THEM**

PROVIDES **SCALABILITY**  
AND **FAULT-TOLERANCE**  
FOR BIG DATA  
PROCESSING JOBS



# DEALING WITH ERRORS

A photograph of a server room with rows of server racks. The room is brightly lit with overhead lights, and the racks are filled with various server components. The perspective is from an elevated position, looking down at the racks.

## SERVER FAILURE

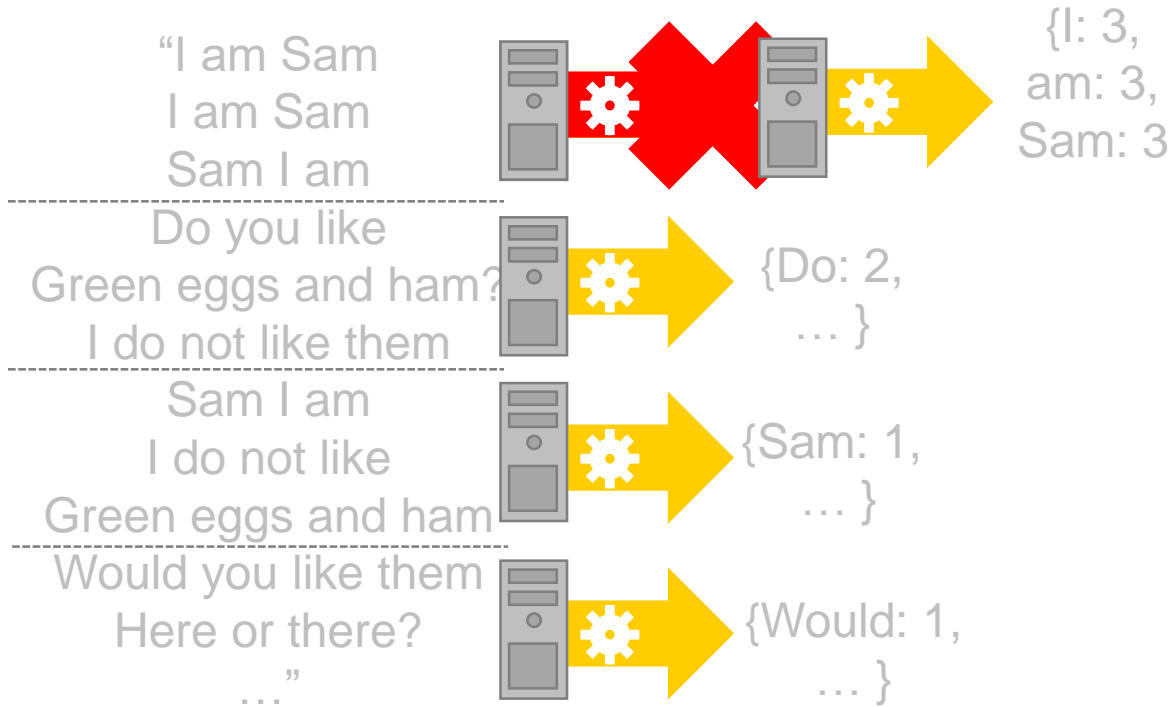
1 server fails every 3 years

→ 10K nodes see 10 faults/day

## STRAGGLERS

Nodes are slow or unresponsive

# JUST LAUNCH A REPLACEMENT

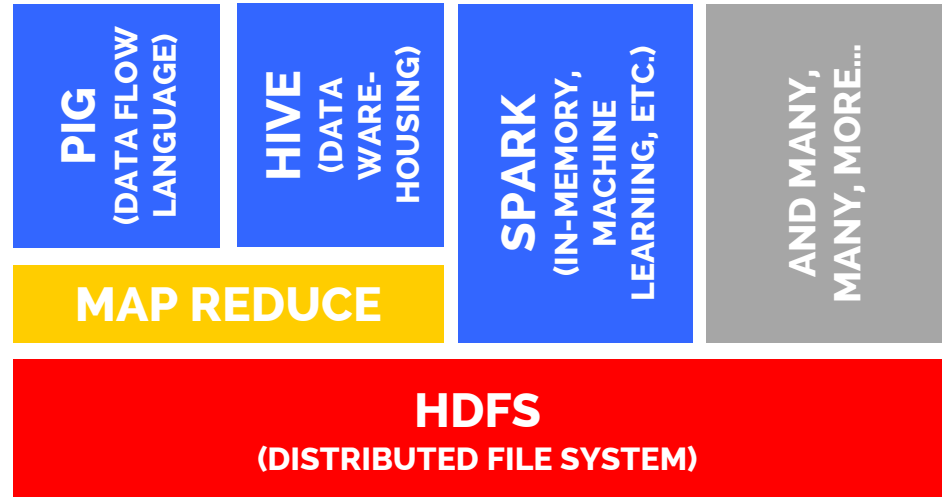


# APACHE HADOOP

**OPEN-SOURCE** DISTRIBUTED FILE SYSTEM  
+ MAP REDUCE **AND MORE**

**INSPIRED** BY GOOGLE'S SYSTEMS

**MANY DATA PROCESSING**  
PIPELINES NOW BUILT  
ON HADOOP INFRASTRUCTURE



# SOME OPTIONS FOR SPECIFYING BIG DATA PROCESSING OPERATIONS

**WRITE YOUR OWN** MAP-REDUCE METHODS

USE A QUERY LANGUAGE LIKE **APACHE PIG**  
THAT CAN COMPILE DOWN TO MAP REDUCE-  
STYLE DISTRIBUTED COMPUTATIONS

```
a = load '/documents';  
b = foreach a generate flatten(TOKENIZE((chararray)$0)) as word;  
c = group b by word;  
d = foreach c generate COUNT(b), group;  
store d into '/pig_wordcount';
```

# BENEFITS AND CHALLENGES

Data manipulation on clusters is now a **big business**.

There is a **huge library of tools** for querying and processing distributed data.

**BUT...**

Most of these tools are **not** real-time or interactive.

**WHAT IF YOU NEED TO INTERACTIVELY  
EXAMINE OR VISUALIZE A BIG DATASET?**



# DATA ANALYSIS AT SCALE

CHALLENGES

ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

PARALLELIZING HUMAN INTELLIGENCE

# **STRATEGIES FOR PROVIDING INTERACTIVITY WITH BIG DATA**

## **1. INTERACTIVITY VIA PRECOMPUTATION**

**(AGGREGATE AND THEN INTERACT)**

## **2. VISUALIZATION AS QUERY SPECIFICATION**

**(LEAVE BIG DATA ON THE SERVERS)**

## **3. SAMPLE INTERACTIVELY**

**(APPROXIMATE FIRST THEN REFINE)**

# **STRATEGIES FOR PROVIDING INTERACTIVITY WITH BIG DATA**

**PARALLELIZE**

**1. INTERACTIVITY VIA PRECOMPUTATION**

**(AGGREGATE AND THEN INTERACT)**

**2. VISUALIZATION AS QUERY SPECIFICATION**

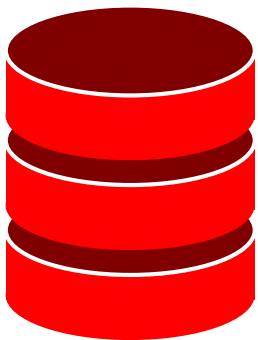
**(LEAVE BIG DATA ON THE SERVERS)**

**3. SAMPLE INTERACTIVELY**

**(APPROXIMATE FIRST THEN REFINE)**

# **SAMPLING FOR INTERACTION**

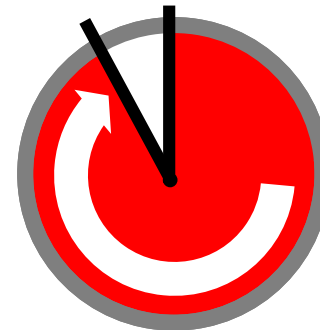
# STANDARD QUERY



**BIG  
DISTRIBUTED  
DATABASE**

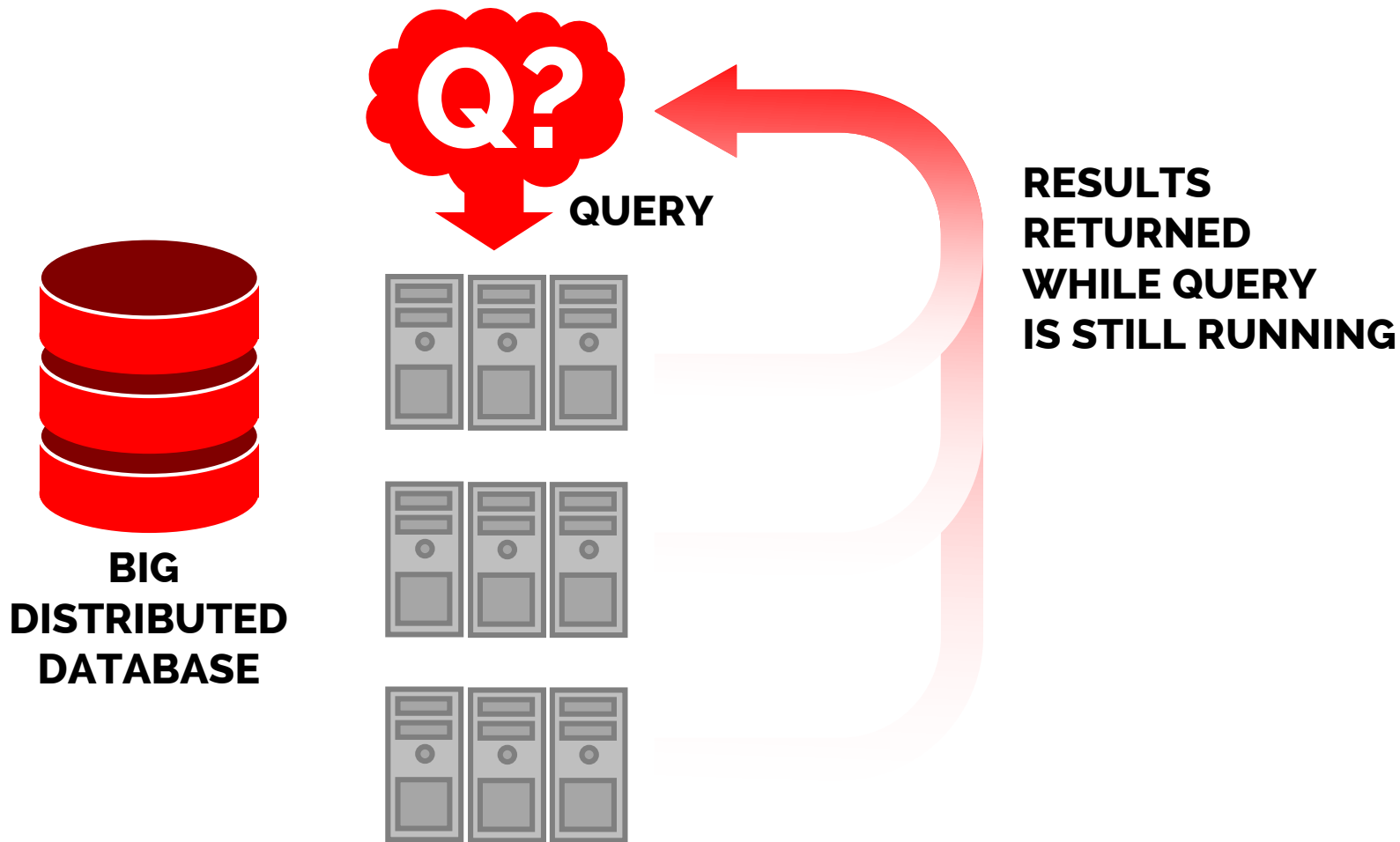


**ALL RESULTS  
RETURNED AT  
THE END**

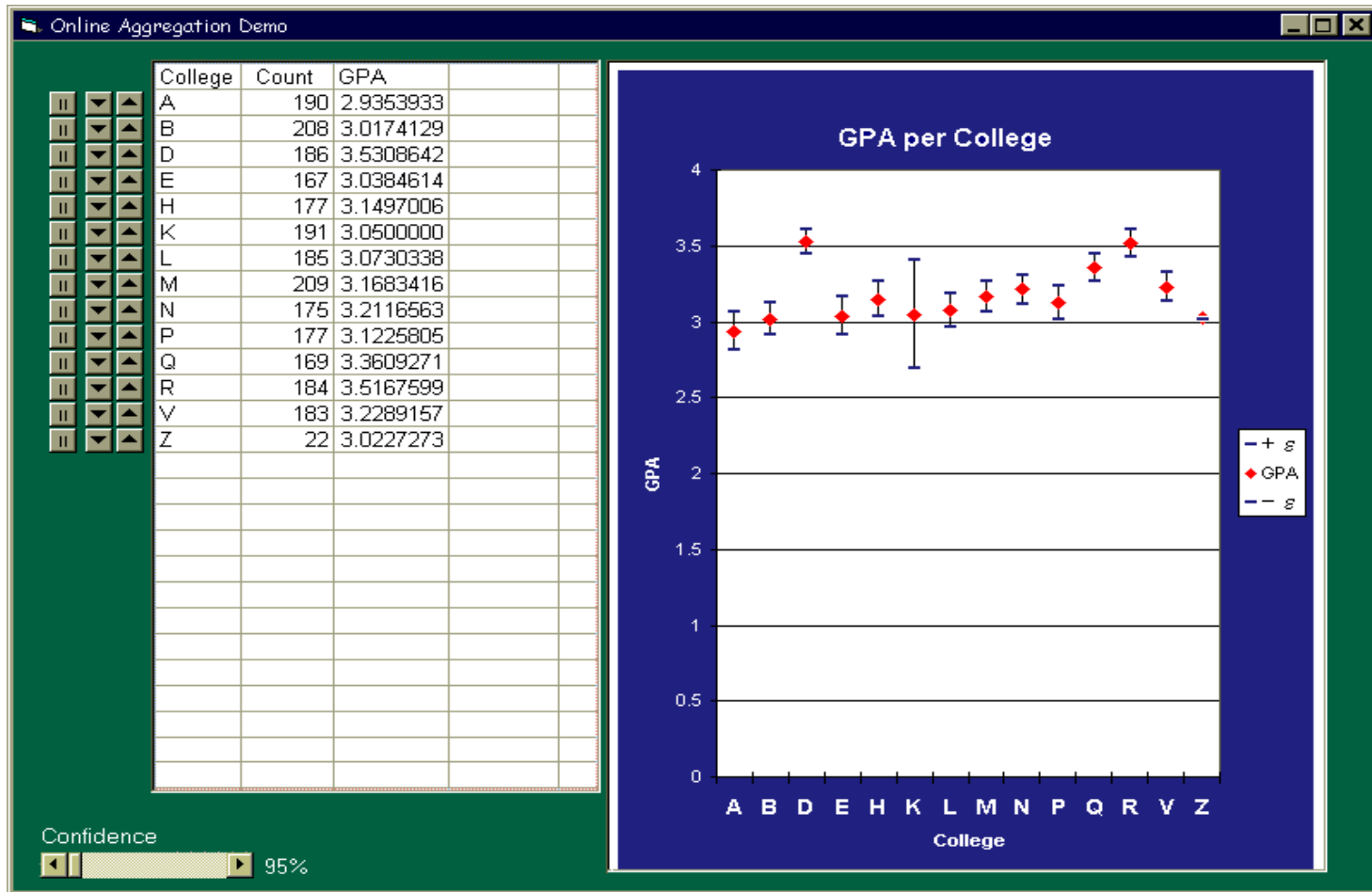


**RESULTS**

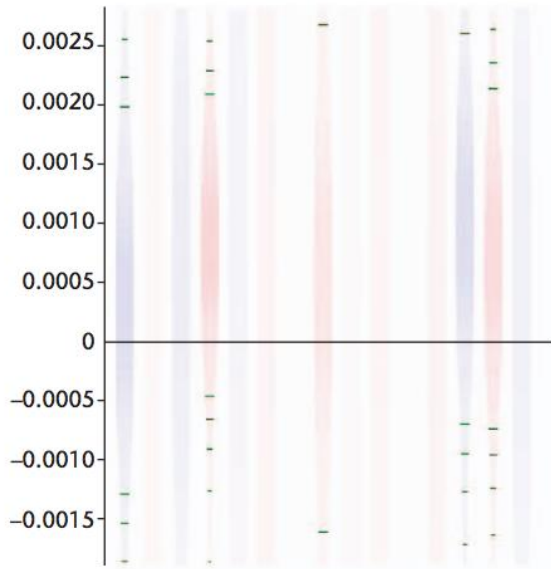
# INTERACTIVE SAMPLING



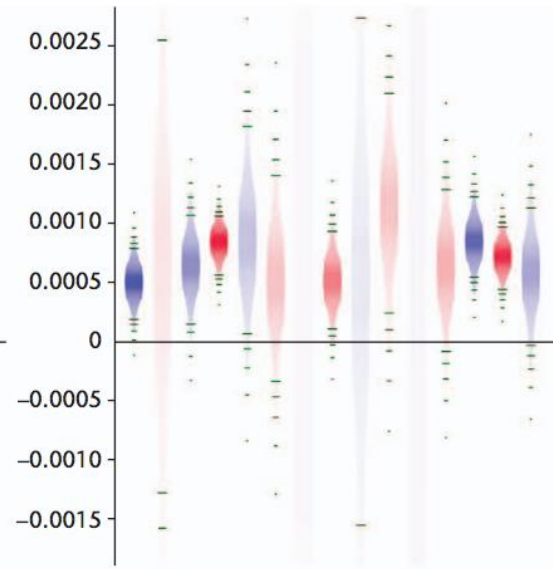
# INTERACTIVE SAMPLING



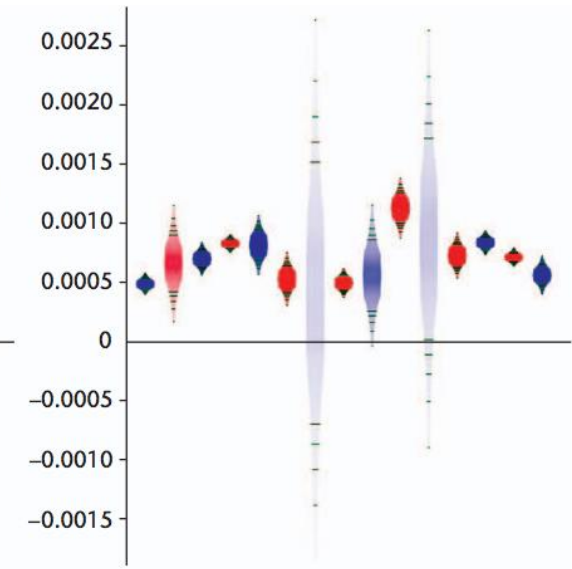
# INTERACTIVE SAMPLING



(a)



(b)



(c)



# INTERACTIVE SAMPLING

**BUT...**

**MOST BACKENDS AREN'T DESIGNED TO  
RETURN PROGRESSIVE RESULTS**

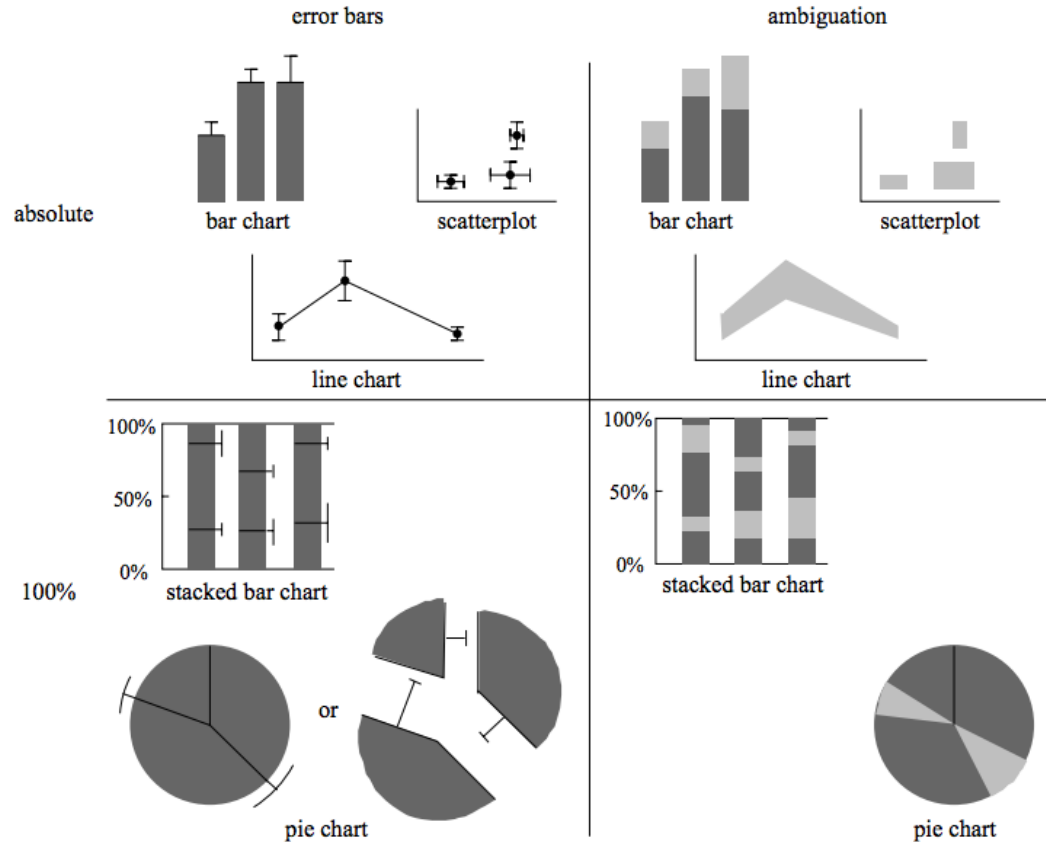
**WE NEED GOOD SAMPLING DISTRIBUTIONS FOR EACH  
FIELD TO PRODUCE MEANINGFUL INTERMEDIATE RESULTS**

**HOW BEST TO VISUALIZE UNCERTAINTY?**

**HOW WELL CAN PEOPLE INTERPRET PARTIAL RESULTS?**

**THIS IS STILL A VERY OPEN RESEARCH AREA!**

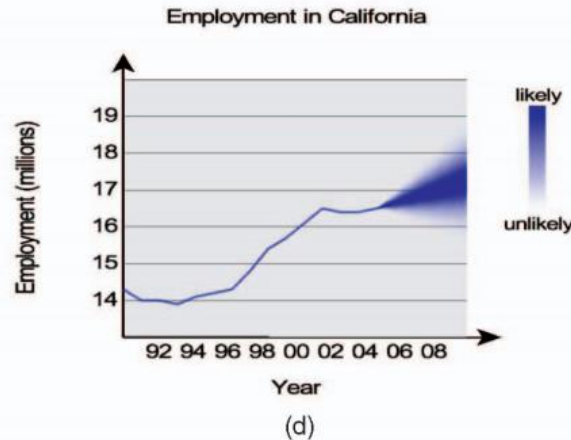
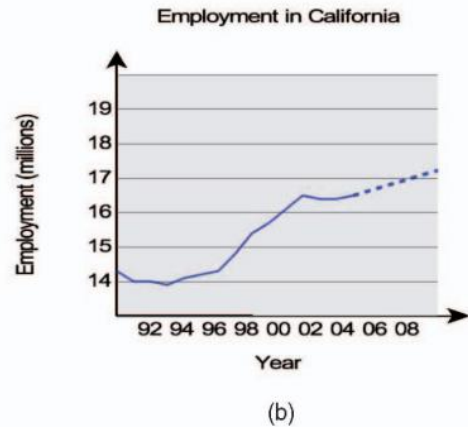
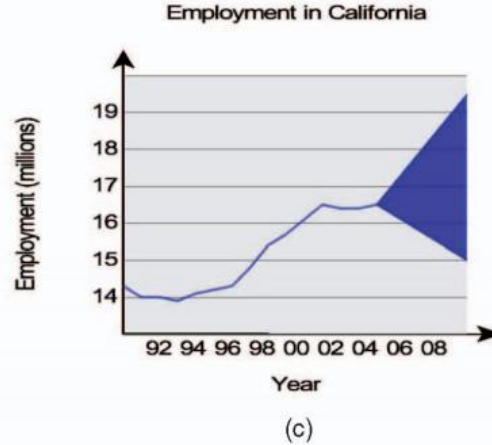
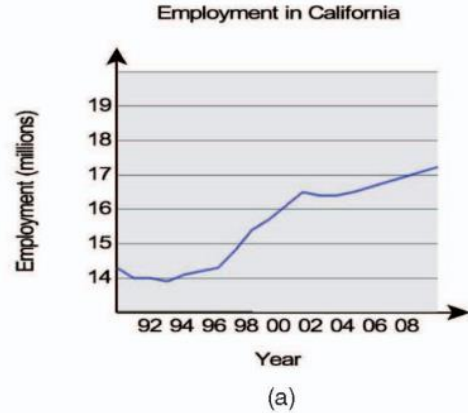
# HOW TO SHOW UNCERTAINTY?



[Olston & Mackinlay, 2002]

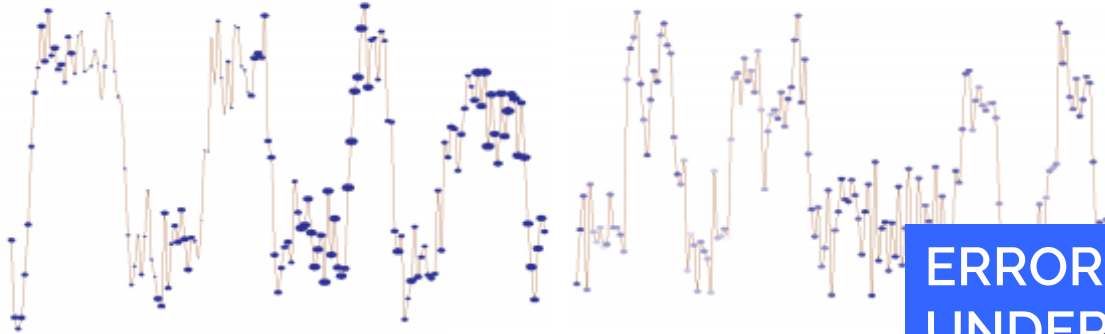
Figure 1: Error bars and ambiguity applied to some common chart types.

# HOW TO SHOW UNCERTAINTY?



[Streit, Pham, & Brown 2008]

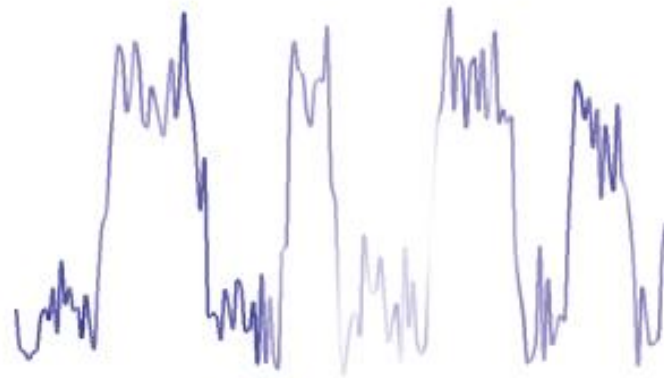
# HOW TO SHOW UNCERTAINTY?



- High uncertainty
- Low uncertainty

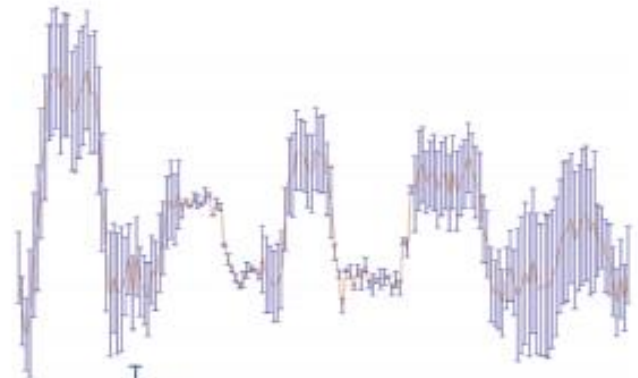
a.

ERROR BARS CONSISTENTLY UNDERPERFORMED



- High uncertainty
- Low uncertainty

c.

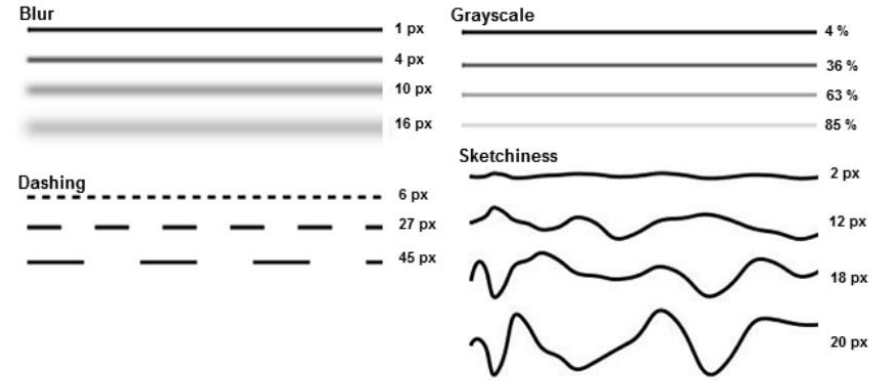
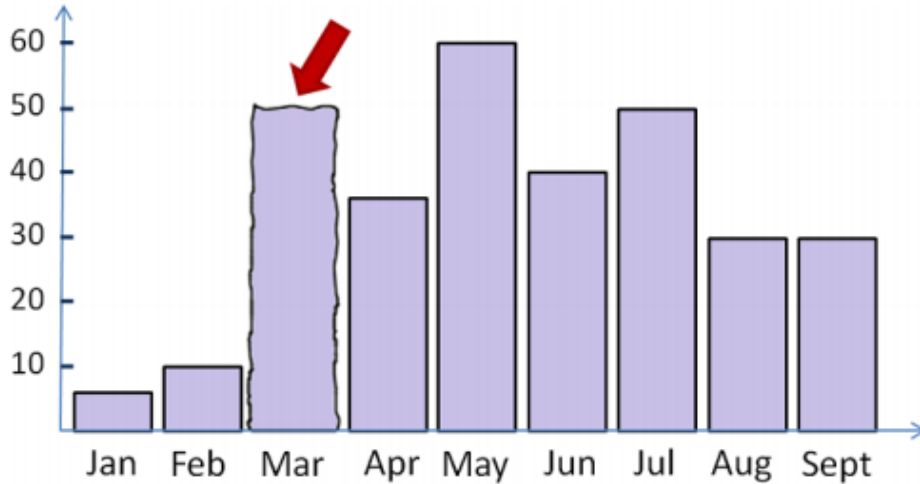


- High uncertainty
- Low uncertainty

d.

[Sanyal, et al. 2009]

# HOW TO SHOW UNCERTAINTY?



[Boukhelifa, et al. 2012]

PEOPLE DON'T ALWAYS  
INTERPRET THESE AS SHOWING  
UNCERTAINTY

# **A FEW INTERESTING RESEARCH PROTOTYPES**

- Datasets
  - Hamster
  - Homes
  - NASDAQ
  - Whifull
- 
- Hello
- hello
  - White House visitors
  - test
  - new page
  - stock market
  - play
  - homes
  - stock analysis**

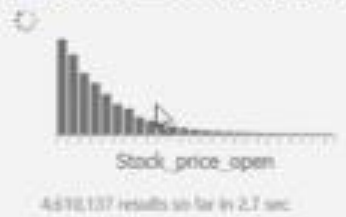


```
var stock = NASDAQ.Where(s=>s.Stock_price_open<100);
```

Stock_price_open	Stock_price_close	Date	Stock_symbol	Exchange	Stock_price_high	Stock_price_low	Stock_volume	Stock_price_adj	close
2.55	2.67	12/9/2009 12:00:00 AM	ABXA	NASDAQ	2.77	2.5	155000	2.67	
2.71	2.88	12/8/2009 12:00:00 AM	ABXA	NASDAQ	2.74	2.52	191700	2.55	
2.65	2.71	12/7/2009 12:00:00 AM	ABXA	NASDAQ	2.76	2.65	174000	2.75	
2.60	2.65	12/6/2009 12:00:00 AM	ABXA	NASDAQ	2.66	2.58	230900	2.65	
2.55	2.6	12/5/2009 12:00:00 AM	ABXA	NASDAQ	2.62	2.51	360900	2.6	
2.41	2.55	12/2/2009 12:00:00 AM	ABXA	NASDAQ	2.59	2.4	287700	2.55	
2.35	2.4	12/1/2009 12:00:00 AM	ABXA	NASDAQ	2.44	2.27	653000	2.4	
2.26	2.25	11/30/2009 12:00:00 AM	ABXA	NASDAQ	2.36	2.11	446100	2.25	
2.35	2.35	11/27/2009 12:00:00 AM	ABXA	NASDAQ	2.42	2.5	121200	2.35	
2.48	2.45	11/25/2009 12:00:00 AM	ABXA	NASDAQ	2.49	2.4	77500	2.45	

Items 1-10 of 6,435,366 results in 1.5 sec

```
var opening = stock.Viz().Histogram(s=>s.Stock_price_open);
```



**STILL USING SIMPLE VISUALS**

**NO UNCERTAINTY INFO**

**NOTE: ANALYSIS NOTEBOOKS AND PROVENANCE**

# DATA ANALYSIS AT SCALE

CHALLENGES

ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

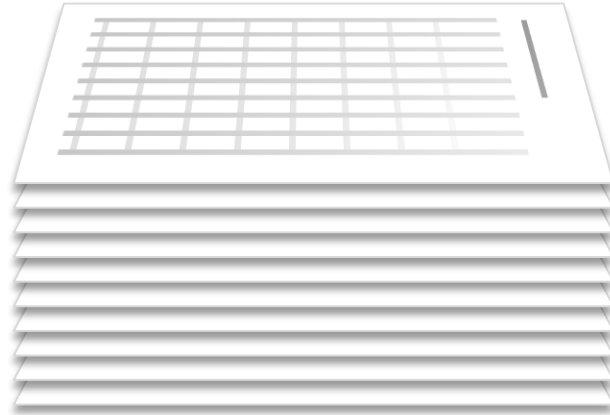
PARALLELIZING HUMAN INTELLIGENCE



**HOW CAN WE LEVERAGE MULTIPLE  
PEOPLE TO EXPEDITE ANALYSIS?**



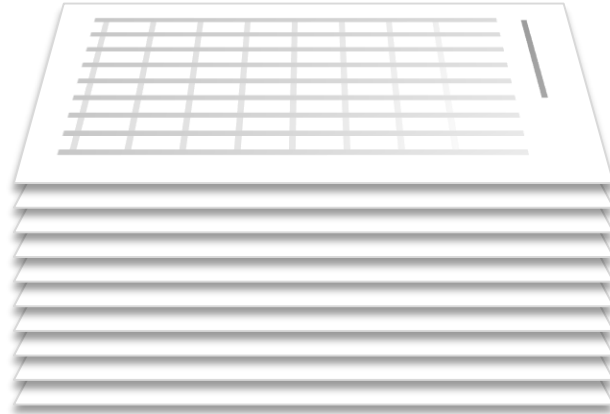
Analyst



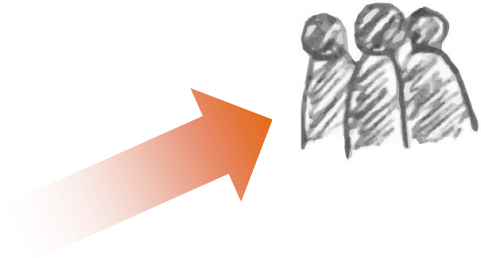
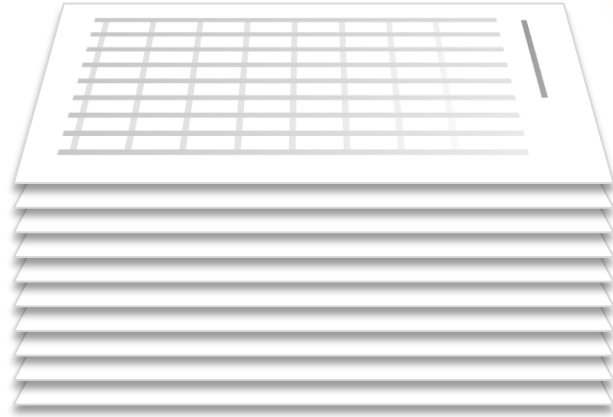
CollegeRankings2013.csv



Analyst



“Can I enlist others to help make sense of my data?”



Crowd

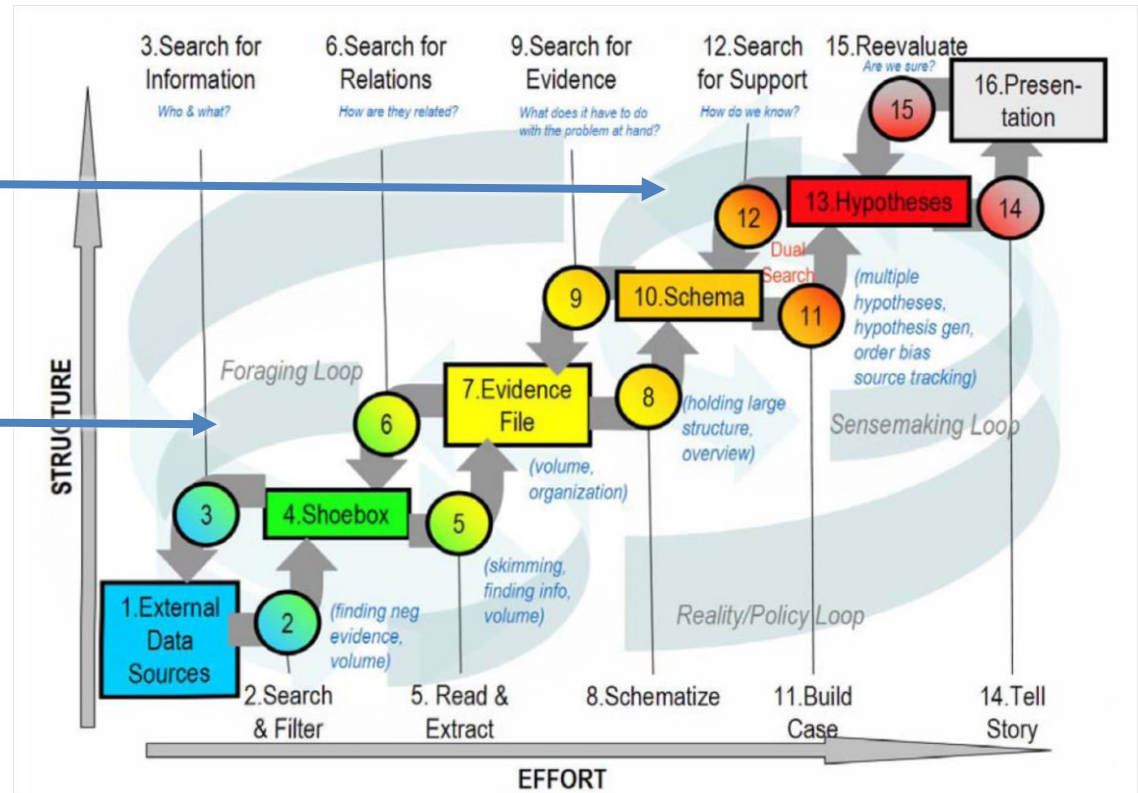


MANY IMPORTANT ANALYSIS TASKS REQUIRE  
HUMAN INTELLIGENCE BUT LEND THEMSELVES  
WELL TO PARALLELIZATION

# MANY IMPORTANT ANALYSIS TASKS REQUIRE HUMAN INTELLIGENCE BUT LEND THEMSELVES WELL TO PARALLELIZATION

Sensemaking Loop

Foraging Loop



[Pirolli & Card 2005]

# MANY EYES

## Explore

- Visualizations
- Data sets
- Comments
- Topic centers

## Participate

- Create a visualization
- Upload a data set
- Create a topic center
- Register

## Learn more

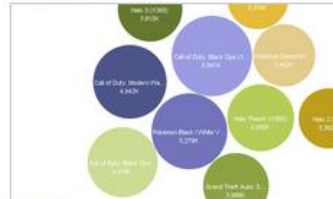
- Quick start
- Visualization types
- About Many Eyes
- Privacy
- Blog

Visualization ▾

Search

## Try our featured visualizations

### Game Sales During First Week of Release



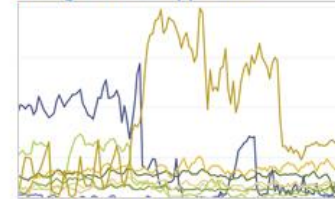
Top 10  
by EmersonM

### Global Surface Temperature



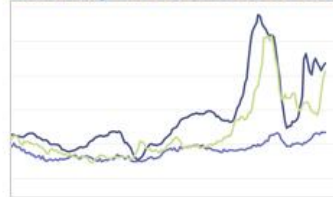
1880-2009 - comparison to global mean.  
by cliffsnellgrove

### Dating Services App Rank



Apr 2011 to Sept 2011  
by kshonbeck

### Meat, Dairy and Cereal Price Indices



1990-2010  
by Anonymous

### World Cancer Drug Market

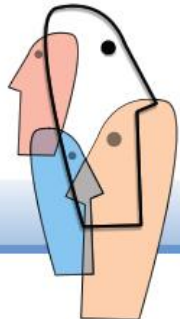


By Product Type, 2005-2015  
by Elsevier Global Medical News

### Steve Jobs Stanford Commencement Address



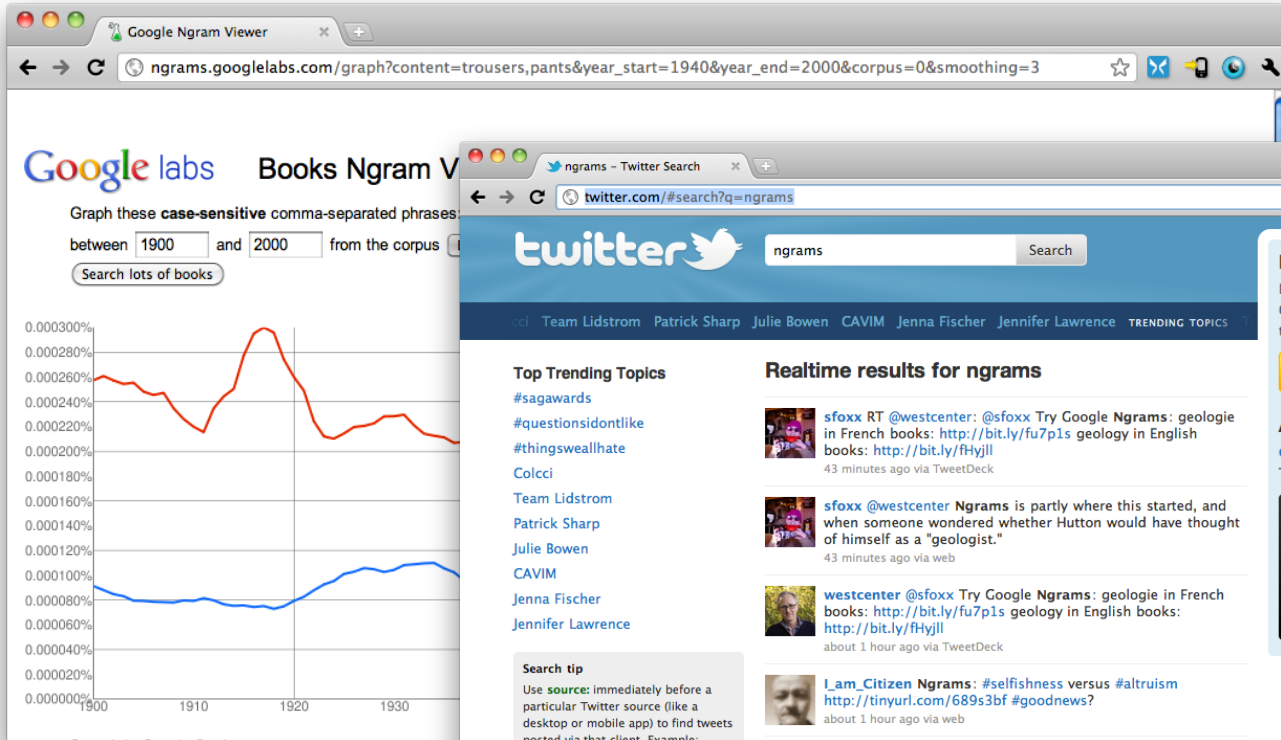
Selection from his address.  
by nrcamp



An experiment brought to you by IBM Research and the IBM Cognos software group

[Viégas, et al. 2007, 2008]

# GOOGLE BOOKS N-GRAMS



Twitter search results for "#ngrams". The search bar contains "ngrams" and the search button is labeled "Search". The results are categorized into "Top Trending Topics" and "Realtime results for ngrams".

**Top Trending Topics**

- #sagawards
- #questionsdontlike
- #thingsweallhate
- Colcci
- Team Lidstrom
- Patrick Sharp
- Julie Bowen
- CAVIM
- Jenna Fischer
- Jennifer Lawrence

**Realtime results for ngrams**

- sfoxx** RT @westcenter: @sfoxx Try Google Ngrams: geologie in French books: <http://bit.ly/fu7p1s> geology in English books: <http://bit.ly/fHyjll> 43 minutes ago via TweetDeck
- sfoxx** @westcenter Ngrams is partly where this started, and when someone wondered whether Hutton would have thought of himself as a "geologist." 43 minutes ago via web
- westcenter** @sfoxx Try Google Ngrams: geologie in French books: <http://bit.ly/fu7p1s> geology in English books: <http://bit.ly/fHyjll> about 1 hour ago via TweetDeck
- I\_am\_Citizen** Ngrams: #selfishness versus #altruism <http://tinyurl.com/689s3bf> #goodnews? about 1 hour ago via web
- I\_am\_Citizen** RT @LondonEvolution: Ngrams: group selection vs. kin selection. <http://bit.ly/ewaAY3> about 1 hour ago via bitly
- gonzo\_pz** Hottest name: [http://ngrams.googlelabs.com/graph?content=Gonzalo%2C+Felipe%2C+Ignacio&year\\_start=1800&ye](http://ngrams.googlelabs.com/graph?content=Gonzalo%2C+Felipe%2C+Ignacio&year_start=1800&ye) about 2 hours ago via web

**Search tip**

Use **source:** immediately before a particular Twitter source (like a desktop or mobile app) to find tweets posted via that client. Example: **weather source:tweetie** will find tweets containing "weather" and entered via Tweetie.

**New to Twitter?**

Easy, free, and instant updates. Get access to the information that interests you most.

**Sign Up >**

**A #NewTwitter**

Catch a glimpse of the new Twitter.com.



# **CROWDSOURCING DATA ANALYSIS**

**DATA COLLECTION & CITIZEN SCIENCE**

**ANALYSIS COMPETITIONS**

**“MICROWORK” AND TASK MARKETS**

**COLLABORATION TOOLS FOR  
ANALYSTS**

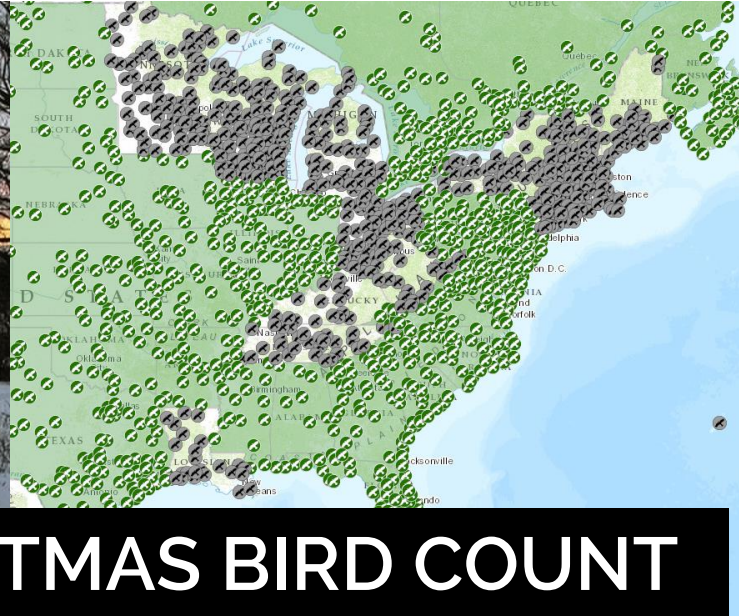
# CITIZEN SCIENCE

## DATA COLLECTION



## CREEK WATCH

[IBM]



## CHRISTMAS BIRD COUNT

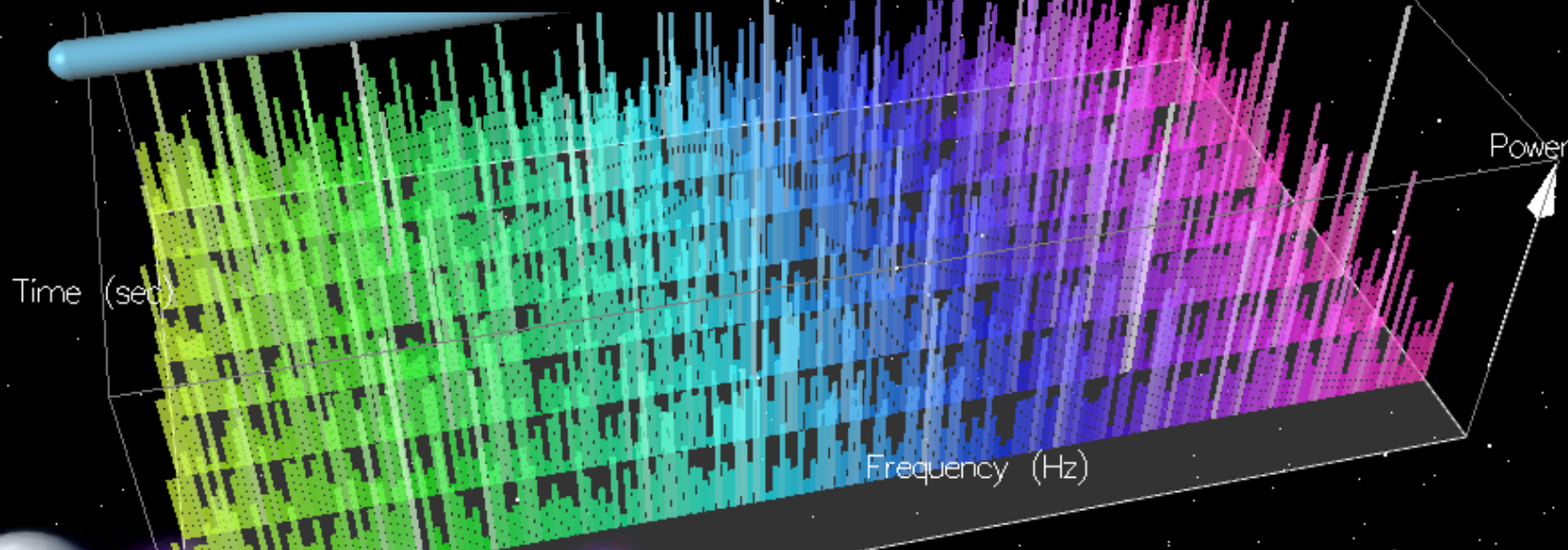
4,000  
Creek Watch  
users

in over  
25  
countries



# CITIZEN SCIENCE

## DATA PROCESSING

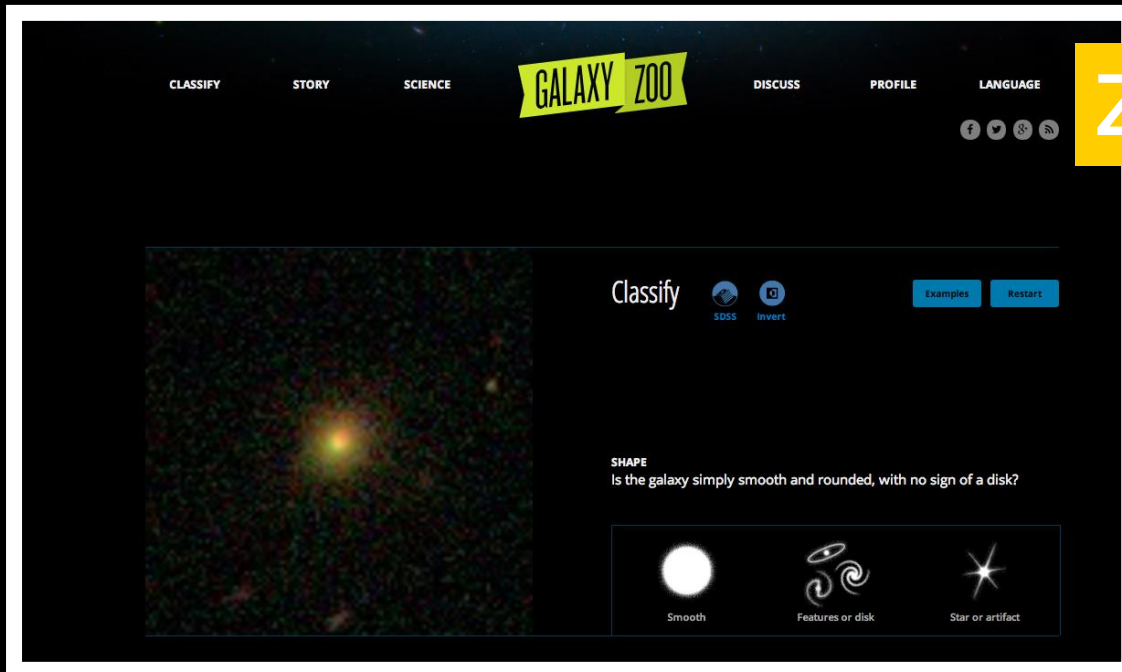


**SETI@home**  
The Search for Extraterrestrial Intelligence

**SETI@Home**

# CITIZEN SCIENCE

## HUMAN VISION & PROBLEM SOLVING



ZOONIVERSE

# CITIZEN SCIENCE

## HUMAN VISION & PROBLEM SOLVING

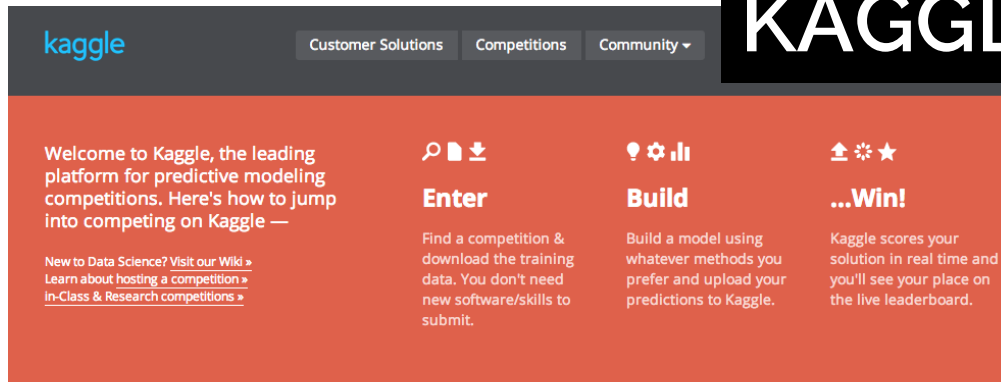
FOLD.IT

The screenshot shows the Fold.it game interface. At the top left, it says "Pull Mode". In the top right corner, a status box displays: "Rank: 4", "Score: 9587.911", "Soleist", "Beginner Puzzle: Streptococcal Protein", "Expires 3/27/2013 17:00 MZ (2 days, 17 hours)", and "No bonuses or conditions". The main area features a 3D model of a protein structure, with green and orange ribbons representing different parts of the protein. A vertical toolbar on the left contains icons for "K", "O", "P", "R", "O", "C", "H", "D". At the bottom, there is a control panel with various actions: "Shake", "Mutate", "Wiggle All", "Wiggle Backbone", "Wiggle Sidechains", "Help", "Glossary", "Freeze Protein", "Remove Bands", "Disable Bands", "Reset Structures", "Reset Puzzle", and "Align Guide". The bottom of the screen shows navigation tabs: "Actions", "Undo", "Social", "Modes", "Behavior", "View", and "Menu". In the bottom right corner, there are chat and notification options: "Chat - Group", "Chat - Puzzle", "Chat - Global", "Notifications", and "auto show" buttons.



# ANALYSIS COMPETITIONS

# KAGGLE



The screenshot shows the Kaggle homepage with a navigation bar containing 'kaggle', 'Customer Solutions', 'Competitions', and 'Community'. The main content area is divided into three columns: 'Enter', 'Build', and 'Win!'. The 'Enter' column includes a search icon and text about finding competitions. The 'Build' column includes a gear icon and text about building models. The 'Win!' column includes an award icon and text about winning prizes.











# NETFLIX PRIZE

## Active Competitions

## Active Competitions

### All Competitions

		<b>Tradeshift Text Classification</b> Classify text blocks in documents	27 days 144 teams \$5,000
		<b>American Epilepsy Society Seizure Prediction ...</b> Predict seizures in intracranial EEG recordings	34 days 279 teams \$25,000
	<b>AfSIS</b>	<b>Africa Soil Property Prediction Challenge</b> Predict physical and chemical properties of soil using spectral measurements	7.4 days 1219 teams \$8,000
		<b>CIFAR-10 - Object Recognition in Images</b> Identify the subject of 60,000 labeled images	4.4 days 224 teams Knowledge
		<b>Learning Social Circles in Networks</b> Model friend memberships to multiple circles	14 days 167 teams Knowledge

# MICROWORK PLATFORMS

SITES WHERE WORKERS PERFORM SMALL PIECES OF WORK ("TASKS") - USUALLY IN EXCHANGE FOR SMALL FINANCIAL REWARDS.

**amazon** **mechanicalturk**<sup>™</sup>  
Artificial Artificial Intelligence

 **CrowdFlower** *mobileworks*

# MICROWORK

USING APIS – DEVELOPERS CAN WRITE  
PROGRAMS THAT INCORPORATE  
HUMAN JUDGEMENT

“HUMAN COMPUTATION”



# **APPLYING MICROWORK TO DATA ANALYSIS**

# **CROWDSOURCING LOW-LEVEL ANALYSIS**

**DATA COLLECTION AND DATA ENTRY**

**LABELING**

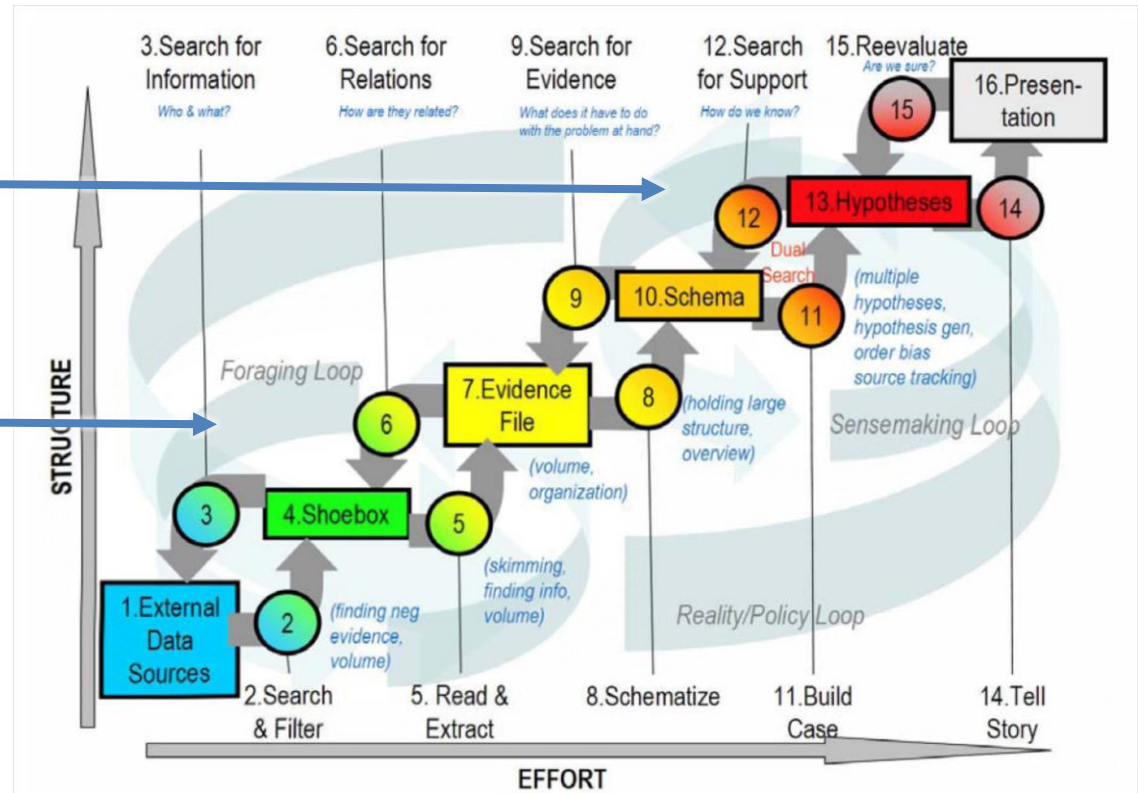
**DATA CLEANING**

**SENTIMENT ANALYSIS**

# MANY IMPORTANT ANALYSIS TASKS REQUIRE HUMAN INTELLIGENCE BUT LEND THEMSELVES WELL TO PARALLELIZATION

Sensemaking Loop

Foraging Loop

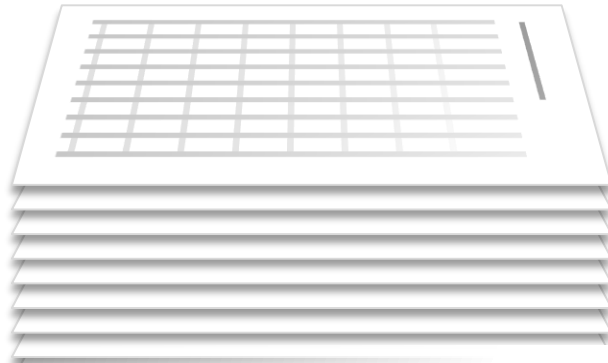


[Pirolli & Card 2005]

# CROWDSOURCING HIGHER-LEVEL ANALYSIS TASKS



Analyst



“Can I screen this dataset to **quickly**  
find the **most interesting** parts?”

# A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



Data

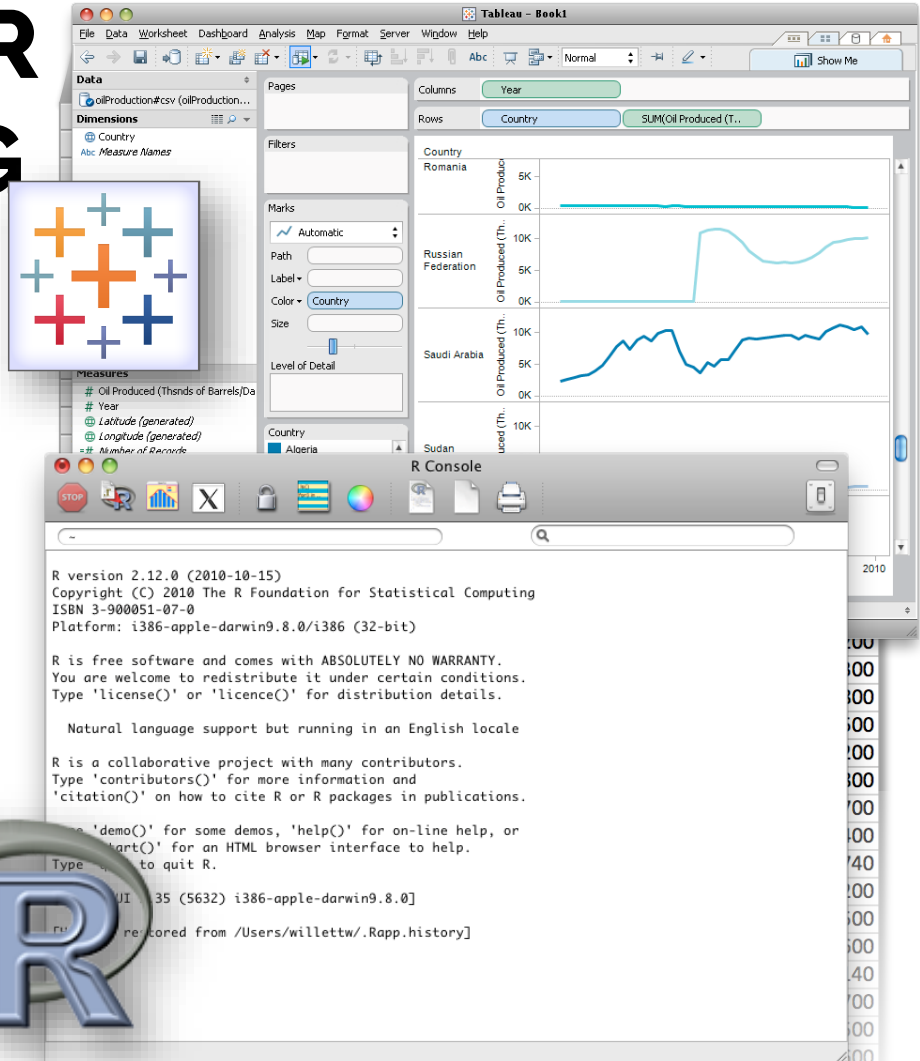


Analyst



Crowd

# A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



Data



Analyst



Crowd



[Willett et al. CHI 2012, VAST 2013]

# A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



Data

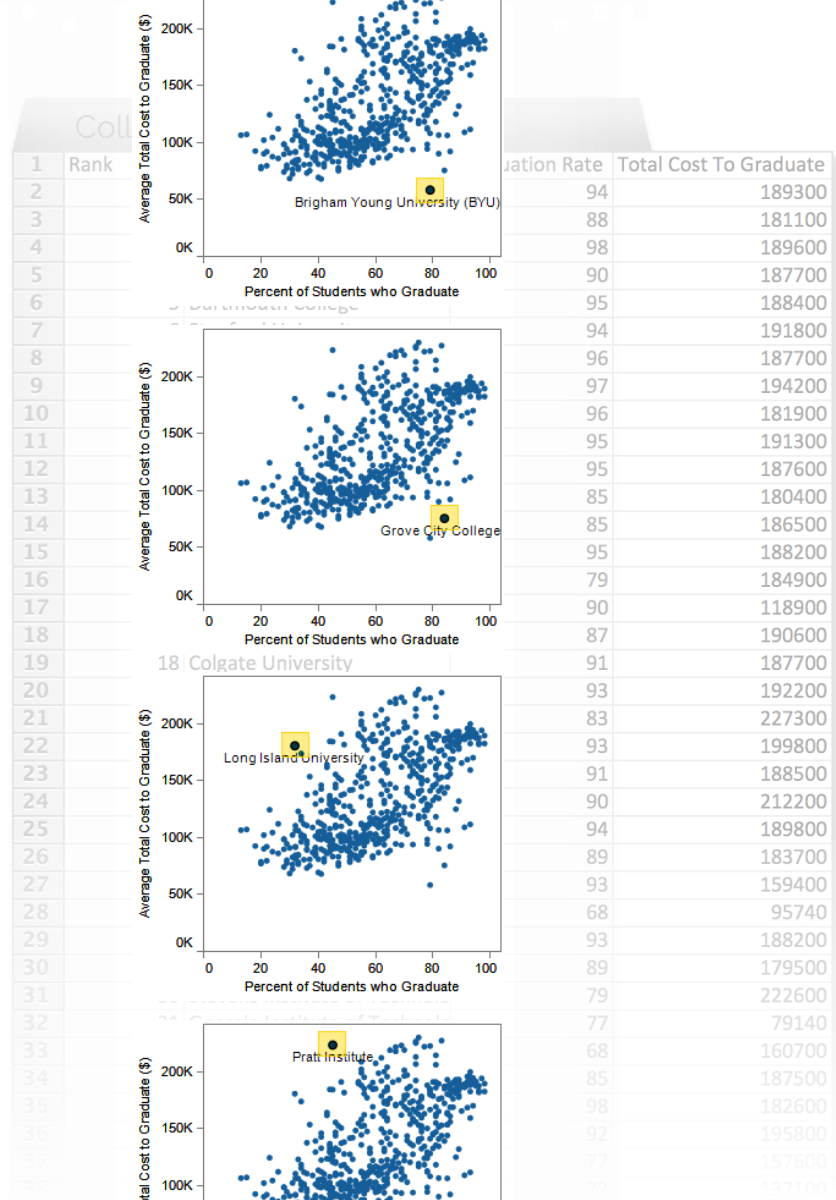


Analyst

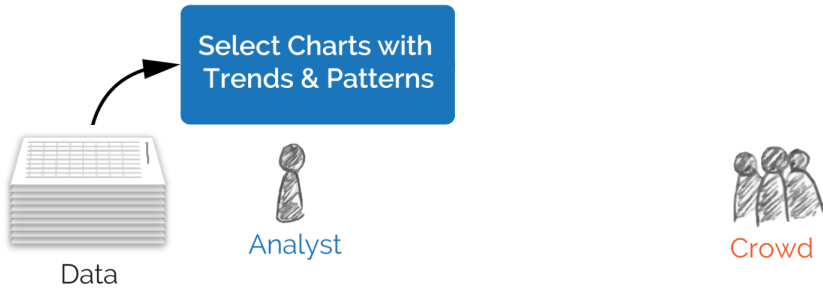


Crowd

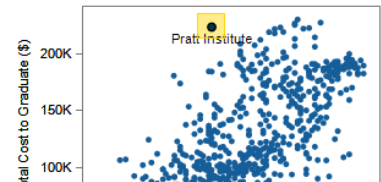
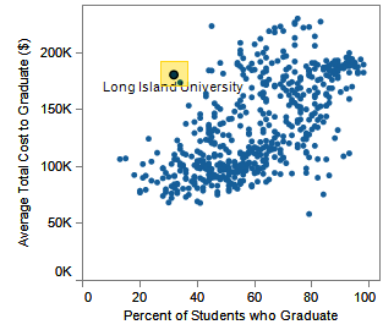
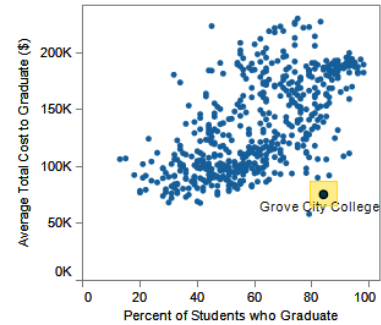
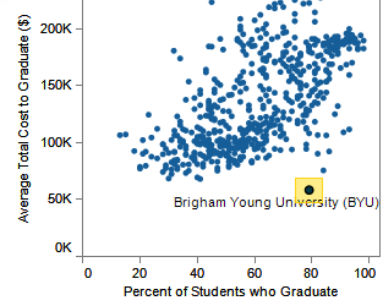
[Willett et al. CHI 2012, VAST 2013]



# A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS

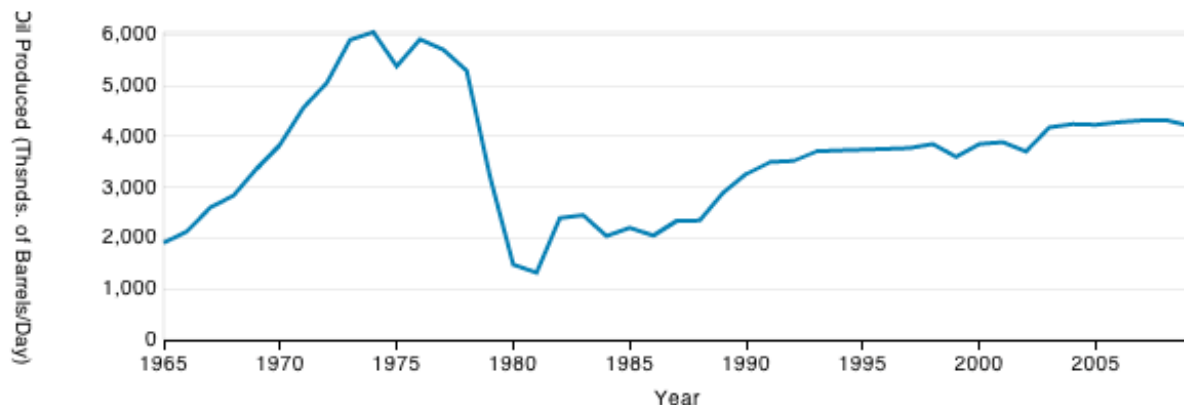


[Willett et al. CHI 2012, VAST 2013]





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

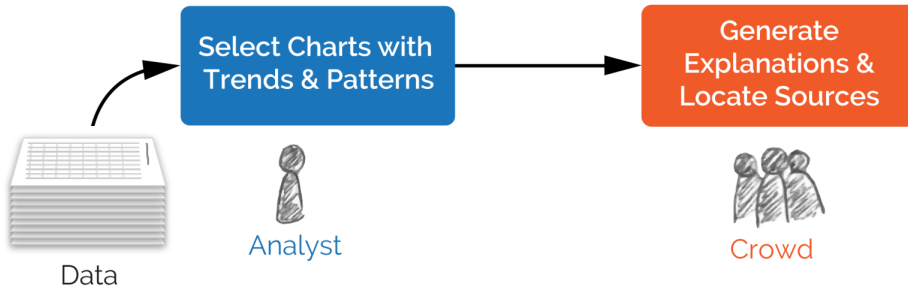


This chart shows **Oil Produced (Thsnds. of Barrels/Day)** by **Year**. The view is filtered by **Country** to show only **"Iran"**.

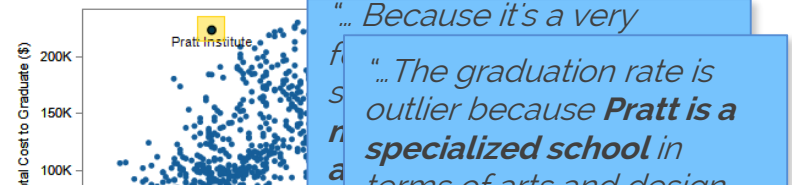
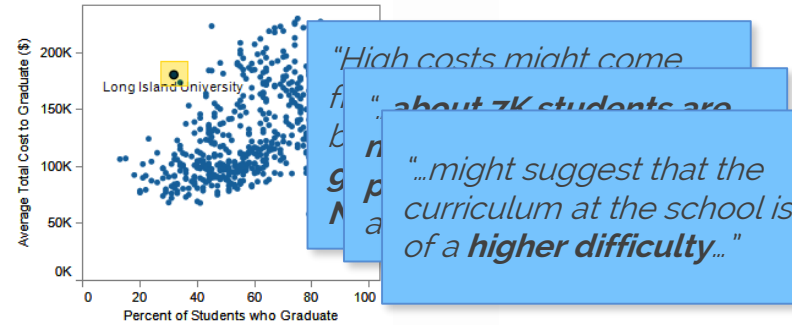
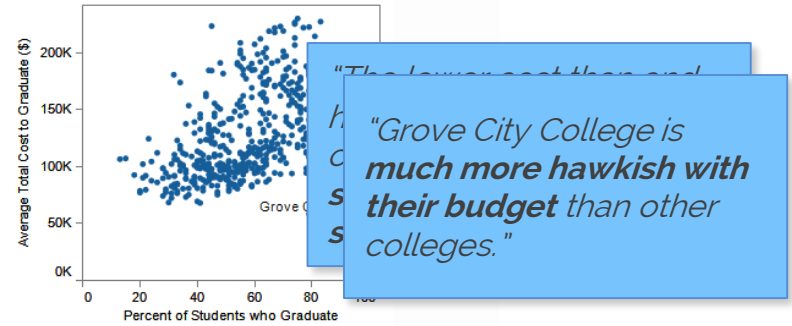
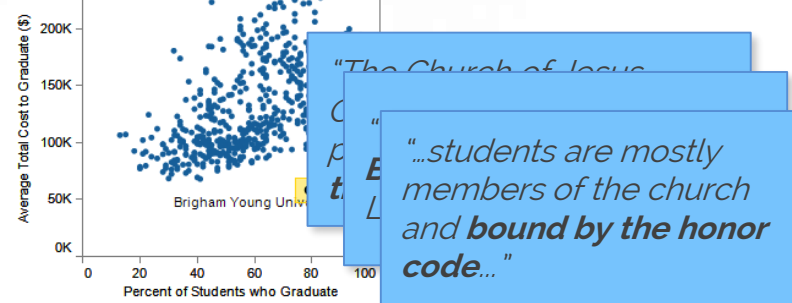
1. Explain **why** the strong **peak or valley** highlighted in the chart might have occurred.

Submit Task

# A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



[Willett et al. CHI 2012, VAST 2013]



**“COULD THIS CREATE  
MORE WORK FOR THE  
ANALYST?”**

# “COULD THIS CREATE MORE WORK FOR THE

## ANALYST?”



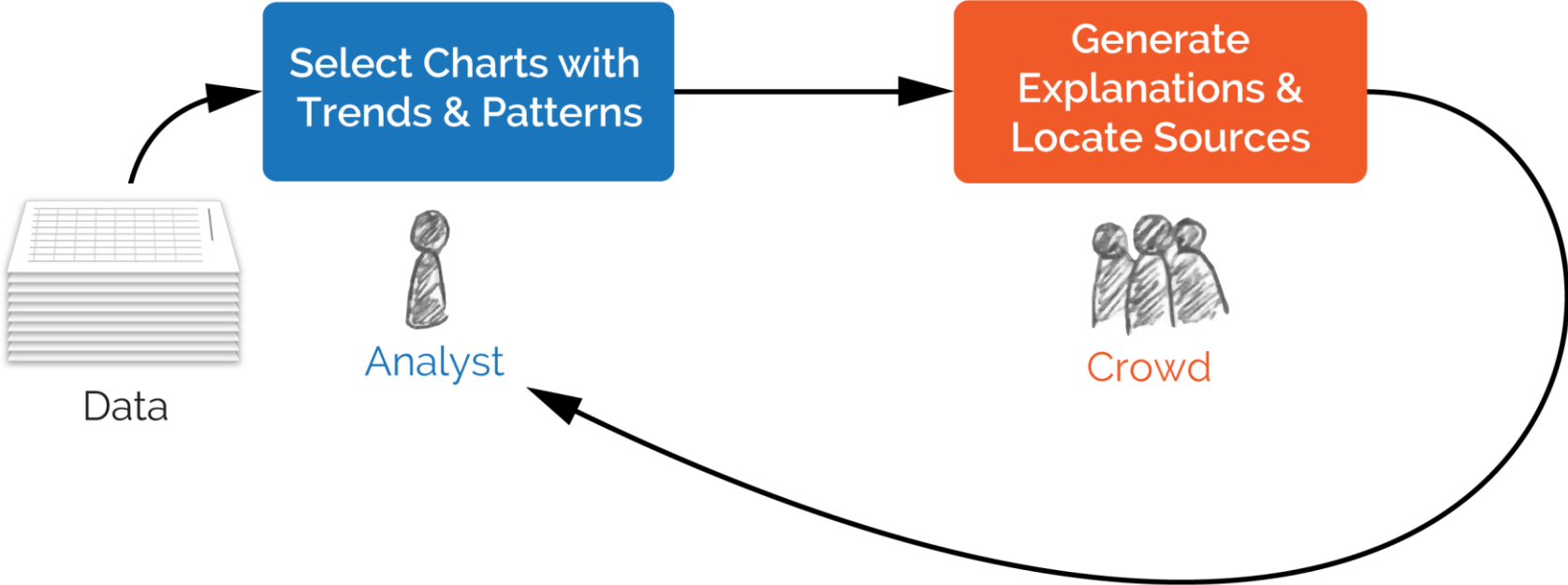
“High costs might come  
from **about 7K students** are  
...might suggest that the  
curriculum at the school is  
of a **higher difficulty**...”

“The lower cost than and  
high cost  
Grove City College is  
**much more hawkish with  
their budget** than other  
colleges.”

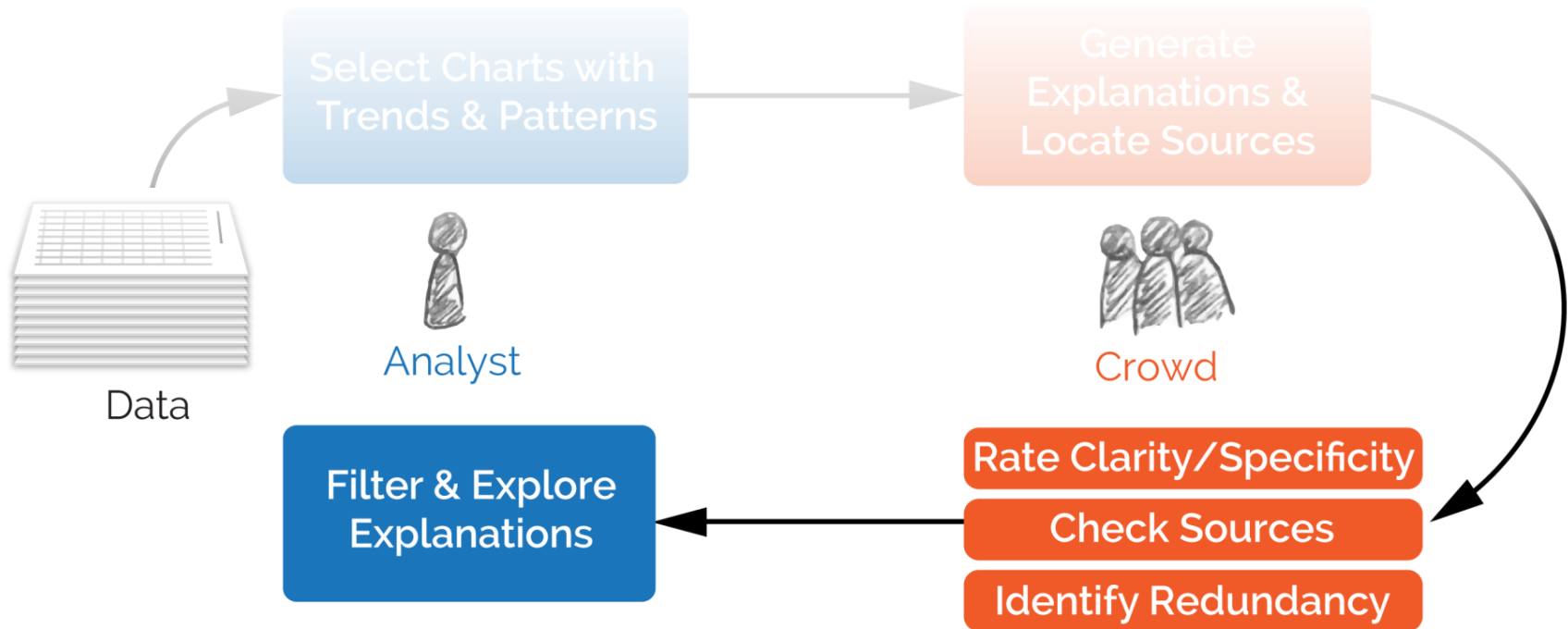
“The Church of Jesus  
Christ of Latter-day Saints  
...students are mostly  
members of the church  
and **bound by the honor  
code**...”

“... Because it's a very  
f  
s  
n  
a  
t  
“... The graduation rate is  
outlier because **Pratt is a  
specialized school** in  
terms of arts and design  
and students...”

# A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



# CROWD-ENABLED EXTENSIONS FOR PROCESSING AND MANAGING RESULTS



# THREE CRITERIA FOR PLAUSIBLE EXPLANATIONS

CLARITY AND SPECIFICITY

PROVENANCE

REDUNDANCY

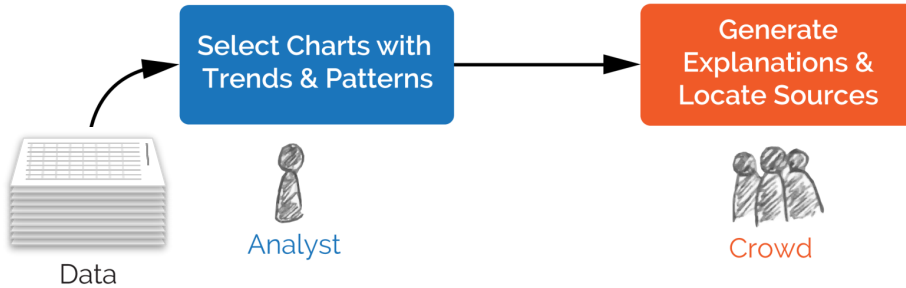
+ AN INTERFACE FOR MANAGING CROWDSOURCED EXPLANATIONS

**CLARITY & SPECIFICITY**



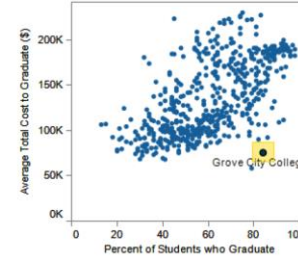
# CLARITY AND SPECIFICITY

## Rating Task



Show Instructions

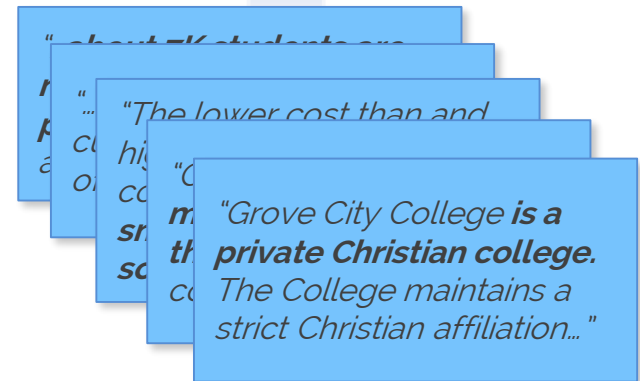
Each of the charts in this hit compares the graduation rate (x-axis) and the total cost to graduate (y-axis) for 554 top US colleges and universities (as ranked by Bloomberg Businessweek in 2010). Each point represents a single college or university.



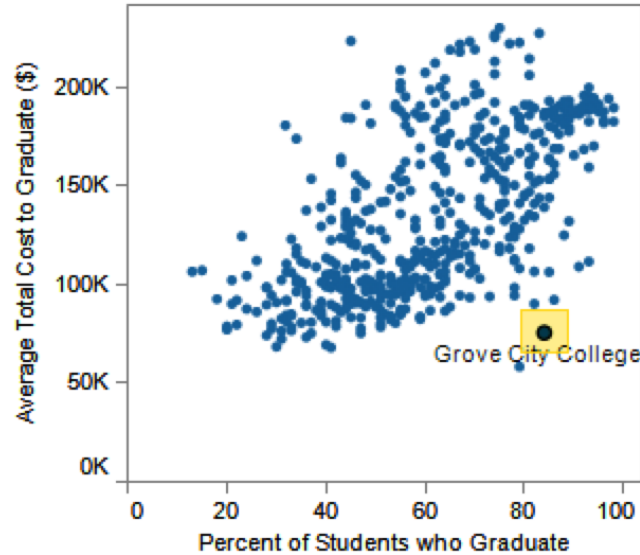
**Prompt:** Explain **why** the **outlier** highlighted in the chart might be different from the other items. (Give **one** specific, well-justified answer.)

**Response R2:** " Grove City College is a private Christian college. The College maintains a strict Christian affiliation, in contrast to many institutions whose religions affiliations have become merely historical in nature. This Christian identity, as well as a heavily politically Conservative identity, on campus may likely attract superior students who would not choose to attend otherwise comparable institutions lacking this culture." (Reference: <http://www.discoverthenetworks.org/Articles/Conservative%20Colleges.htm> )

- Does this response provide an explanation for **why** the highlighted outlier in the chart might have occurred?  
 Yes  No  None Present
- How **clear** and **specific** is the response?  
 Clear/Specific ←  1  2  3  4  5 → (Very Clear/Specific)



Each of the charts in this hit compares the graduation rate (x-axis) and the total cost to graduate (y-axis) for 554 top US colleges and universities (as ranked by Bloomberg Businessweek in 2010). Each point represents a single college or university.



**Prompt:** Explain **why** the **outlier** highlighted in the chart might be different from the other items. (Give **one** specific, well-justified answer.)

**Response R2:** " Grove City College is a private Christian college. The College maintains a strict Christian affiliation, in contrast to many institutions whose religions affiliations have become merely historical in nature. This Christian identity, as well as a heavily politically Conservative identity, on campus may likely attract superior students who would not choose to attend otherwise comparable institutions lacking this culture."(Reference: <http://www.discoverthenetworks.org/Articles/Conservative%20Colleges.htm> )

1. Does this response provide an explanation for **why** the highlighted outlier 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)  
Clear/Specific)

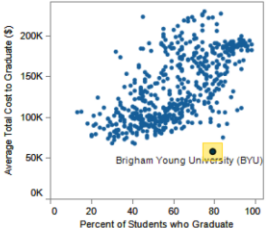
**PROVENANCE**

# PROVENANCE

Explanation Task

proxy.commentspace.net/explainTask?studyName=CrowdAnalytics-CollegeROI-MW3&assignme...

Each of the charts in this hit compares the graduation rate (x-axis) and the total cost to graduate (y-axis) for 554 top US colleges and universities (as ranked by Bloomberg Businessweek in 2010). Each point represents a single college or university.



Brigham Young University (BYU)

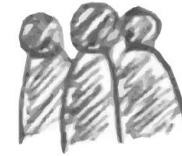
Show Instructions

- Demographic information (Asked on first HIT only).**
  - What is your nationality?
  - What level of schooling have you completed?
  - What is your native language?
  - How comfortable are you with reading charts and graphs?
  - Are you familiar with college rankings?
- What college or university is highlighted in this chart?
- Explain **why** the **outlier** highlighted in the chart might be different from the other items. (Give **one** specific, well-justified answer.)
- Provide the **url of a specific web page** (not just a site) that supports your explanation.

Submit Task

Explanation Task

What are our



workers doing?

# PROVENANCE

Explanation Task

proxy.commentspace.net/explainTask?studyName=CrowdAnalytics-CollegeROI-MW3&assignme...

Each of the charts in this hit compares the graduation rate (x-axis) and the total cost to graduate (y-axis) for 554 top US colleges and universities (as ranked by Bloomberg Businessweek in 2010). Each point represents a single college or university.

Average Total Cost to Graduate (\$)

Percent of Students who Graduate

Brigham Young University (BYU)

Show Instructions

- Demographic information (Asked on first HIT only).**
  - What is your nationality?
  - What level of schooling have you completed?
  - What is your native language?
  - How comfortable are you with reading charts and graphs?
  - Are you familiar with college rankings?
- What college or university is highlighted in this chart?
- Explain **why** the **outlier** highlighted in the chart might be different from the other items. (Give **one** specific, well-justified answer.)
- Provide the url of a specific web page (not just a site) that supports your explanation.

Submit Task

## Explanation Task

brigham young university

https://www.google.fr/search?q=brigham+young+university&aq=brigham+young+university&aq=chrome..69157...

Web Images Maps Shopping Plus Outils de recherche

Environ 15 600 000 résultats (0,30 secondes)

Université Brigham Young

fr.wikipedia.org/wiki/Université\_Brigham\_Young

WIKIPÉDIA L'encyclopédie libre

Mois international de la contribution francophone 2013

Une série d'ateliers est organisée dans la francophonie et durant lesquels des contributeurs expérimentés de Wikipédia, des étudiants et toute personne intéressée à enrichir Wikipédia se rassemblent.

Brigham Young University Admissions

https://saas.byu.edu/tools/b4byu/sites/b4byu/visiting-student/how-much-does-it-cost/

Audience Type | The content on this page applies to a Visiting Student

b4 BYU Brigham Young University Admissions

Why BYU

How to Get In

How to Pay For It

How Much Does It Cost?

Tuition Charges

Part-time Work

More...

Where to Live

New Admits

Contact Us

Facebook Comments

Rate this Page

### How Much Does it Cost?

Counting the cost of BYU

2013-2014 Total Undergraduate Charges

Category	Charge
Public In-State	\$9,447
Public Out-of-State	\$13,971
Private	\$43,214
BYU (USD)	\$1,000

Students at BYU enjoy affordable prices that allow them access to a high quality education at a great price. In 2013, *US News & World Report* ranked BYU in the top 20 for "Great Schools, Great Prices."

### Tuition

As BYU's sponsor, The Church of Jesus Christ of Latter-day Saints subsidizes tuition prices with its members' tithing funds. In principle, each student attending BYU is on scholarship.

# INSTRUMENTING EXPLANATION TASKS

Examine a line chart showing employment change in a US city and briefly explain it.

Requester: visualizationlab.ucb

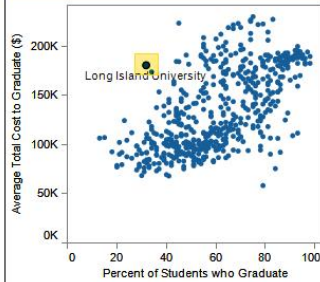
Reward: \$0.40 per HIT

HITs Available: 10

Duration: 30 minutes

Qualifications Required: Location is US

Each of the charts in this hit compares the graduation rate (x-axis) and the total cost to graduate (y-axis) for 554 US colleges and universities (as ranked by Bloomberg Businessweek in 2010). Each point represents a single college or university.



1. What college or university is highlighted in this chart?

2. Explain **why** the strong **outlier** highlighted in the chart might be different from the other items. (Try to give **one** specific, well-justified answer per text box.)

If there are multiple explanations, enter each one in a separate text box.

Using the browser to the right, find text on a web page that justifies each explanation. Select the text and click the "mark as source" button to add it.

Explanation 1

Source:

+ Add Another Explanation



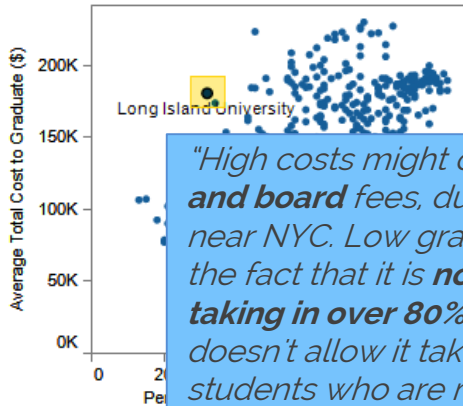
Finished with this HIT? Let someone else do it?

Submit HIT

Return HIT

# PROVENANCE

## Paragraph-level citations



*"High costs might come from its **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."*



### #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](tel:5162992000)

**2013-2014 Tuition**  
\$34,070 tuition and fees

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

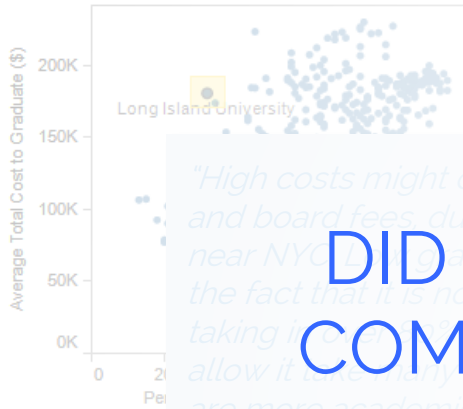
**Admissions**  
rolling admission  
78.8% accepted

[▶ More Information](#)

## Visitation logs

2011-12-11 09:22:04 [google.com](http://google.com)  
2011-12-11 09:22:04 [sqr:help](http://sqr:help)  
2011-12-11 09:23:08 [google.com/search?hl=en&source=hp](http://google.com/search?hl=en&source=hp)  
2011-12-11 09:23:11 [google.com/search?hl=en&q=Long Isl](http://google.com/search?hl=en&q=Long+Isl)  
2011-12-11 09:23:13 [google.com/search?q=Long Island Un](http://google.com/search?q=Long+Island+Un)  
2011-12-11 09:23:31 [google.com/search?q=Long Island Un](http://google.com/search?q=Long+Island+Un)  
2011-12-11 09:23:38 [google.com/search?q=Long Island Un](http://google.com/search?q=Long+Island+Un)  
2011-12-11 09:23:43 [google.com/search?q=Long Island Un](http://google.com/search?q=Long+Island+Un)  
2011-12-11 09:23:54 [google.com/search?q=Long Island Un](http://google.com/search?q=Long+Island+Un)  
2011-12-11 09:24:09 [colleges.usnews.rankingsandreviews.c](http://colleges.usnews.rankingsandreviews.c)

# PROVENANCE



*"High costs might come from its high room and board fees, due to its geographic location near NYC. LIU Post is a private institution taking in more money than it spends, which allow it to be more expensive than other are more academically focused."*

DID THE FACTS AND INFERENCE  
COME FROM THE SOURCE OR DID  
THE WORKER ADD THEM?

Paragraph-level citations



Regional Universities (North)

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](tel:5162992000)

**2013-2014 Tuition**  
\$34,070 tuition and fees

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

**Admissions**  
rolling admission  
78.8% accepted

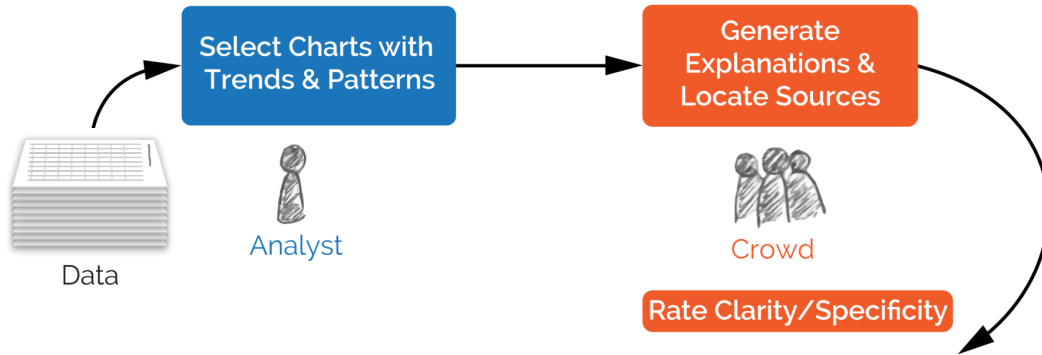
[More Information](#)

## Visitation logs

2011-12-11 09:22:04 [google.com](http://google.com)  
2011-12-11 09:22:04 [sqr:help](http://sqr:help)  
2011-12-11 09:23:08 [google.com/search?hl=en&source=hp](http://google.com/search?hl=en&source=hp)  
2011-12-11 09:23:11 [google.com/search?hl=en&q=Long+Island+Univ](http://google.com/search?hl=en&q=Long+Island+Univ)  
2011-12-11 09:23:13 [google.com/search?q=Long+Island+Univ](http://google.com/search?q=Long+Island+Univ)  
2011-12-11 09:23:31 [google.com/search?q=Long+Island+Univ](http://google.com/search?q=Long+Island+Univ)  
2011-12-11 09:23:38 [google.com/search?q=Long+Island+Univ](http://google.com/search?q=Long+Island+Univ)  
2011-12-11 09:23:43 [google.com/search?q=Long+Island+Univ](http://google.com/search?q=Long+Island+Univ)  
2011-12-11 09:23:54 [google.com/search?q=Long+Island+Univ](http://google.com/search?q=Long+Island+Univ)  
2011-12-11 09:24:09 [colleges.usnews.rankingsandreviews.com](http://colleges.usnews.rankingsandreviews.com)



# SOURCE-CHECKING MICROTASKS

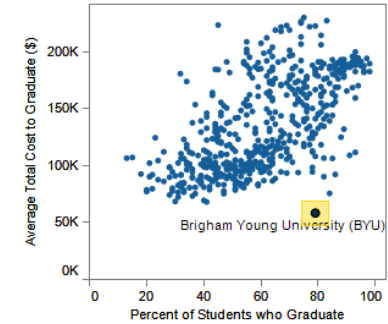


A second group of workers verifies links and attributes explanations to the source or the worker. (75% accurate in our preliminary tests )

**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."*

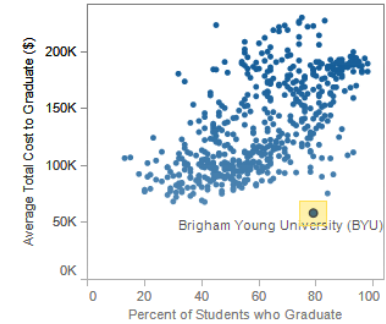
*"98% of their students are members of LDS and they have lowered tuition..."*

# REDUNDANCY

Many explanations provided by workers are redundant.

— Duplicate results for analysts to examine.

+ Redundancy can signal high support and corroborating sources.



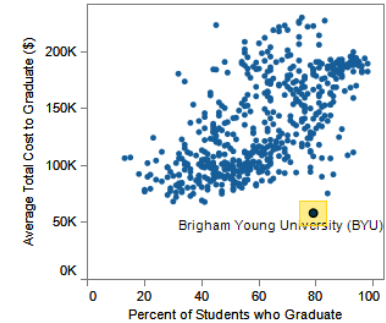
*"The Church of Jesus Christ of Latter Day Saints pays a significant part of*

*"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

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

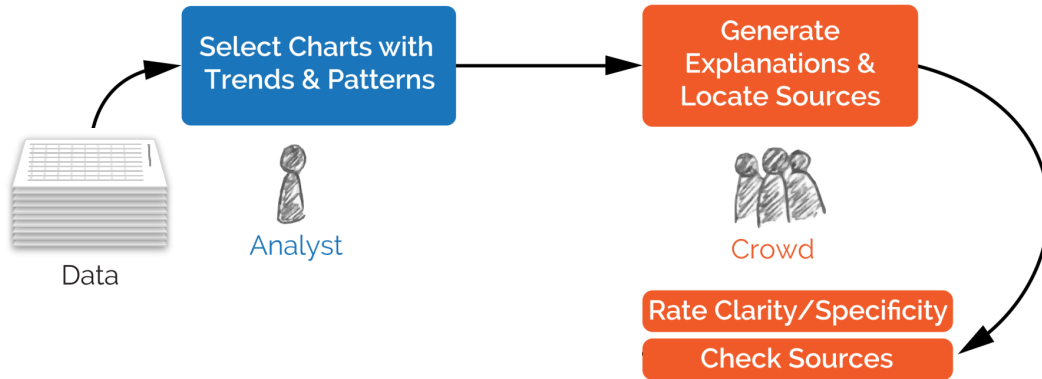


*"The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs..."*

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

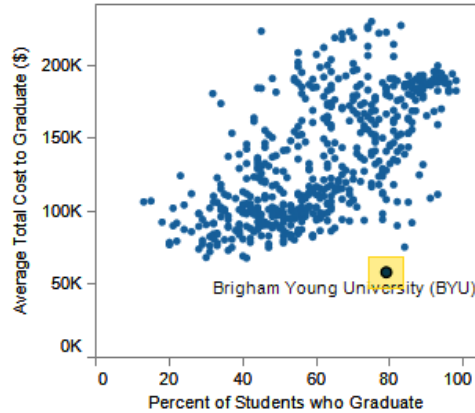
*"98% of their students are members of LDS and they have lowered tuition..."*

# REDUNDANCY



Can we crowdsource  
redundancy  
detection?

# CLUSTERING VIA DISTRIBUTED COMPARISON

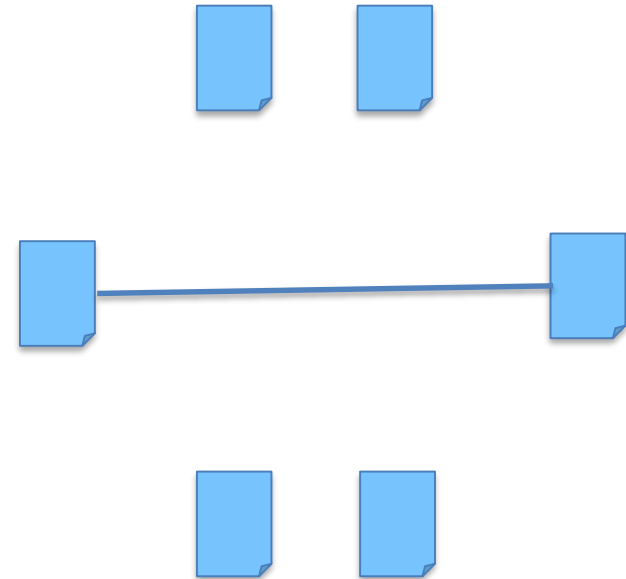


*"98% of their students are members of LDS and they have lowered tuition..."*

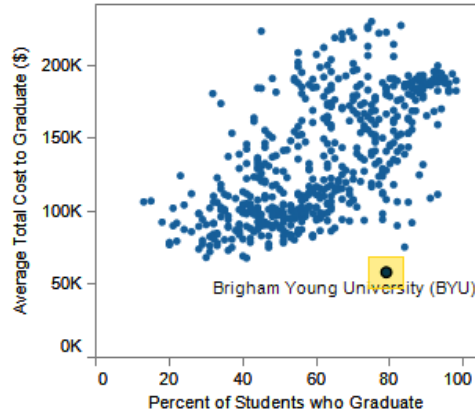
*"The cost of attendance at BYU is subsidized by the LDS church."*

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

*"The Church of Jesus Christ of Latter Day*

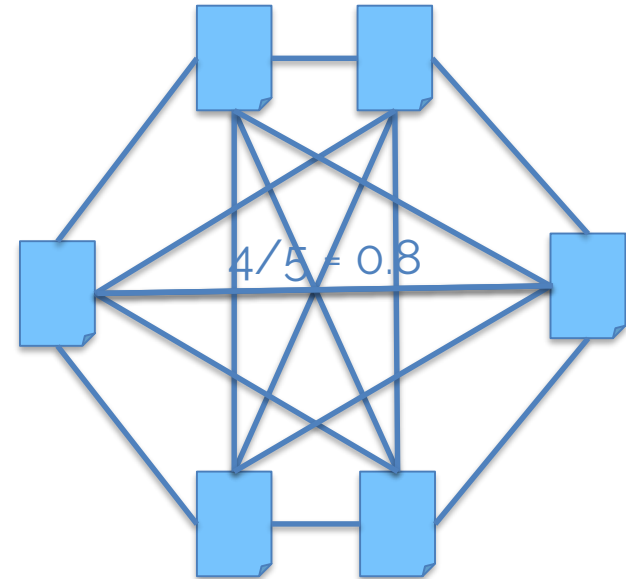


# CLUSTERING VIA DISTRIBUTED COMPARISON



*"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."*

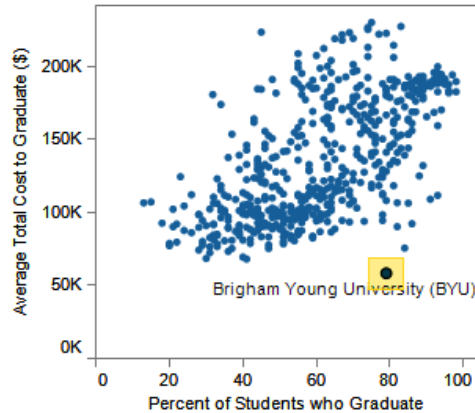


Do these two responses give the same general explanation for the peaks and valleys in the chart?

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



# CLUSTERING VIA DISTRIBUTED COMPARISON

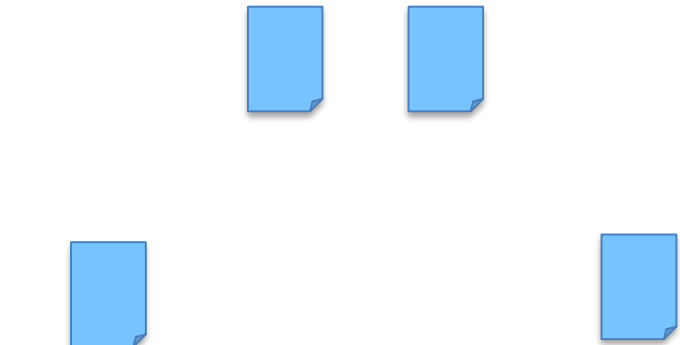


*"98% of their students are members of LDS and they have lowered tuition..."*

*"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 poorly  
Sensitive to clustering method  
Workers have little context

# CLUSTERING VIA COLOR-CODING

**Prompt:** Explain **why** the strong **peak or valley** highlighted in the chart might have occurred.

**Response R2:** "A new medical school is providing jobs"(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



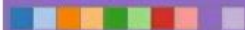
**Response R7:** "The Medical Center of the Americas opened a new medical school and in 2008 construction on a new series of projects began at the University of Texas El Paso. "(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R3:** "Expansion of Fort Bliss"(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R1:** "Increase of construction jobs."(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R4:** "It would appear that the marked growth in jobs up until 2008 coincides with growth of businesses in the area. Notable amongst these businesses are the three school districts that service the city and growth in the health services industry."(Reference: www.google.com/search?&q=el paso employers 2007 )



**Response R5:** "The high peak in 2008 was during the time when the economy was overheated. After that time the economy slipped into a recession which caused the employment status of many people to change. This is why after 2008 the graph shows a sharp drop in employment. " (Reference: www.google.com/url?q=http://en.wikipedia.org/wiki/Late-2000s\_recession&sa=U&ei=ae5qT6yoBMaosQKGI0CWCA&ved=0CBQOFIAB&usq=AFOiCNGuzI5xk-iiEUTtOIK4C8Gi6DP0FQ )



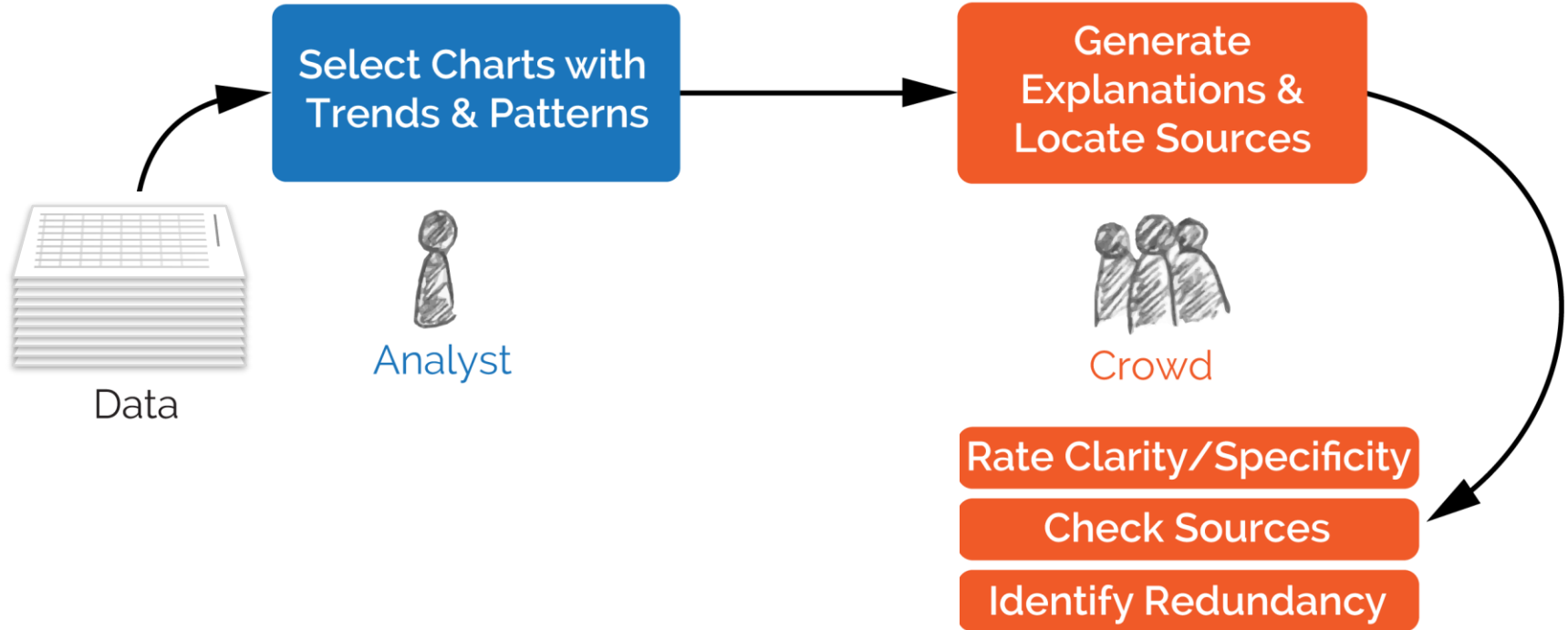
MULTIPLE WORKERS  
INDEPENDENTLY CLUSTER  
THE WHOLE SET.

USE COMPUTATIONAL  
SIMILARITY METRICS TO  
SELECT THE BEST,  
CONSISTENT CLUSTERING.

FINDING THE RIGHT BALANCE OF  
HUMAN AND AUTOMATED EFFORT

# MANAGING THE CROWD'S WORK

# MANAGING THE CROWD'S WORK



# EXPLANATION MANAGEMENT INTERFACE

The screenshot displays the Analyst UI interface within a browser window. The browser address bar shows the URL: proxy.commentspace.net/media/html/AnalystUITest/?clusters=data/All-Split-Clusters(Shiry).tab.txt. The interface includes a control bar with a quality slider set to (2.5-5), a 'Group by' dropdown menu set to 'Chart > Cluster', and a 'Sort by' dropdown menu set to 'Quality'. On the left, a scatter plot titled '3/6' shows 'Average Total Cost to Graduate (\$)' on the y-axis (0 to 200K) and 'Percent of Students who Graduate' on the x-axis (0 to 100). A single data point for Brigham Young University (BYU) is highlighted. The main content area lists several articles under the heading 'church subsidizes' (5/5 items). Each article includes a quality score in a yellow circle, a brief description, and a source link. The articles are: 1) 'The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs...' (score 5); 2) 'Brigham Young University is a private not for profit school which is also funded by The Church of Jesus Christ of Latter-Day Saints...' (score 4); 3) 'BYU is funded through the LDS church so they are able to subsidize their final needs and release the financial burden from the students tuition...' (score 3.5); 4) 'The church subsidizes the tuition of its students BYU students tuitions, therefore the true cost of tuition is not reflected in the graph...' (score 3.5); 5) 'The cost of attendance at BYU is subsidized by the LDS church...' (score 3). Below this, the 'honor code' section (1/1 item) has a score of 4, and the 'athletics' section (1/1 item) has a score of 2.5. A large empty white box is visible on the right side of the interface.

# CROWDSOURCING HIGH-LEVEL ANALYSIS

HUMAN COMPUTATION CAN BE A USEFUL  
COMPLEMENT TO AUTOMATED PROCESSING

EVEN MORE INTERESTING WITH EXPERTISE



cheap low-skill crowds

vs.

more knowledgeable trusted ones

UNDERSTANDING HOW TO PARALLELIZE  
**ANALYSIS PROCESSES** MAY BE AS  
IMPORTANT AS PARALLELIZING  
COMPUTATION HAS BEEN.

# DATA ANALYSIS AT SCALE

CHALLENGES

ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

PARALLELIZING HUMAN INTELLIGENCE





**UP NEXT**

**AFTER THE BREAK**

**TUTORIAL**



# **BONUS MATERIAL**

MORE DETAILS ON CROWDSOURCED DATA ANALYSIS

# CLUSTERING VIA COLOR-CODING

**Prompt:** Explain **why** the strong **peak or valley** highlighted in the chart might have occurred.

**Response R2:** "A new medical school is providing jobs"(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R7:** "The Medical Center of the Americas opened a new medical school and in 2008 construction on a new series of projects began at the University of Texas El Paso. "(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R3:** "Expansion of Fort Bliss"(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R1:** "Increase of construction jobs."(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R4:** "It would appear that the marked growth in jobs up until 2008 coincides with growth of businesses in the area. Notable amongst these businesses are the three school districts that service the city and growth in the health services industry."(Reference: www.google.com/search?&q=el paso employers 2007 )



**Response R5:** "The high peak in 2008 was during the time when the economy was overheated. After that time the economy slipped into a recession which caused the employment status of many people to change. This is why after 2008 the graph shows a sharp drop in employment. " (Reference: www.google.com/url?q=[http://en.wikipedia.org/wiki/Late-2000s\\_recession&sa=U&ei=ae5qT6yoBMaosQKGI0CWCA&ved=0CBQQFJAB&usq=AFOiCNGuzT5xk-iiEUTtOIK4C8Gi6DP0FQ](http://en.wikipedia.org/wiki/Late-2000s_recession&sa=U&ei=ae5qT6yoBMaosQKGI0CWCA&ved=0CBQQFJAB&usq=AFOiCNGuzT5xk-iiEUTtOIK4C8Gi6DP0FQ) )



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?

**Prompt:** Explain **why** the strong **peak or valley** highlighted in the chart might have occurred.

**Response R2:** "A new medical school is providing jobs"(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R7:** "The Medical Center of the Americas opened a new medical school and in 2008 construction on a new series of projects began at the University of Texas El Paso. "(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R3:** "Expansion of Fort Bliss"(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R1:** "Increase of construction jobs."(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus )



**Response R4:** "It would appear that the marked growth in jobs up until 2008 coincides with growth of businesses in the area. Notable amongst these businesses are the three school districts that service the city and growth in the health services industry."(Reference: www.google.com/search?&q=el paso employers 2007 )

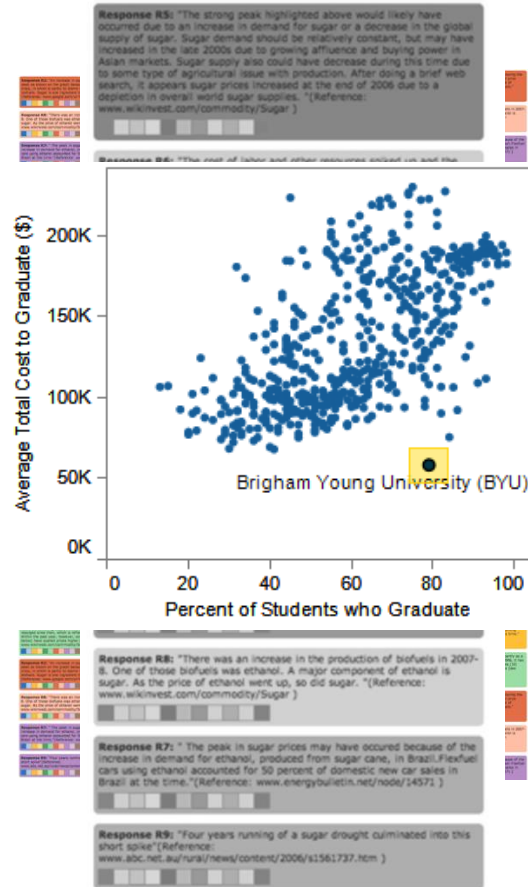


**Response R5:** "The high peak in 2008 was during the time when the economy was overheated. After that time the economy slipped into a recession which caused the employment status of many people to change. This is why after 2008 the graph shows a sharp drop in employment. " (Reference: www.google.com/url?q=http://en.wikipedia.org/wiki/Late-2000s\_recession&sa=U&ei=ae5qT6yoBMAosQKGI0CWCA&ved=0CBQOFIAB&usq=AFOiCNGuzT5xk-iiEUTtOIK4C8GI6DP0FQ )



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

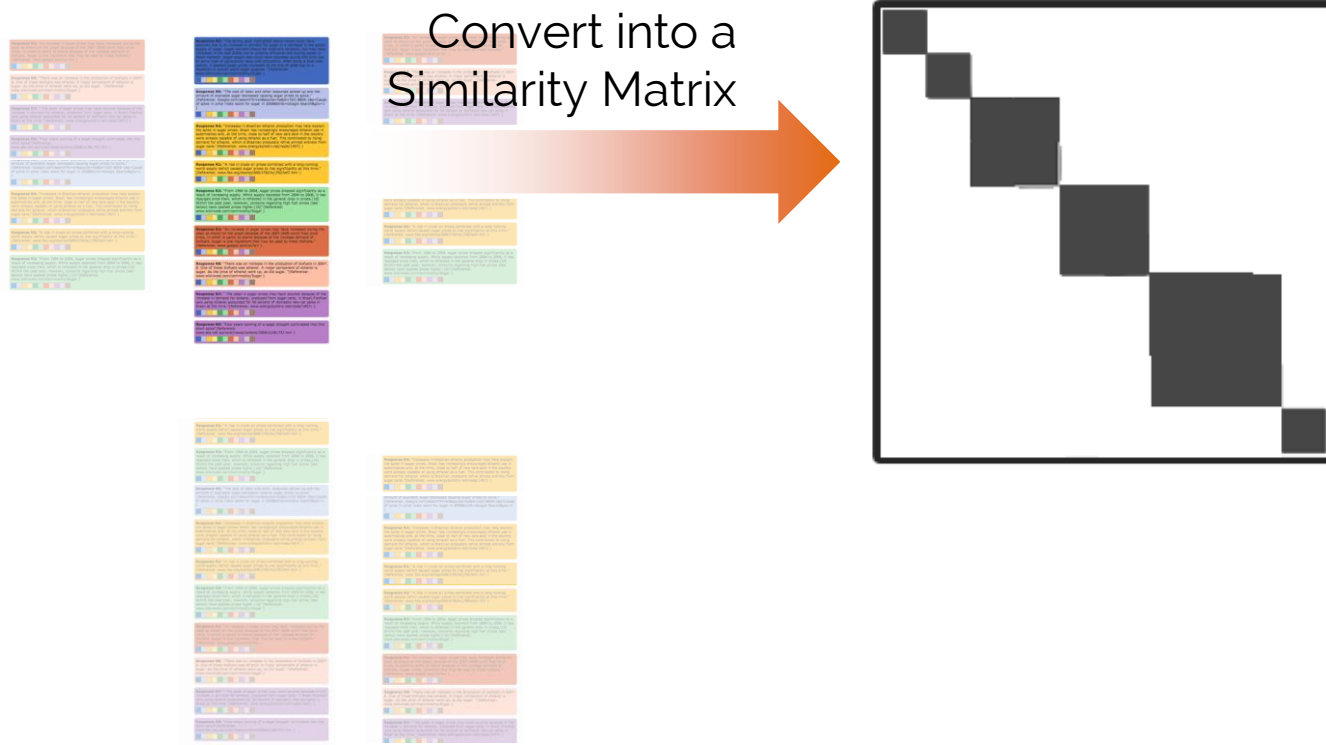
# HOW TO INTEGRATE COLOR-CLUSTERINGS?



Selecting the Most-Representative Clustering

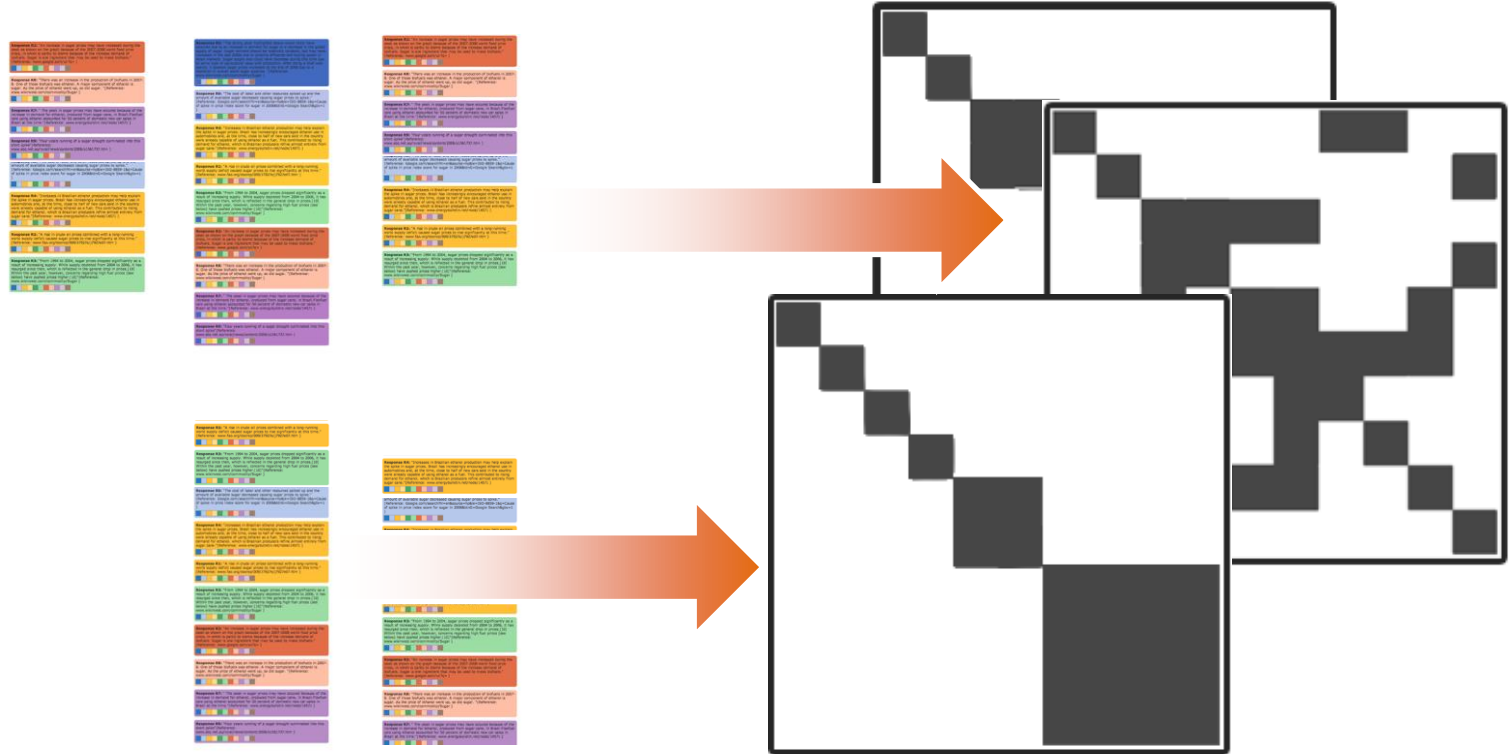


# SELECTING THE MOST-REPRESENTATIVE CLUSTERING

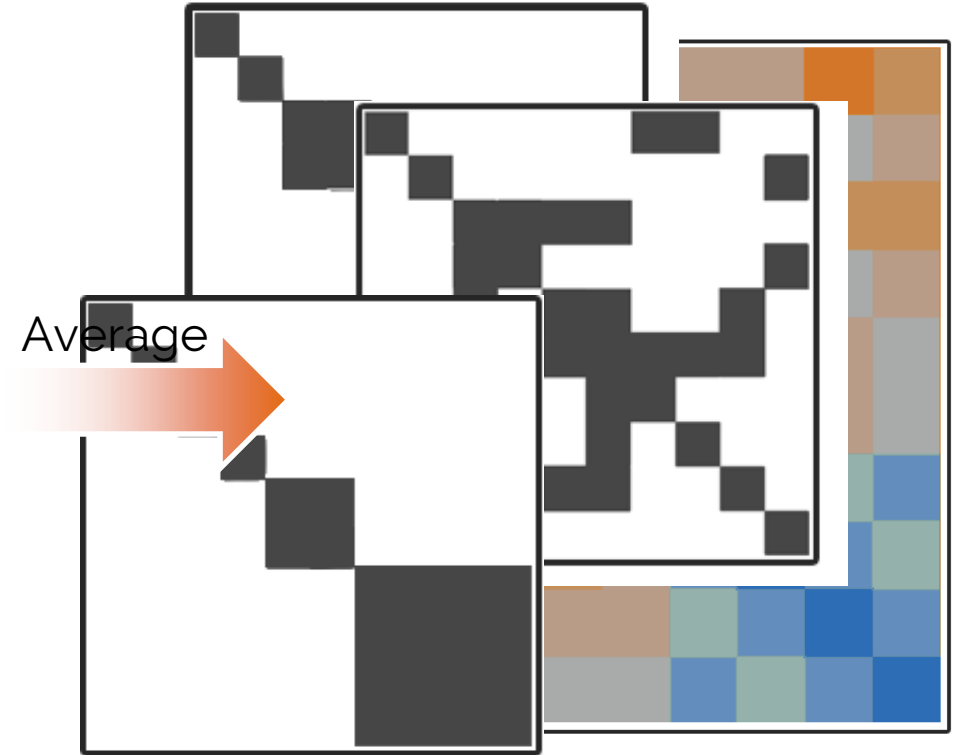




# SELECTING THE MOST-REPRESENTATIVE CLUSTERING



# SELECTING THE MOST-REPRESENTATIVE CLUSTERING



# SELECTING THE MOST-REPRESENTATIVE CLUSTERING



...

...

Select  
Highest  
Scoring



- Response R0:** "The strong gain experienced since about 1990 has been caused due to an increase in demand for sugar or a decrease in the price level of sugar. Sugar demand should be increasing steadily, but must have increased in the late 2000s due to growing ethanol and biofuel power in other markets. Sugar supply also must have decreased during this time due to some type of agricultural issue with production. After doing a brief web search, I noticed sugar prices increased in the end of 2008 due to a restriction in certain world sugar supplies." (Reference: [www.enr.com/commodity/sugar](http://www.enr.com/commodity/sugar))
- Response R1:** "The cost of labor and other resources spiked up and the amount of available sugar decreased causing sugar prices to spike." (Reference: [Google.com/search?q=reasons+for+the+2008-2009+sharp+drop+of+sugar+in+price+index+score+for+sugar+in+2008&rlz=CGoogle+Search&hl=en](https://www.google.com/search?q=reasons+for+the+2008-2009+sharp+drop+of+sugar+in+price+index+score+for+sugar+in+2008&rlz=CGoogle+Search&hl=en))
- Response R4:** "Increases in Brazilian ethanol production may help explain the spike in sugar prices. Brazil has increasingly encouraged ethanol use in automobiles and, at the time, close to half of new cars sold in the country were already capable of using ethanol as a fuel. This contributed to rising demand for ethanol, which is Brazilian producers refine almost entirely from sugar cane." (Reference: [www.energysolutions.net/node/14371](http://www.energysolutions.net/node/14371))
- Response R1:** "A rise in crude oil prices combined with a long-running wind supply deficit caused sugar prices to rise significantly at this time." (Reference: [www.fao.org/docrep/009/x7927/x792707.htm](http://www.fao.org/docrep/009/x7927/x792707.htm))
- Response R3:** "From 1994 to 2004, sugar prices dropped significantly as a result of increasing supply. While supply decreased from 2004 to 2006, it has rebounded since then, which is reflected in the general drop in prices. (2) Within the past year, however, concerns regarding High Fuel prices (see below) have pushed prices higher. (10)" (Reference: [www.wickwest.com/commodity/Sugar](http://www.wickwest.com/commodity/Sugar))
- Response R2:** "An increase in sugar prices may have increased during the spike in prices on the grain business of the 2007-2008 world food price crisis, in which is partly to blame because of the increase demand of biofuels. Sugar is most important fuel may be used to make biofuels." (Reference: [www.google.com/url?q=](http://www.google.com/url?q=))
- Response R8:** "There was an increase in the production of biofuels in 2007-8. One of those biofuels was ethanol. A major component of ethanol is sugar. As the price of ethanol went up, so did sugar." (Reference: [www.wickwest.com/commodity/Sugar](http://www.wickwest.com/commodity/Sugar))
- Response R7:** "The peak in sugar prices may have occurred because of the increase in demand for ethanol, produced from sugar cane, in Brazil. Ethanol cars using ethanol accounted for 50 percent of domestic new car sales in Brazil at the time." (Reference: [www.energysolutions.net/node/14371](http://www.energysolutions.net/node/14371))
- Response R9:** "Four years running of a sugar drought culminated into this price spike! (Reference: [www.dcc.net.au/news/news/stories/2009/x1361737.htm](http://www.dcc.net.au/news/news/stories/2009/x1361737.htm))

# 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 per 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)  $[0 \longleftrightarrow 1]$  (identical)

# EVALUATING REDUNDANCY-DETECTION

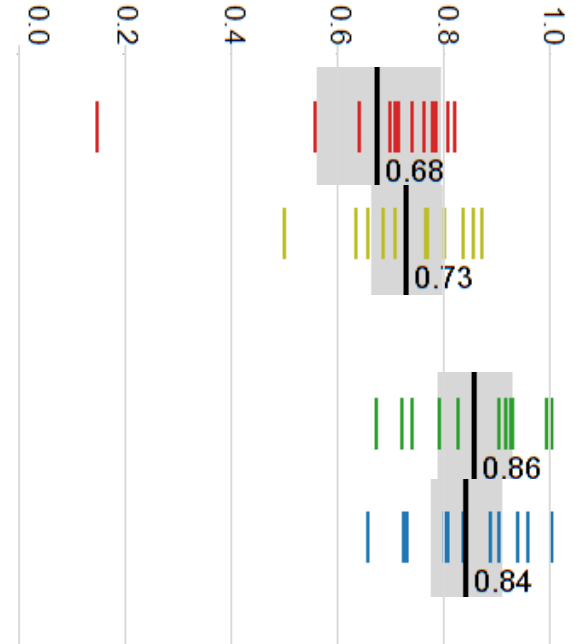
Average F-measure Score (vs. Experts)

Unclustered Results **F=0.68**

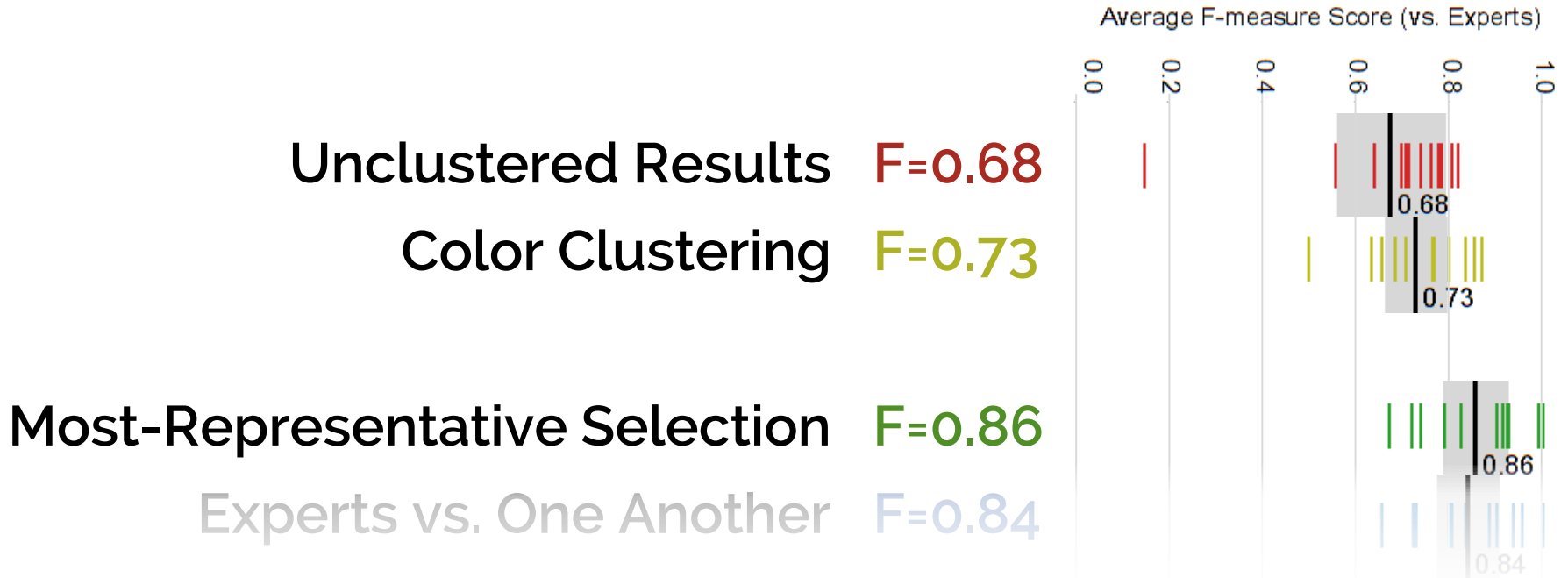
Color Clustering **F=0.73**

Most-Representative Selection **F=0.86**

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



# EVALUATING REDUNDANCY-DETECTION



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