DATA ANALYSIS AT SCALE

WESLEY WILLETT

VISUAL ANALYTICS

15 OCT 2014

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

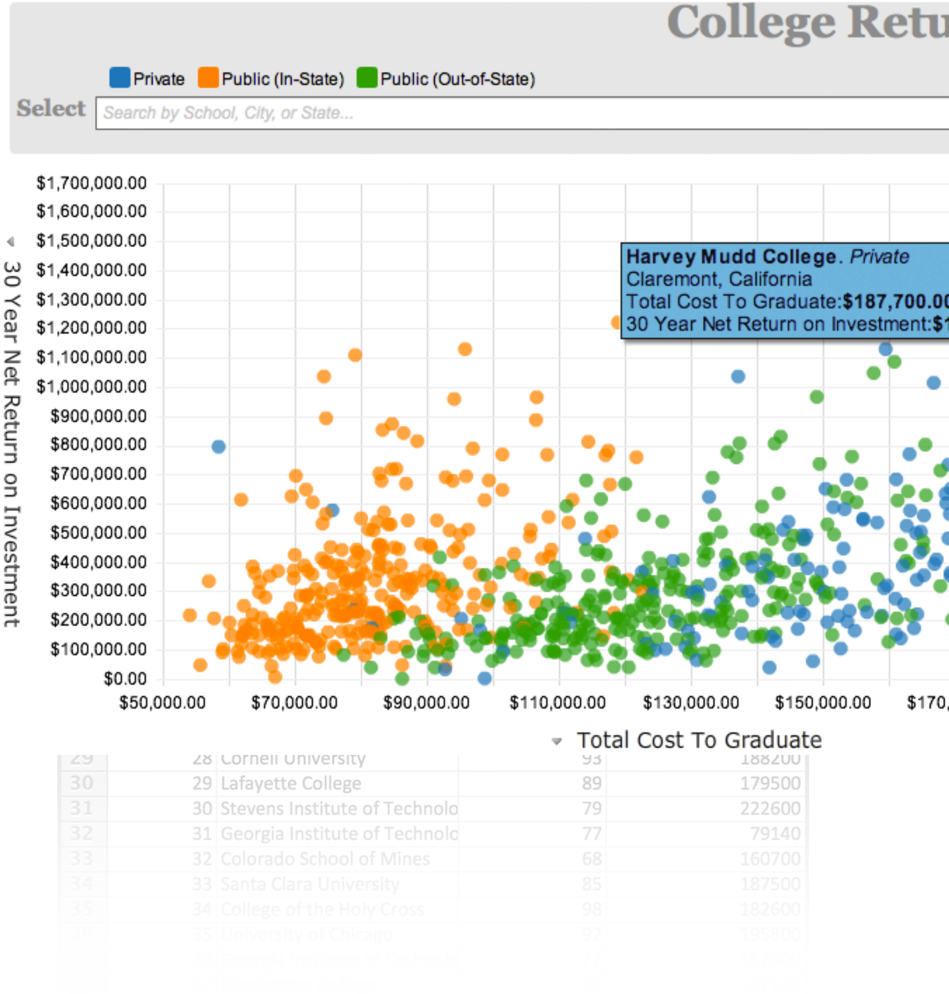
MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

PETABYTES OF DATA

...



Harvey Mudd College. *Private*
Claremont, California
Total Cost To Graduate:**$187,700.0**
30 Year Net Return on Investment:**$1**

29	28	Cornell University	93	188200
30	29	Lafayette College	89	179500
31	30	Stevens Institute of Technolo	79	222600
32	31	Georgia Institute of Technolo	77	79140
33	32	Colorado School of Mines	68	160700
34	33	Santa Clara University	85	187500
35	34	College of the Holy Cross	98	182600
36	35	University of Chicago	92	195800

SIZE

KILOBYTES OF DATA

MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

PETABYTES OF DATA

...

2560 X 1600 = 4,096,000 PIXELS

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

SIZE

KILOBYTES OF DATA

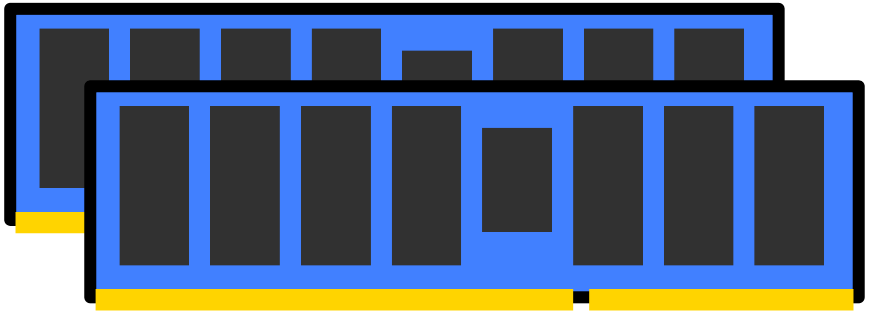
MEGABYTES OF DATA

GIGABYTES OF DATA

TERABYTES OF DATA

PETABYTES OF DATA

...



MORE DATA THAN CAN FIT IN MEMORY

SIZE

KILOBYTES OF DATA

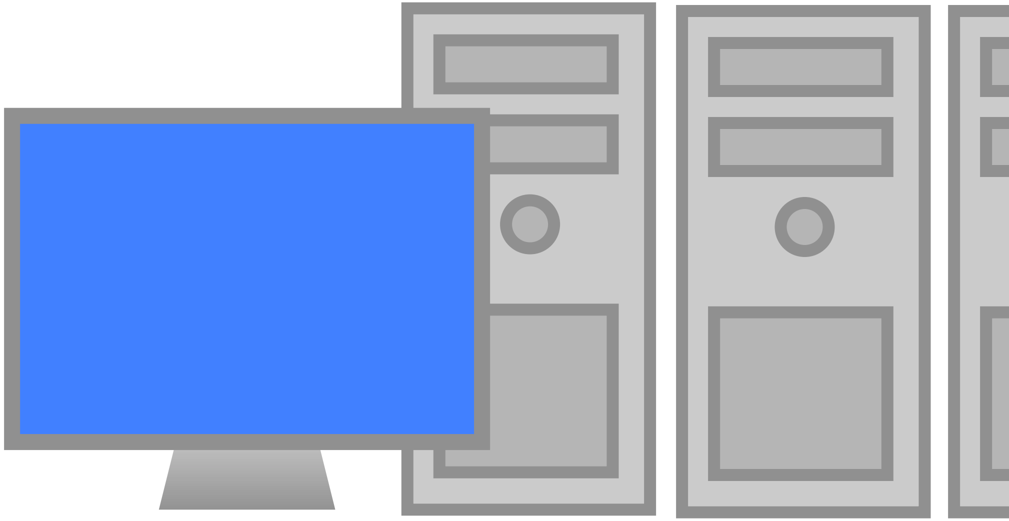
MEGABYTES OF DATA

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

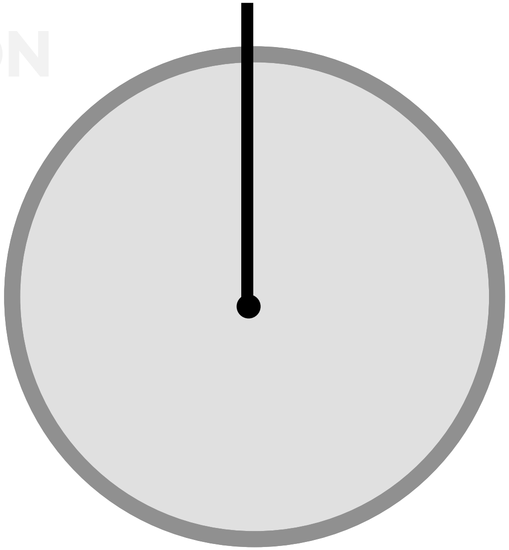
TERABYTES OF DATA

PETABYTES OF DATA

(MANY BIG DATA-DRIVEN QUESTIONS TODAY)

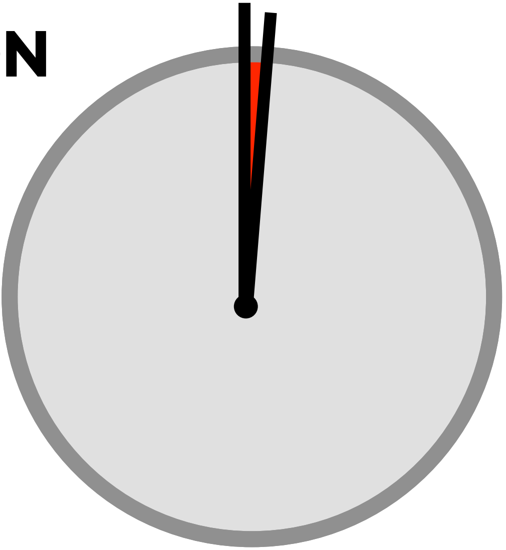
SPEED

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



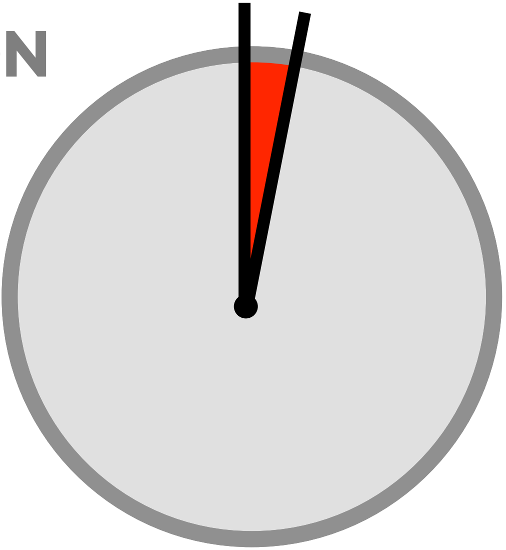
SPEED

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



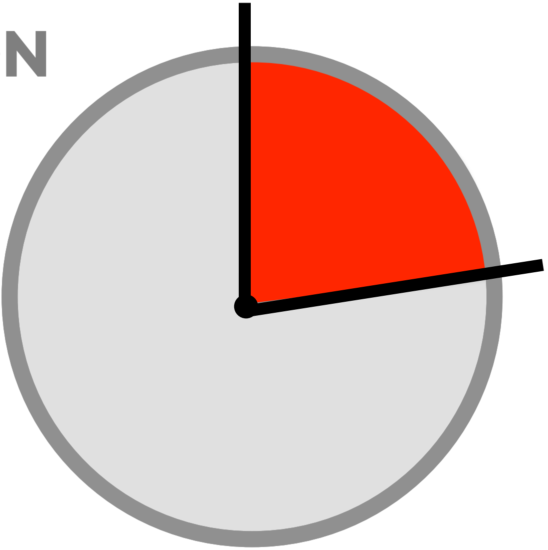
SPEED

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



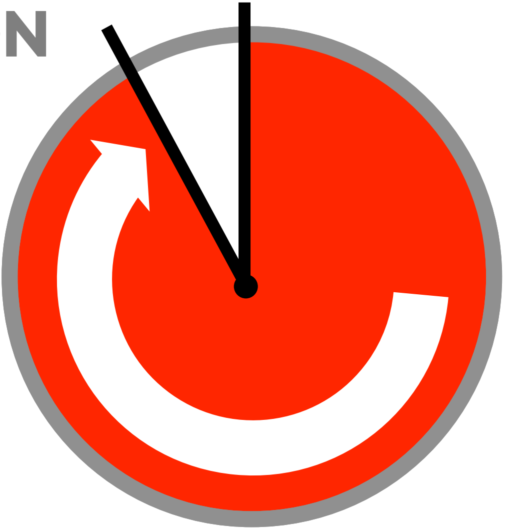
SPEED

~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) [OECD 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

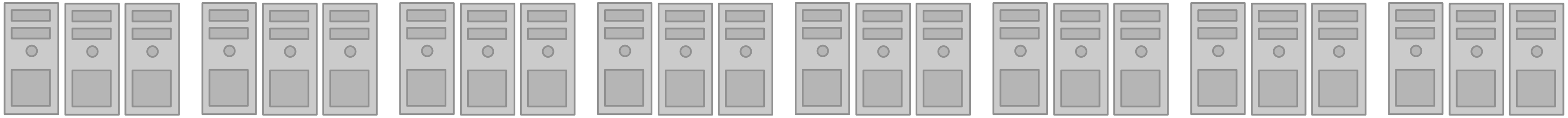
ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

PARALLELIZING HUMAN INTELLIGENCE

ANALYSIS & CLUSTER COMPUTING

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



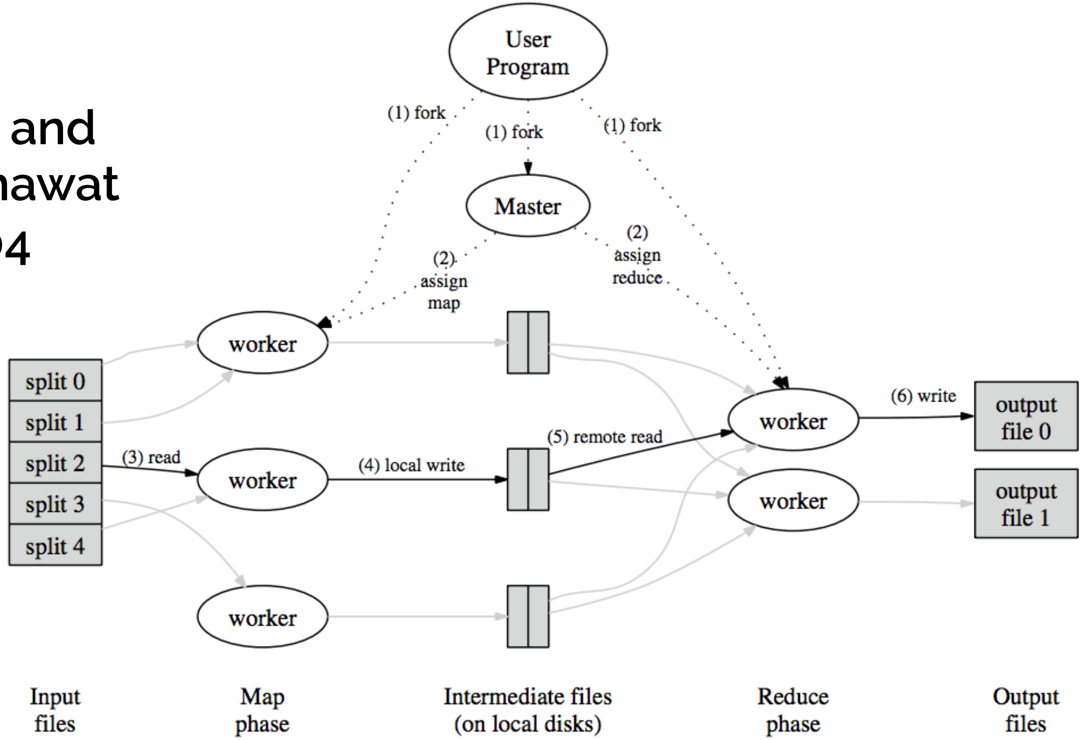
ANALYSIS REQUIRES **DISTRIBUTED** DATA PROCESSING

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
…

[K. Ousterhout - UCB 194-16]

JUST A HASH TABLE

"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham?"

{ }

[K. Ousterhout - UCB 194-16]

JUST A HASH TABLE

"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham?"

{I:1}

[K. Ousterhout - UCB 194-16]

JUST A HASH TABLE

"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham?"

{I:1,
am:1 }

[K. Ousterhout - UCB 194-16]

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}

[K. Ousterhout - UCB 194-16]

JUST A HASH TABLE

"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham?"

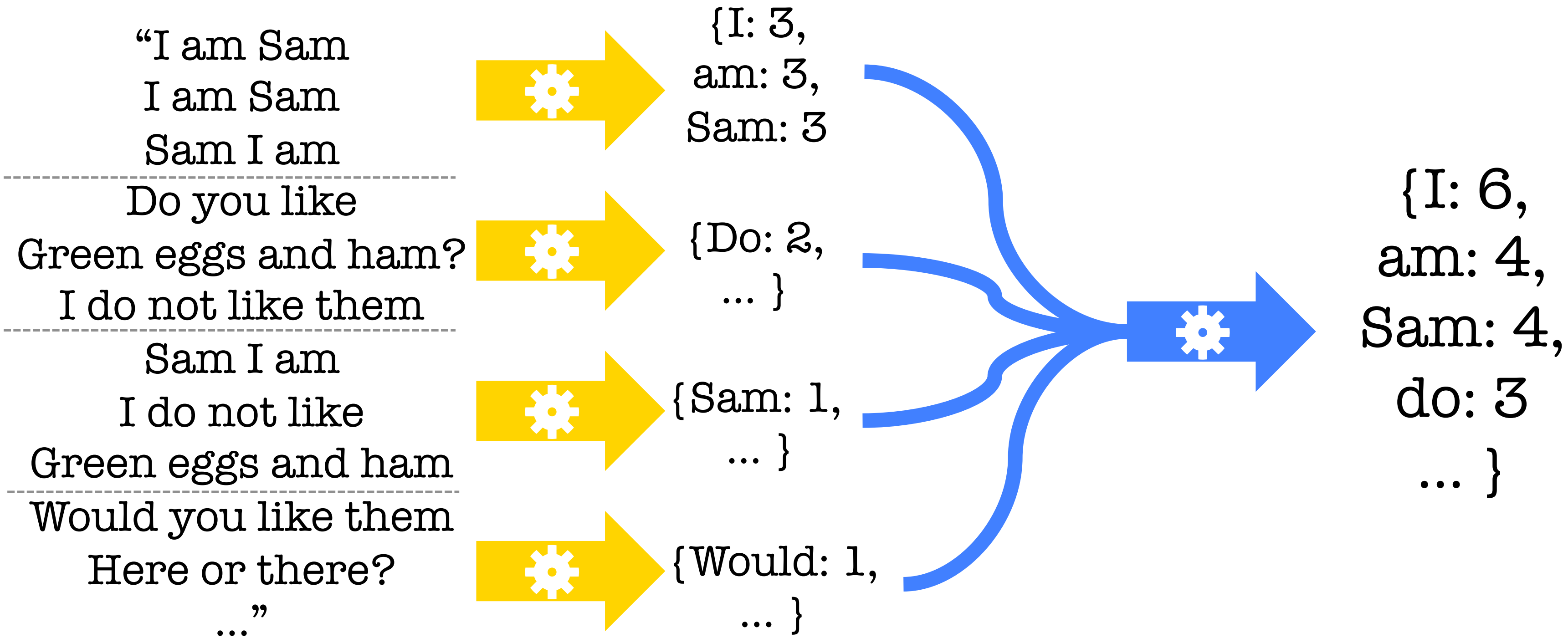
{I:2,
am:1,
Sam:1}

[K. Ousterhout - UCB 194-16]

BUT YOU SAID THE DOCUMENT IS REALLY BIG?

[K. Ousterhout - UCB 194-16]

COMPUTE IN PARALLEL



"I am Sam
I am Sam
Sam I am

Do you like
Green eggs and ham?
I do not like them

Sam I am
I do not like
Green eggs and ham

Would you like them
Here or there?
..."

{I: 3,
am: 3,
Sam: 3

{Do: 2,
... }

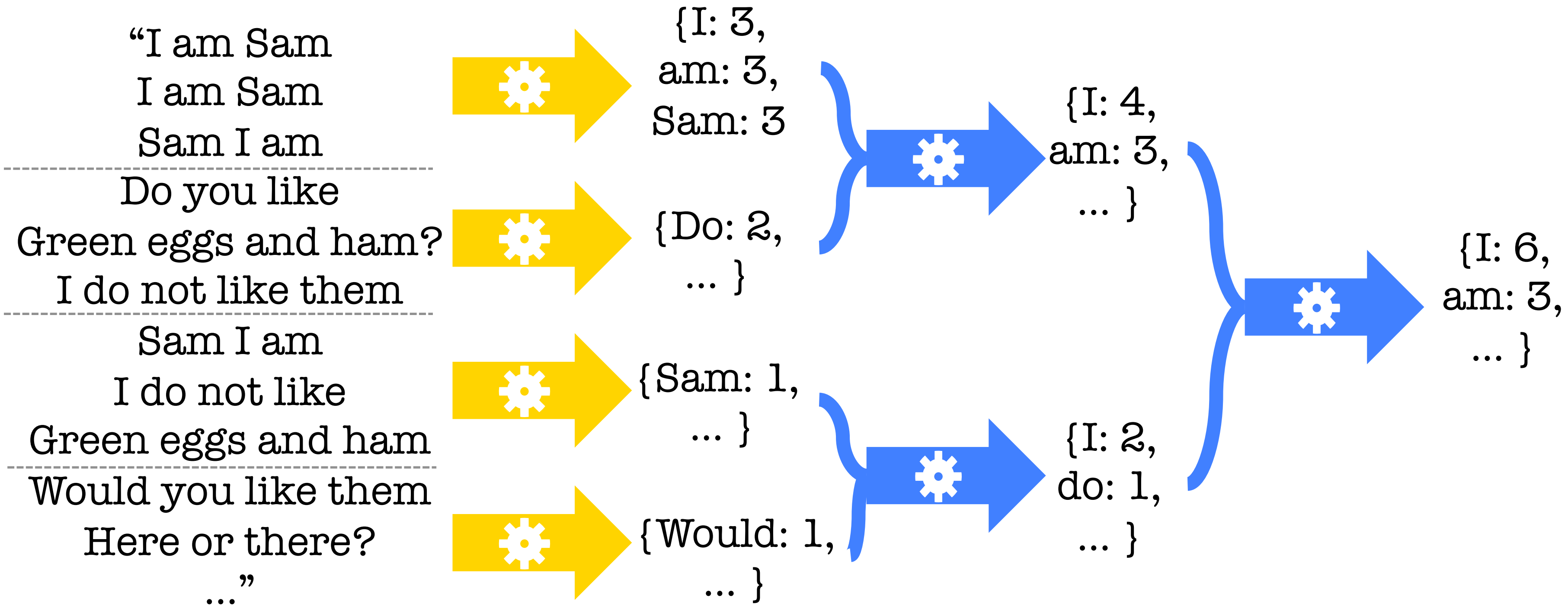
{Sam: 1,
... }

{Would: 1,
... }

{I: 6,
am: 4,
Sam: 4,
do: 3
... }

[K. Ousterhout - UCB 194-16]

COMPUTE IN PARALLEL



“I am Sam
I am Sam
Sam I am

Do you like
Green eggs and ham?
I do not like them

Sam I am
I do not like
Green eggs and ham

Would you like them
Here or there?
…”

{I: 3,
am: 3,
Sam: 3

{Do: 2,
… }

{Sam: 1,
… }

{Would: 1,
… }

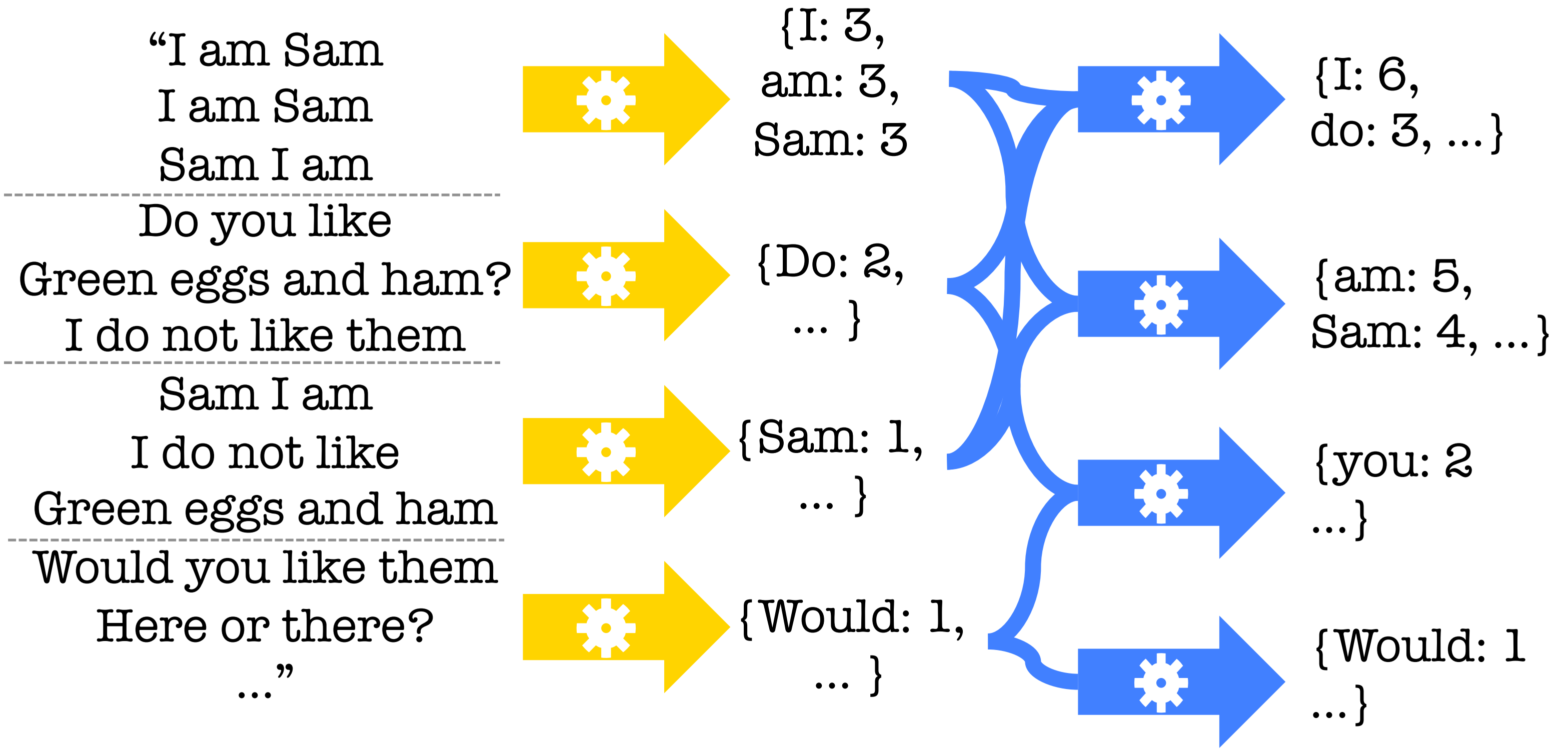
{I: 4,
am: 3,
… }

{I: 2,
do: 1,
… }

{I: 6,
am: 3,
… }

[K. Ousterhout - UCB 194-16]

COMPUTE IN PARALLEL



"I am Sam
I am Sam
Sam I am

Do you like
Green eggs and ham?
I do not like them

Sam I am
I do not like
Green eggs and ham

Would you like them
Here or there?
…"

{I: 3,
am: 3,
Sam: 3

{Do: 2,
… }

{Sam: 1,
… }

{Would: 1,
… }

{I: 6,
do: 3, …}

{am: 5,
Sam: 4, …}

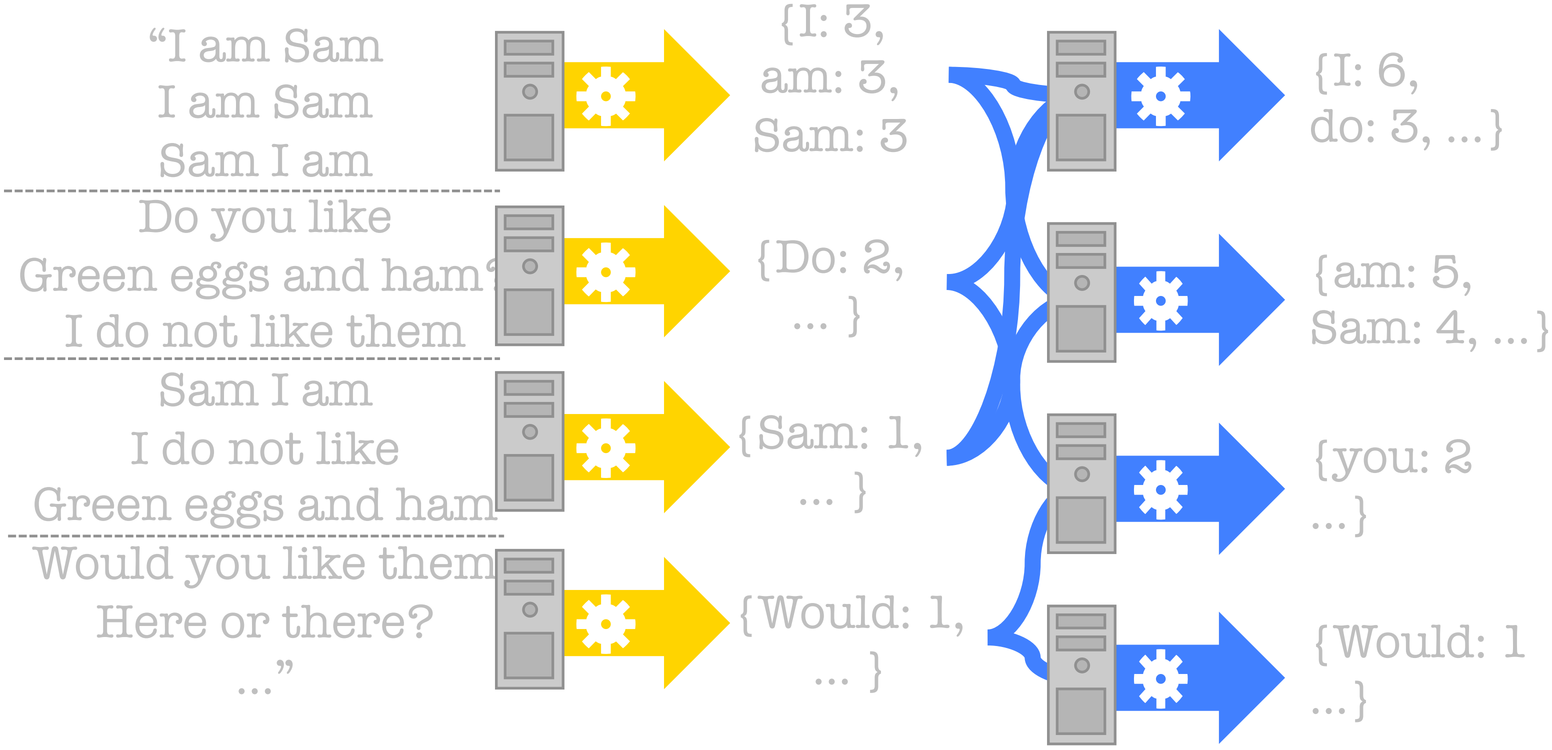
{you: 2
…}

{Would: 1
…}

[K. Ousterhout - UCB 194-16]

MAP

REDUCE



“I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham?
I do not like them
Sam I am
I do not like
Green eggs and ham
Would you like them
Here or there?
…”

{I: 3,
am: 3,
Sam: 3

{Do: 2,
… }

{Sam: 1,
… }

{Would: 1,
… }

{I: 6,
do: 3, …}

{am: 5,
Sam: 4, …}

{you: 2
…}

{Would: 1
…}

[K. Ousterhout - UCB 194-16]

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)
```

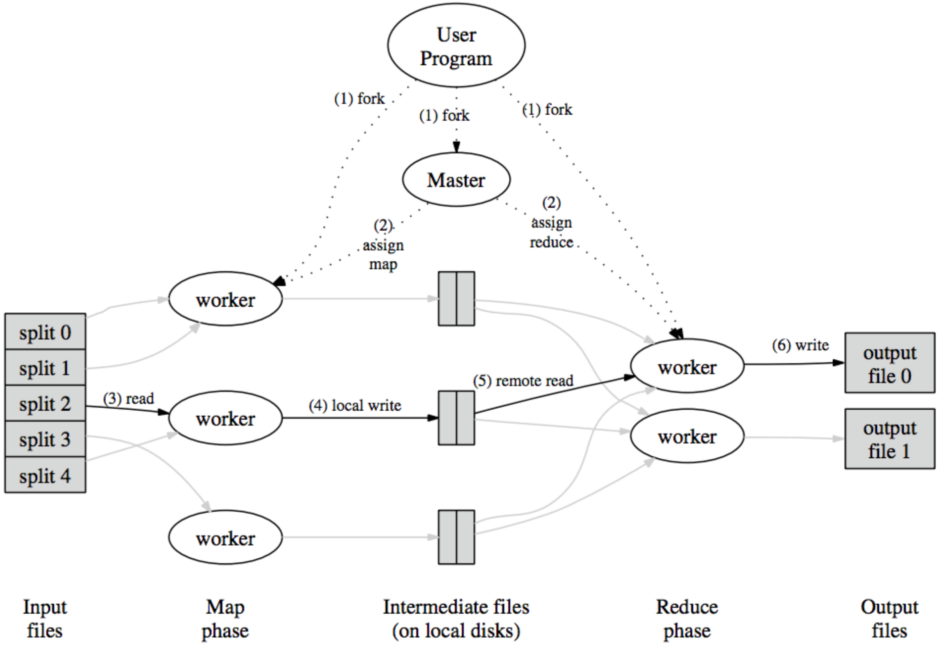
REDUCE

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

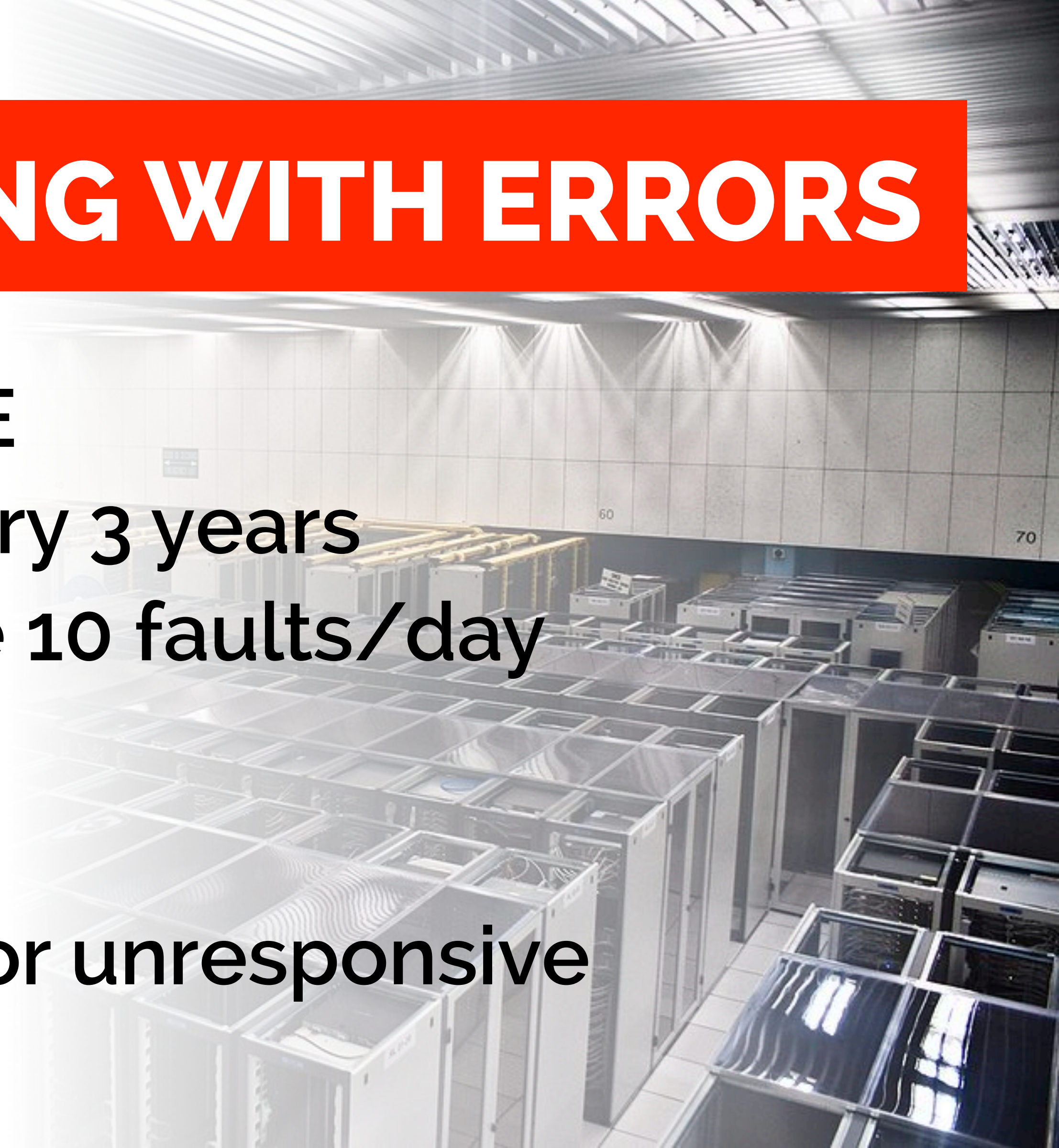
MAP-REDUCE

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



SERVER FAILURE

1 server fails every 3 years

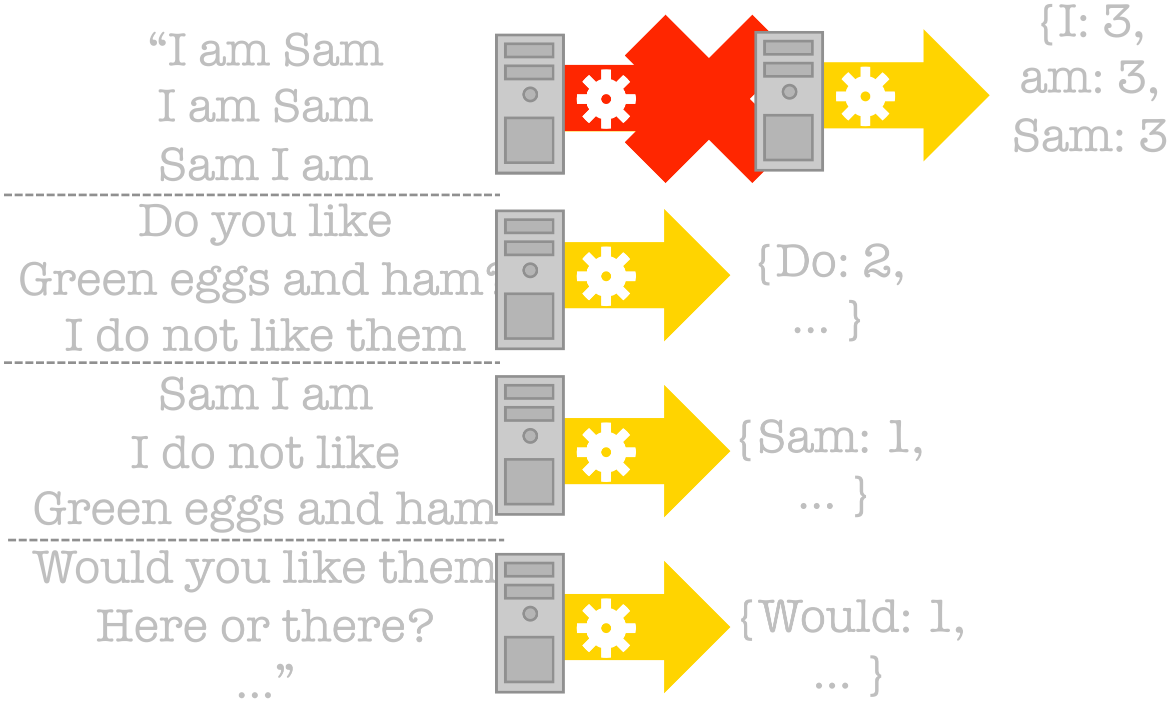
➔10K nodes see 10 faults/day

STRAGGLERS

Nodes are slow or unresponsive

[K. Ousterhout - UCB 194-16]

JUST LAUNCH A REPLACEMENT



"I am Sam
I am Sam
Sam I am

Do you like
Green eggs and ham?
I do not like them

Sam I am
I do not like
Green eggs and ham

Would you like them
Here or there?
…"

{I: 3,
am: 3,
Sam: 3

{Do: 2,
… }

{Sam: 1,
… }

{Would: 1,
… }

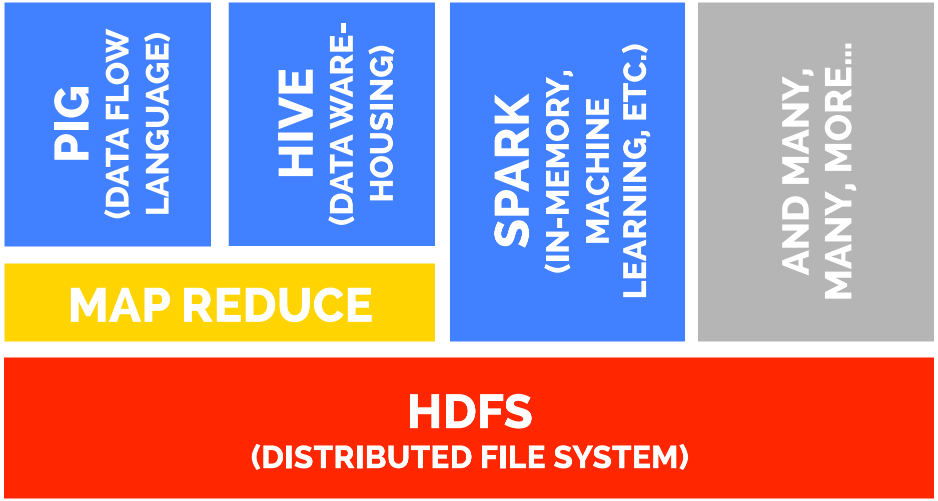
[K. Ousterhout - UCB 194-16]

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



PIG (DATA FLOW LANGUAGE)

HIVE (DATA WARE-HOUSING)

SPARK (IN-MEMORY, MACHINE LEARNING, ETC.)

AND MANY, MANY, MORE...

MAP REDUCE

HDFS (DISTRIBUTED FILE SYSTEM)

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

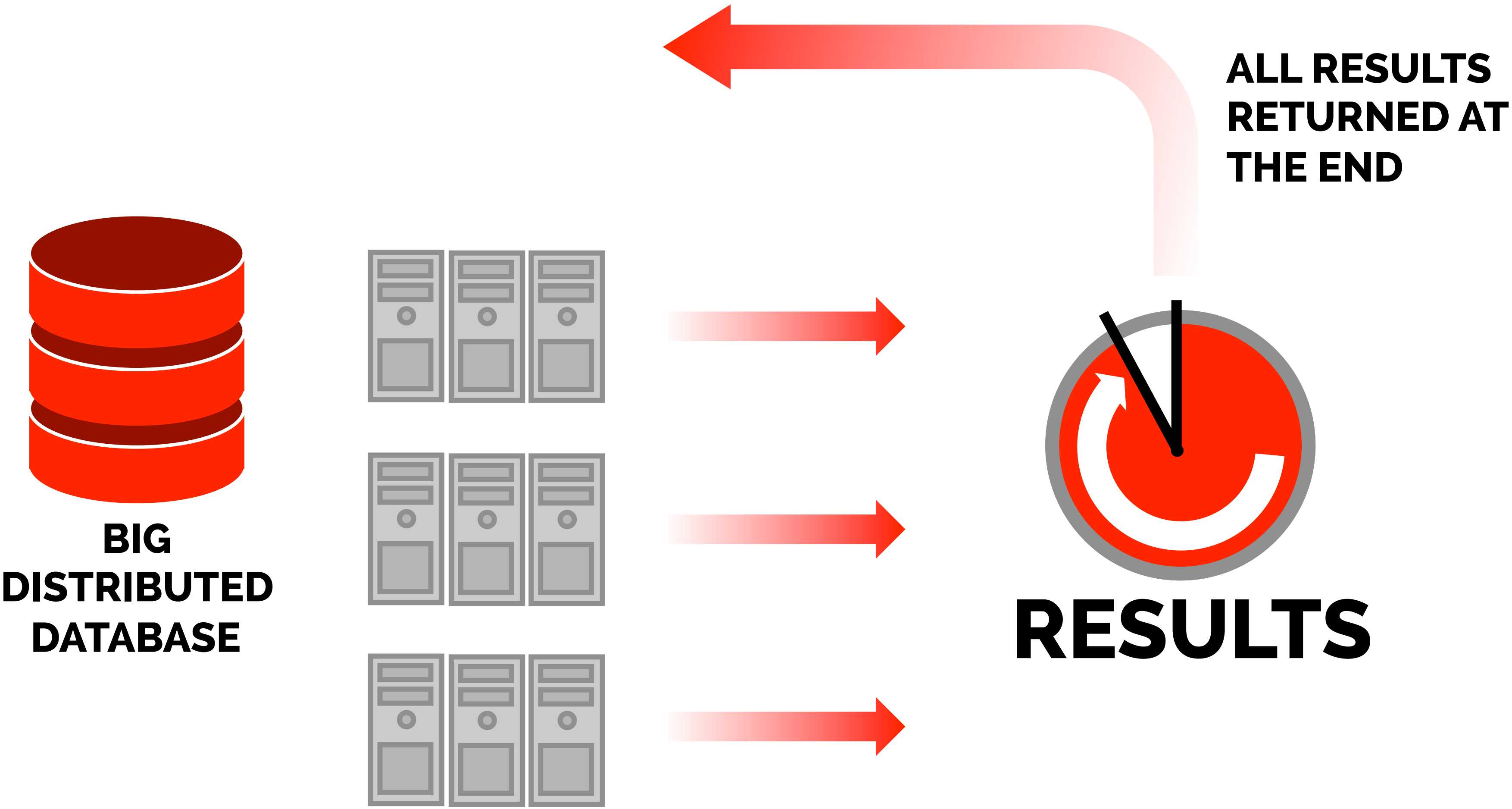
(LEAVE BIG DATA ON THE SERVERS)

3. SAMPLE INTERACTIVELY

(APPROXIMATE FIRST THEN REFINE)

SAMPLING FOR INTERACTION

STANDARD QUERY

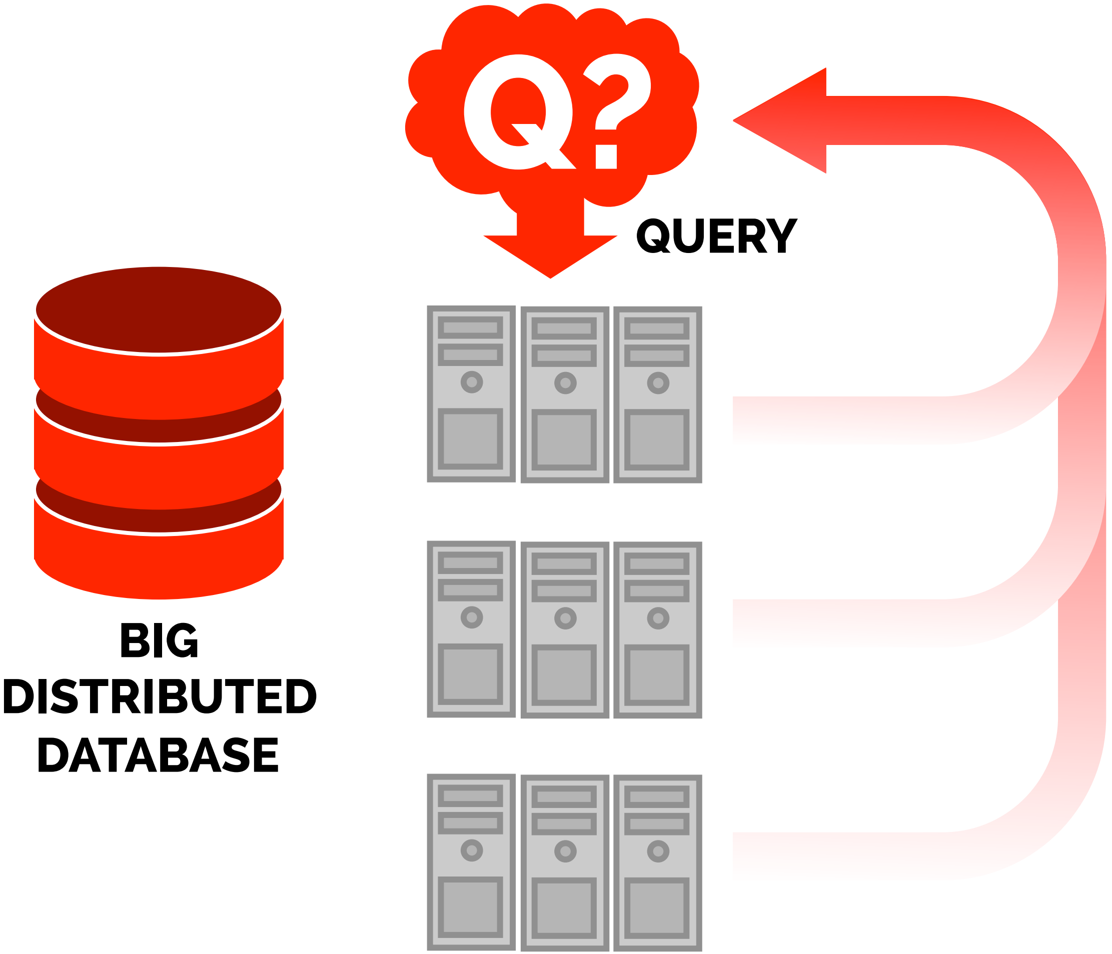


ALL RESULTS RETURNED AT THE END

BIG DISTRIBUTED DATABASE

RESULTS

INTERACTIVE SAMPLING

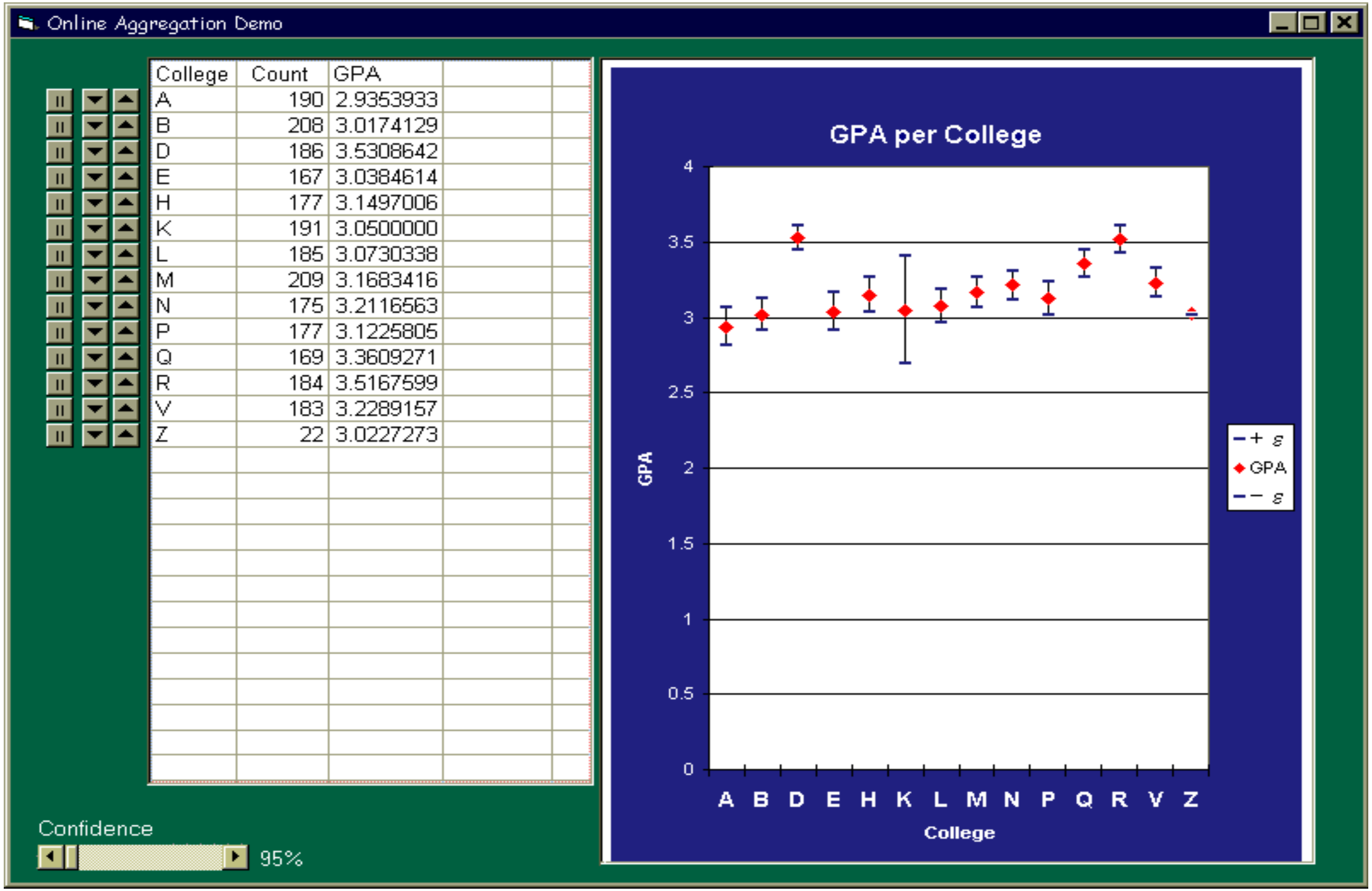


QUERY

BIG DISTRIBUTED DATABASE

RESULTS RETURNED WHILE QUERY IS STILL RUNNING

INTERACTIVE SAMPLING



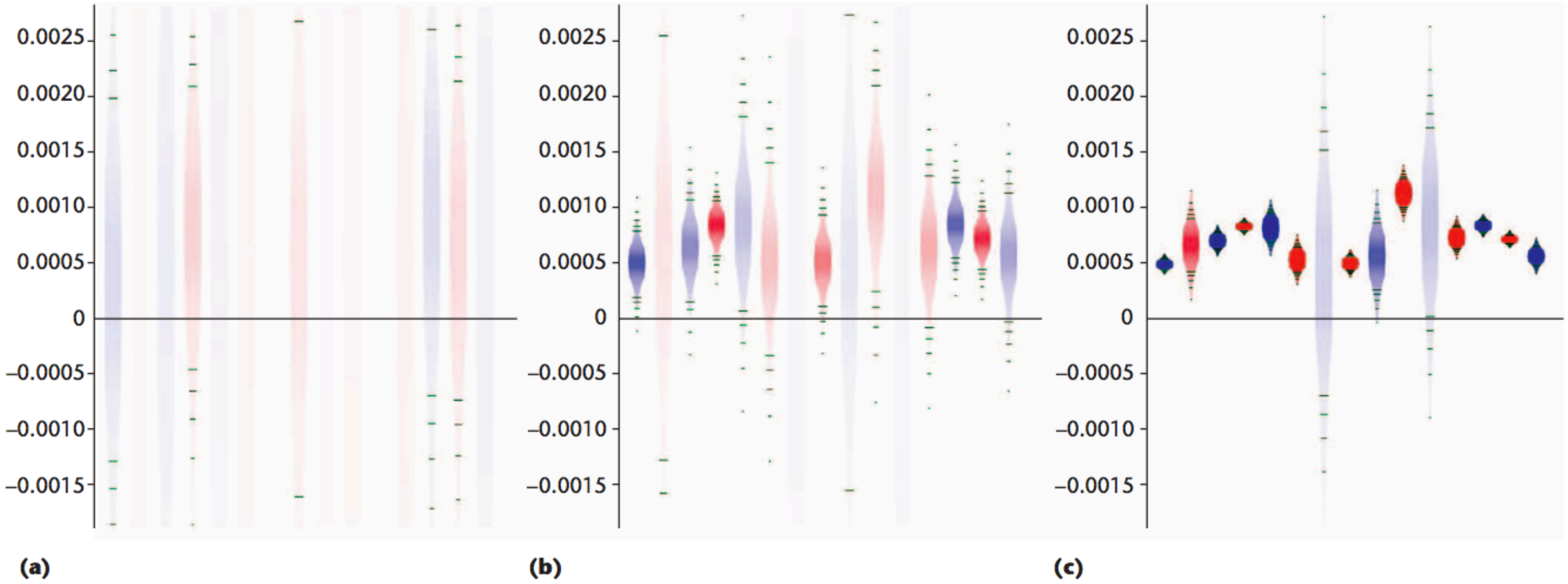
Online Aggregation Demo

College	Count	GPA
A	190	2.9353933
B	208	3.0174129
D	186	3.5308642
E	167	3.0384614
H	177	3.1497006
K	191	3.0500000
L	185	3.0730338
M	209	3.1683416
N	175	3.2116563
P	177	3.1225805
Q	169	3.3609271
R	184	3.5167599
V	183	3.2289157
Z	22	3.0227273

Confidence: 95%

CONTROL [HELLERSTEIN ET AL. 1999]

INTERACTIVE SAMPLING



SAMPLEACTION [FISHER ET AL. 2012]

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?

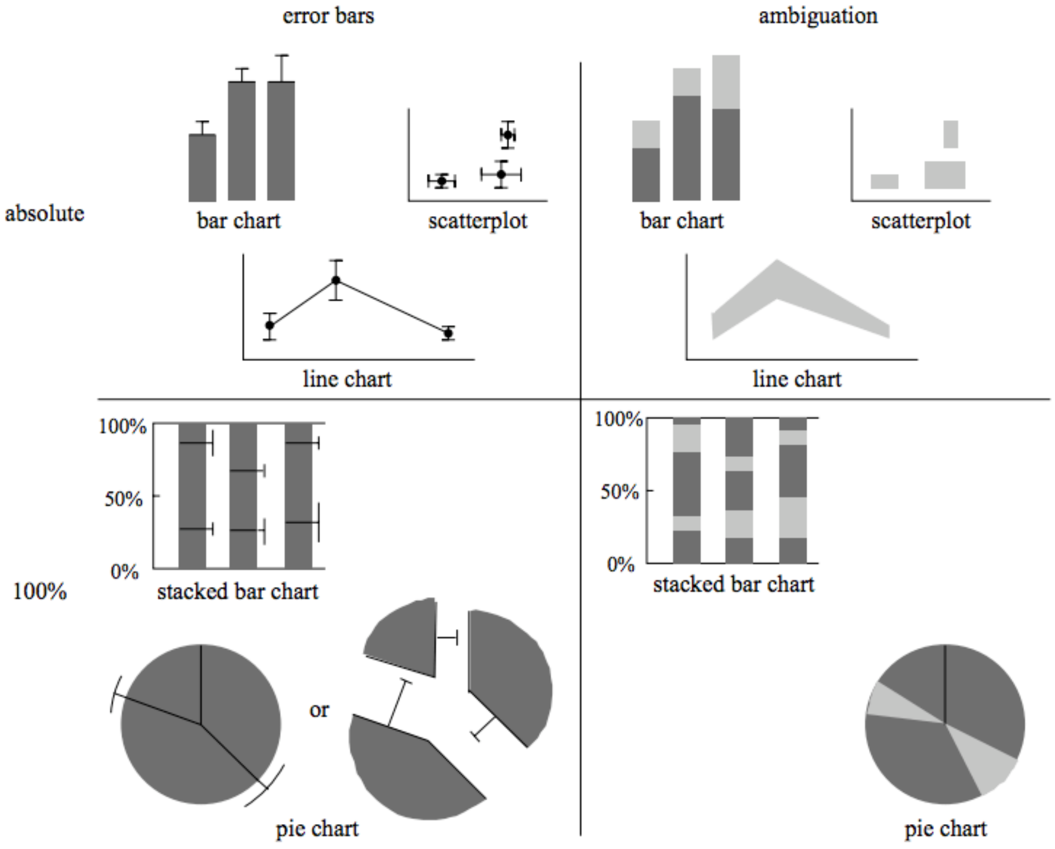
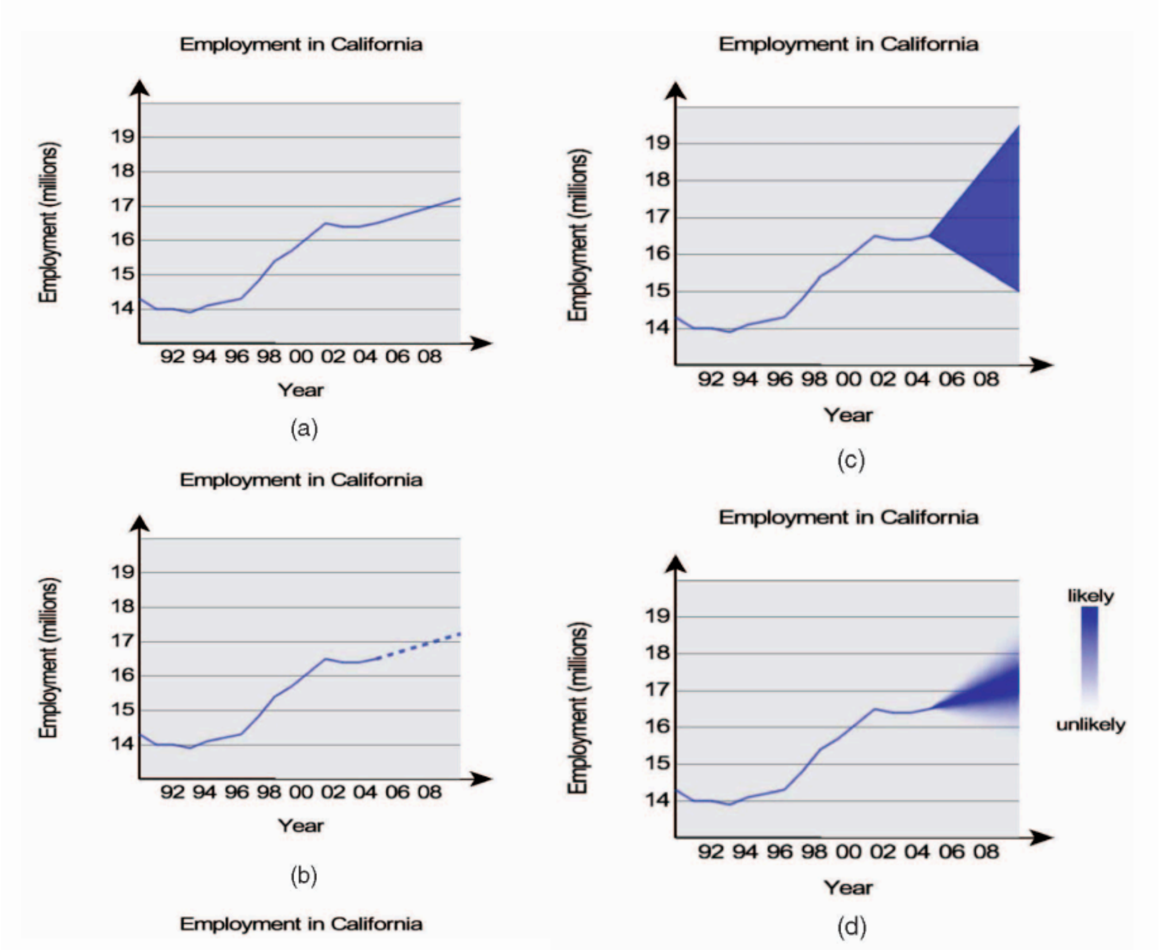


Figure 1: Error bars and ambiguation applied to some common chart types.

[Olston & Mackinlay, 2002]

HOW TO SHOW UNCERTAINTY?



(a)

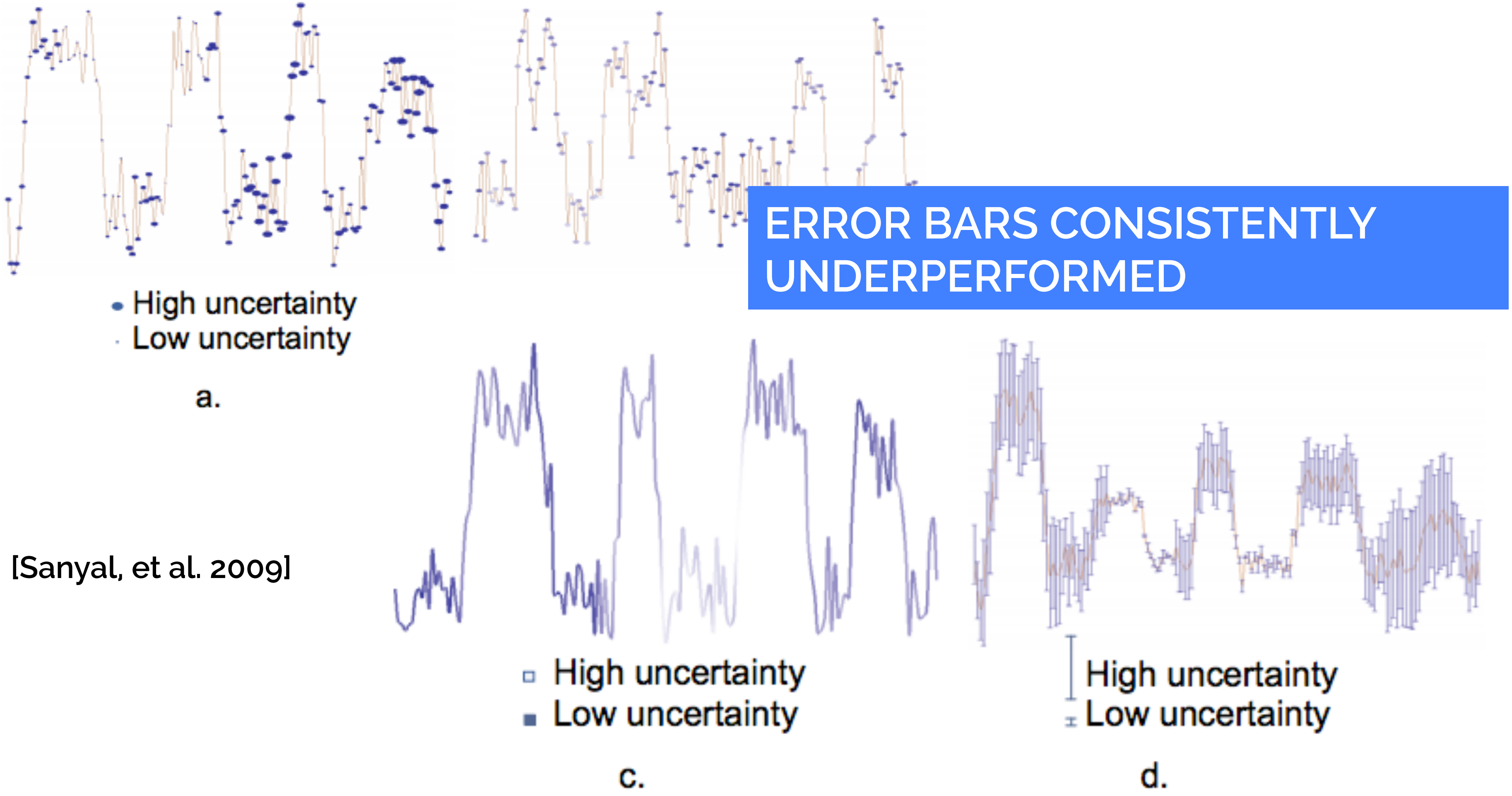
(b)

(c)

(d)

[Streit, Pham, & Brown 2008]

HOW TO SHOW UNCERTAINTY?



a.

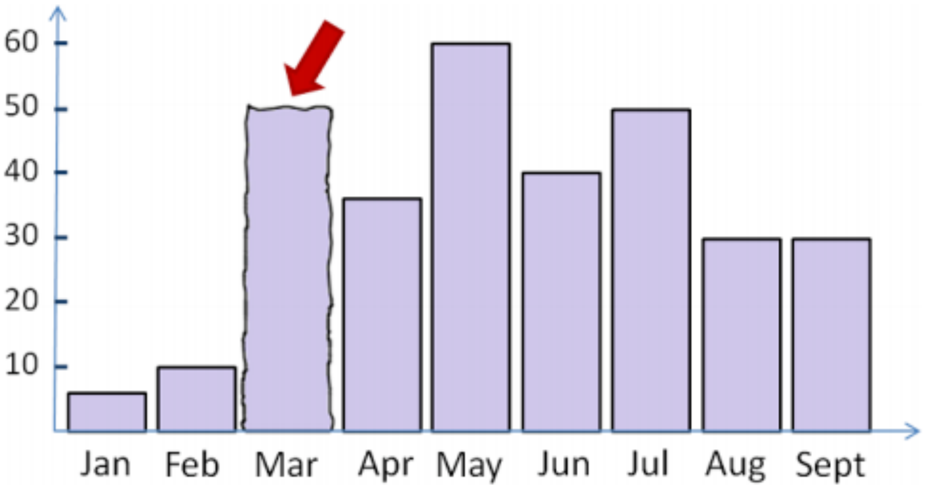
c.

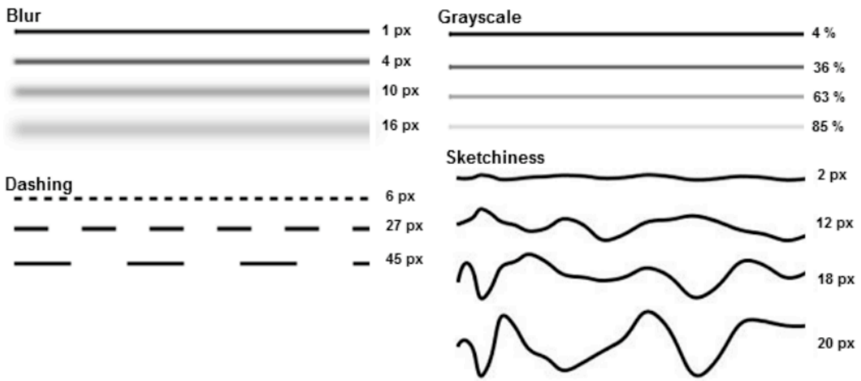
d.

ERROR BARS CONSISTENTLY UNDERPERFORMED

[Sanyal, et al. 2009]

HOW TO SHOW UNCERTAINTY?

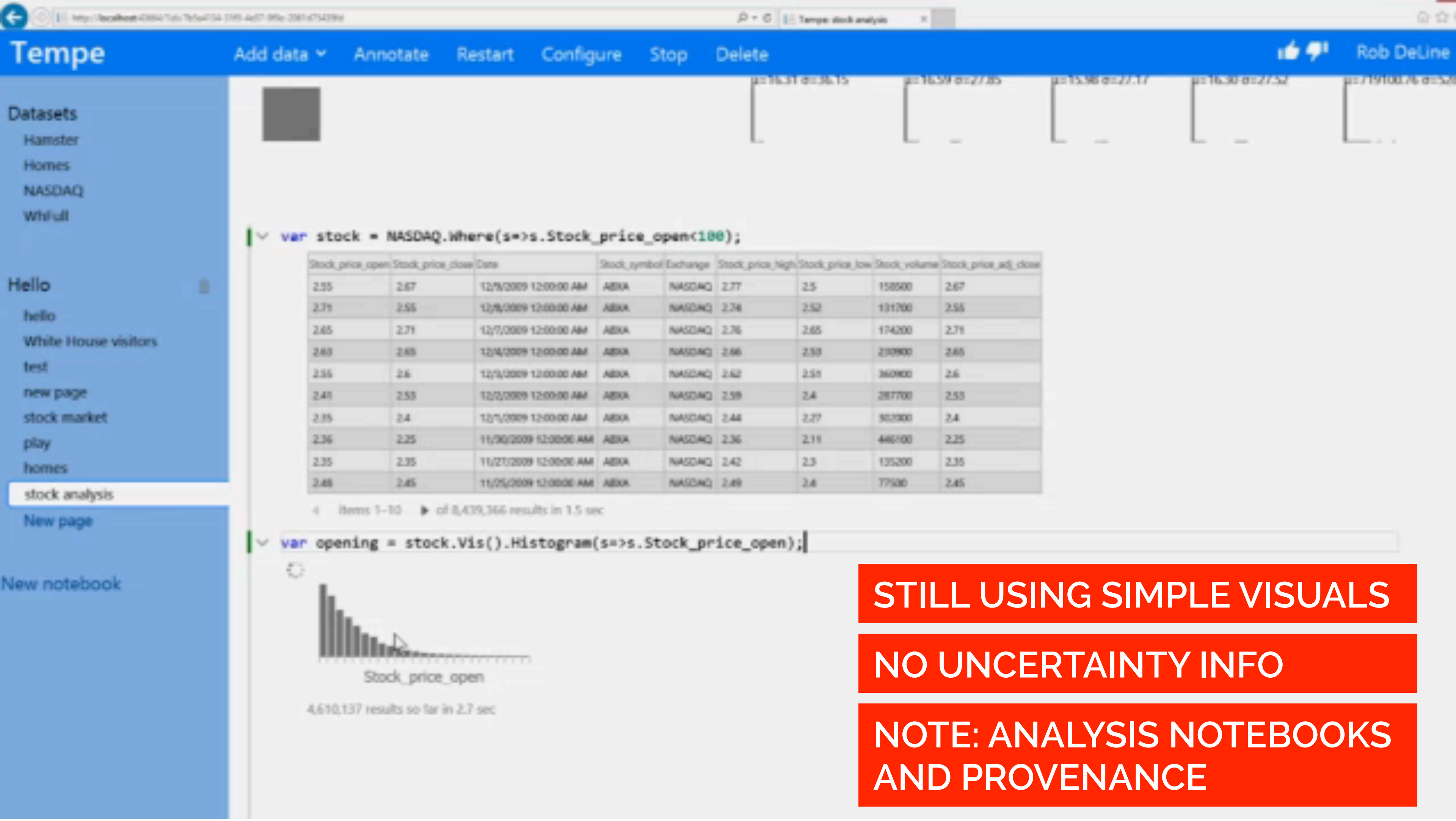




[Boukhelifa, et al. 2012]

PEOPLE DON'T ALWAYS INTERPRET THESE AS SHOWING UNCERTAINTY

A FEW INTERESTING RESEARCH PROTOTYPES



```
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	158500	2.67
2.71	2.55	12/8/2009 12:00:00 AM	ABXA	NASDAQ	2.74	2.52	131700	2.55
2.65	2.71	12/7/2009 12:00:00 AM	ABXA	NASDAQ	2.76	2.65	174200	2.71
2.63	2.65	12/4/2009 12:00:00 AM	ABXA	NASDAQ	2.66	2.53	230900	2.65
2.55	2.6	12/3/2009 12:00:00 AM	ABXA	NASDAQ	2.62	2.51	360900	2.6
2.41	2.53	12/2/2009 12:00:00 AM	ABXA	NASDAQ	2.59	2.4	287700	2.53
2.35	2.4	12/1/2009 12:00:00 AM	ABXA	NASDAQ	2.44	2.27	302000	2.4
2.36	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	2.3	135200	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 8,439,366 results in 1.5 sec

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

4,610,137 results so far in 2.7 sec

STILL USING SIMPLE VISUALS

NO UNCERTAINTY INFO

NOTE: ANALYSIS NOTEBOOKS AND PROVENANCE

TEMPE [Microsoft Research 2014]

DATA ANALYSIS AT SCALE

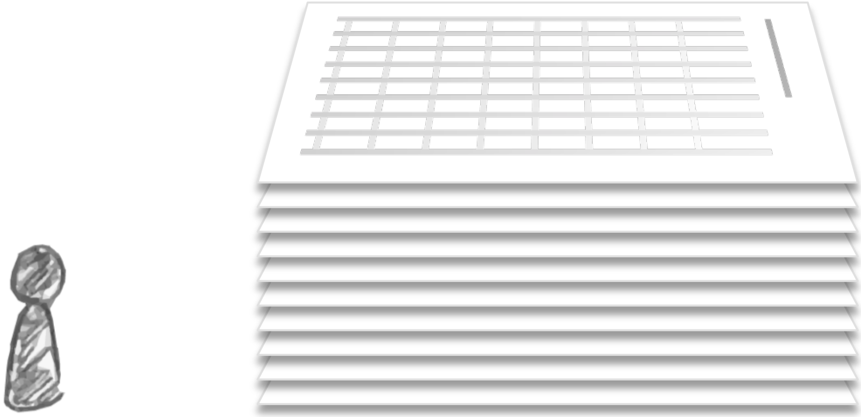
CHALLENGES

ANALYSIS AND CLUSTER COMPUTING

INTERACTING WITH BIG DATA

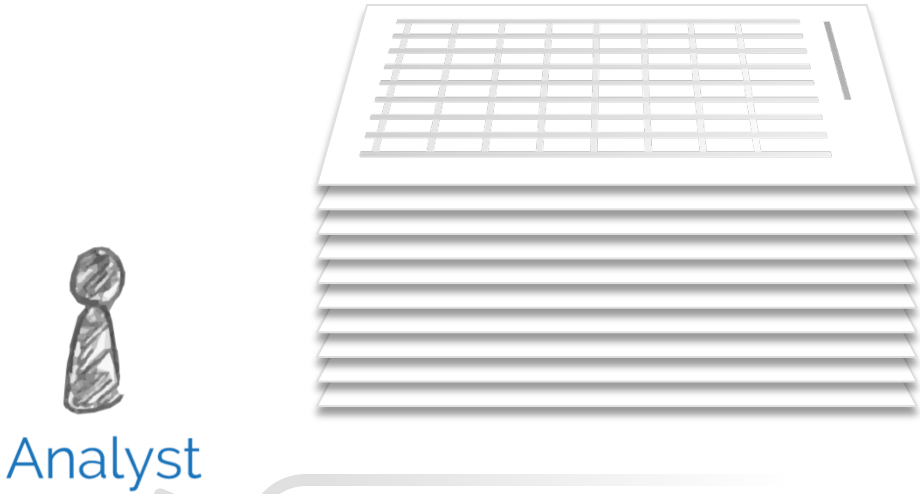
PARALLELIZING HUMAN INTELLIGENCE

HOW CAN WE LEVERAGE MULTIPLE PEOPLE TO EXPEDITE ANALYSIS?

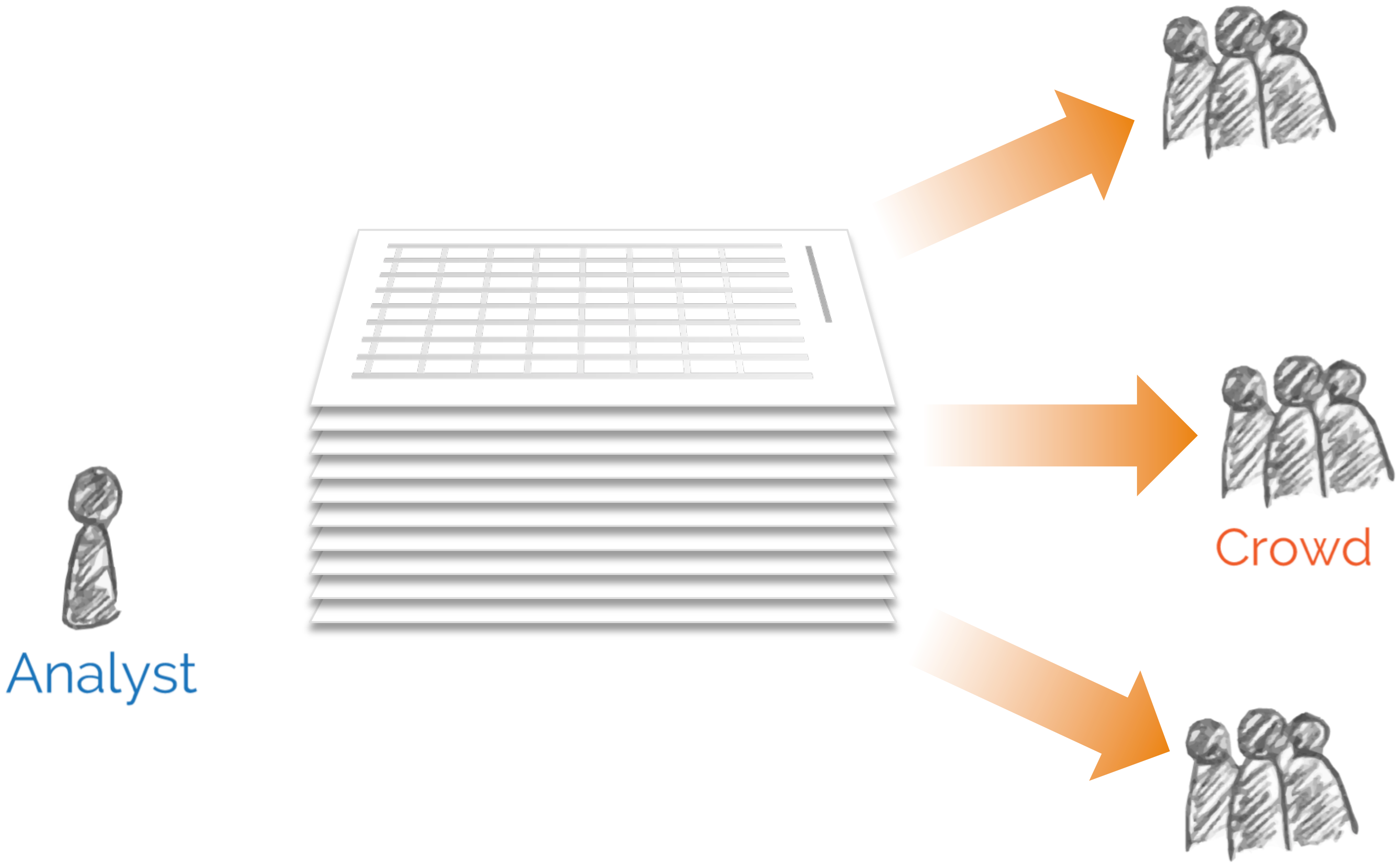


Analyst

CollegeRankings2013.csv



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



Analyst

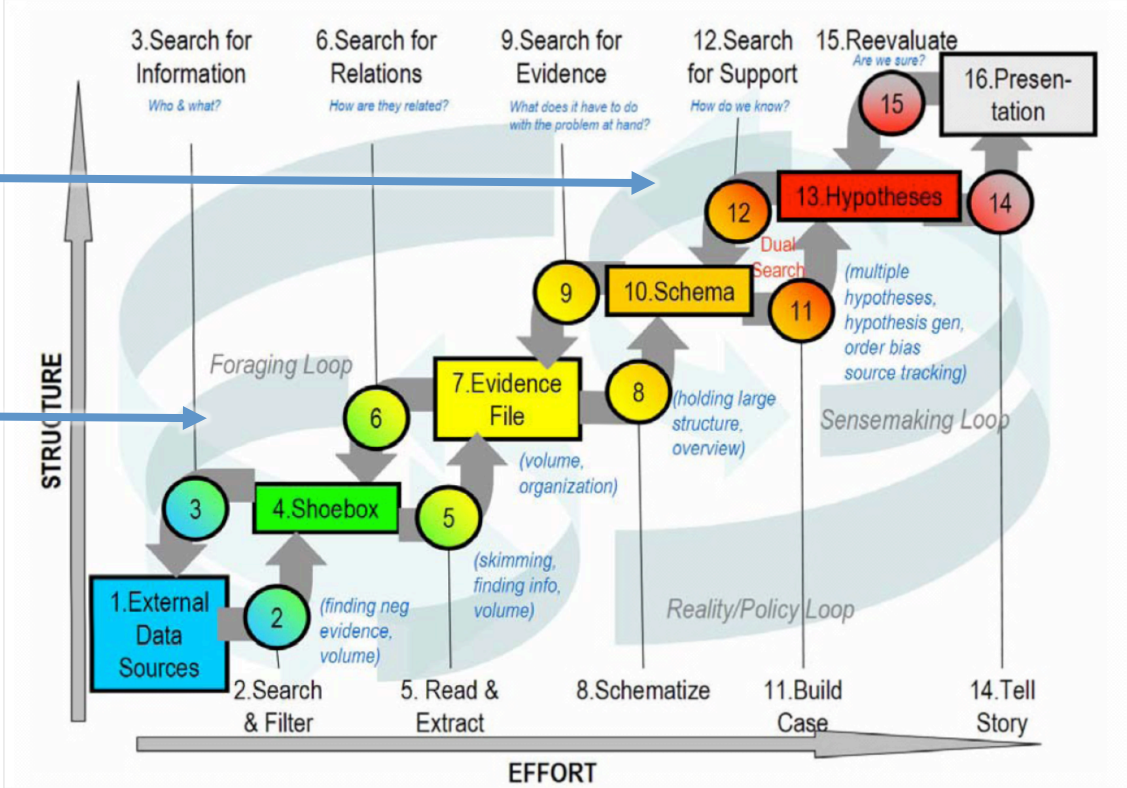
Crowd

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**

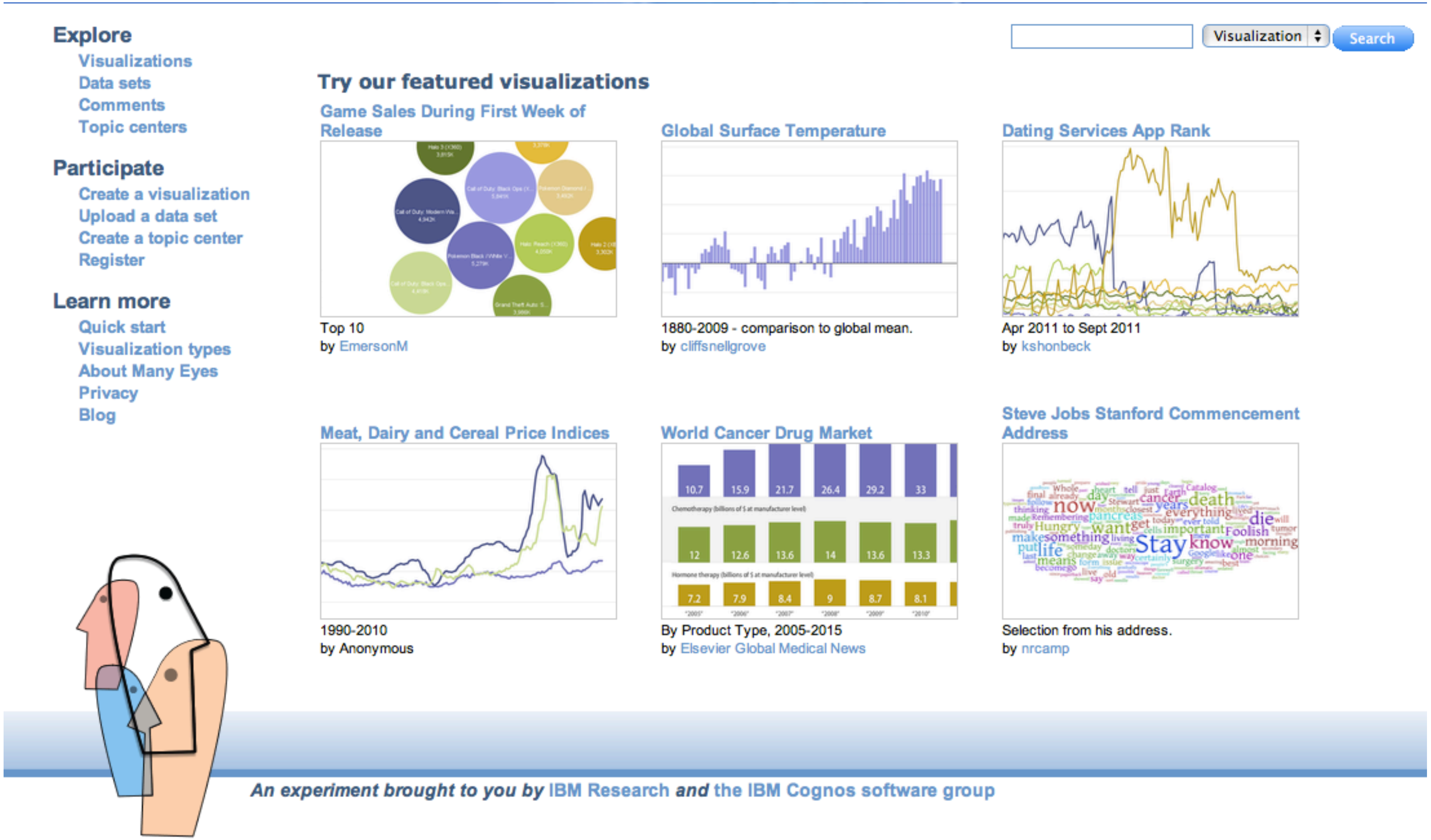
Sensemaking
Loop

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

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

10.7
15.9
21.7
26.4
29.2
33
Chemotherapy (billions of $ at manufacturer level)
12
12.6
13.6
14
13.6
13.3
Hormone therapy (billions of $ at manufacturer level)
7.2
7.9
8.4
9
8.7
8.1

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

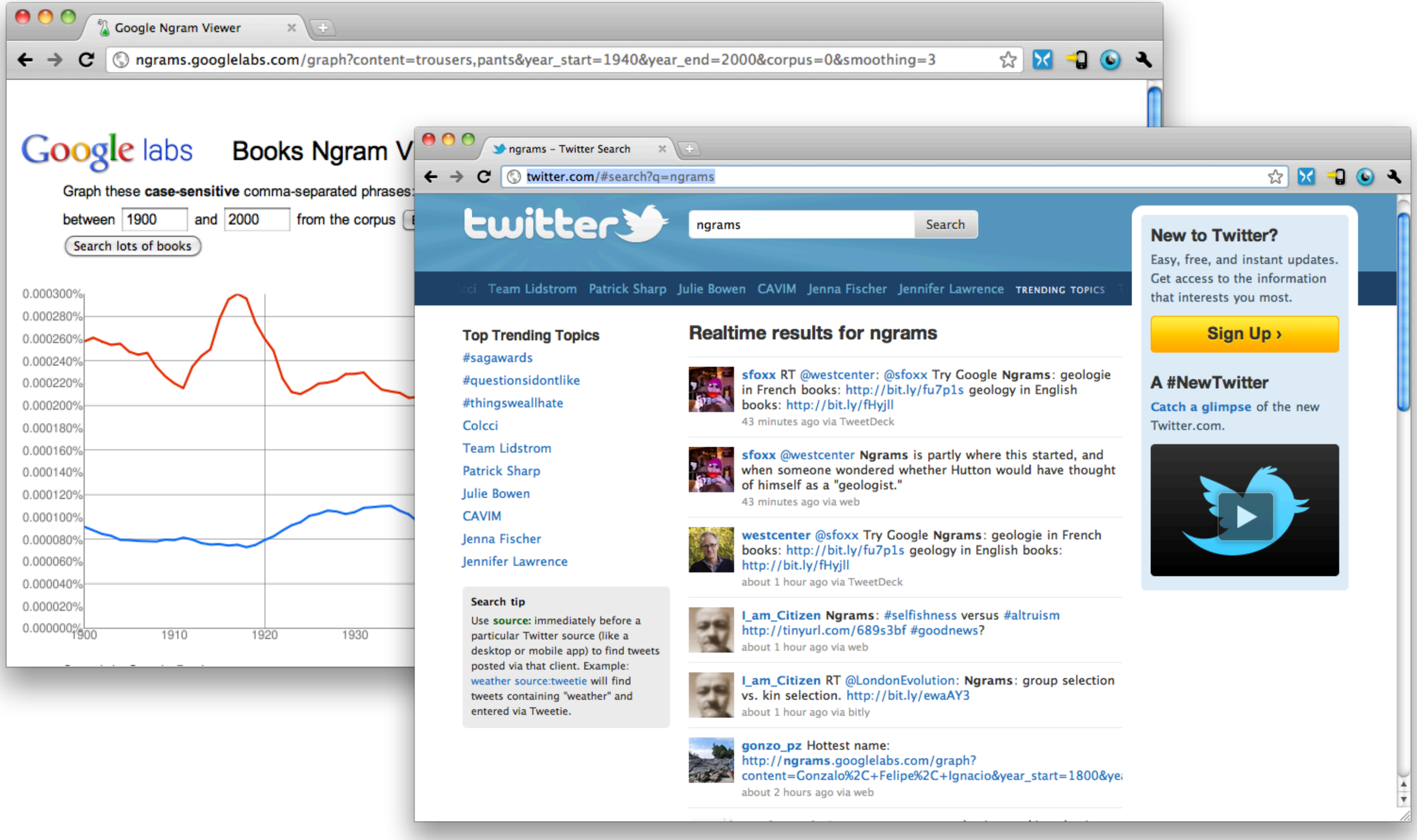
Steve Jobs Stanford Commencement Address

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



Google Ngram Viewer
ngrams.googlelabs.com/graph?content=trousers,pants&year_start=1940&year_end=2000&corpus=0&smoothing=3
Google labs
Books Ngram V

Graph these **case-sensitive** comma-separated phrases:

between 1900 and 2000 from the corpus

Search lots of books

Top Trending Topics

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

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.

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
about 2 hours ago via web

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

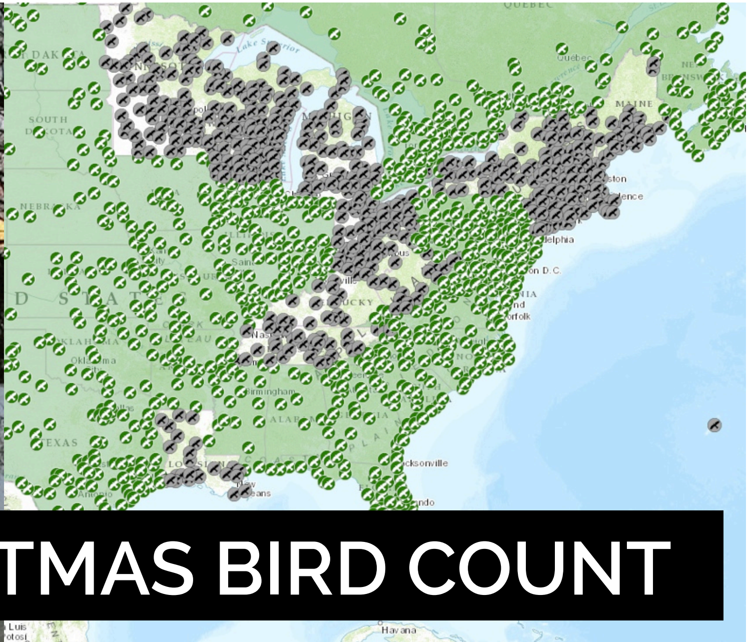
CITIZEN SCIENCE

DATA COLLECTION

CREEK WATCH

[IBM]

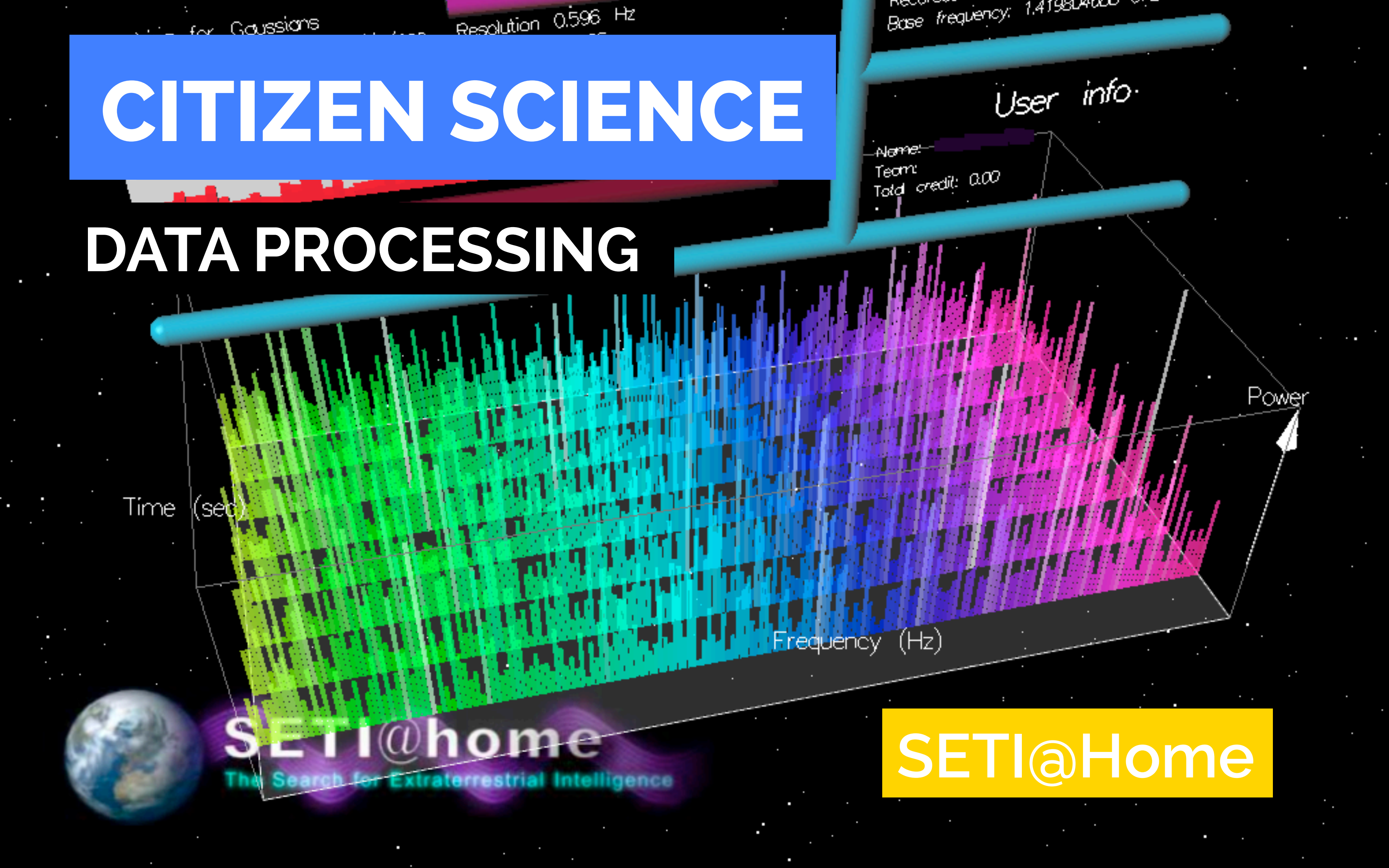
4,000
Creek Watch
users
in over
25
countries



CHRISTMAS BIRD COUNT

CITIZEN SCIENCE

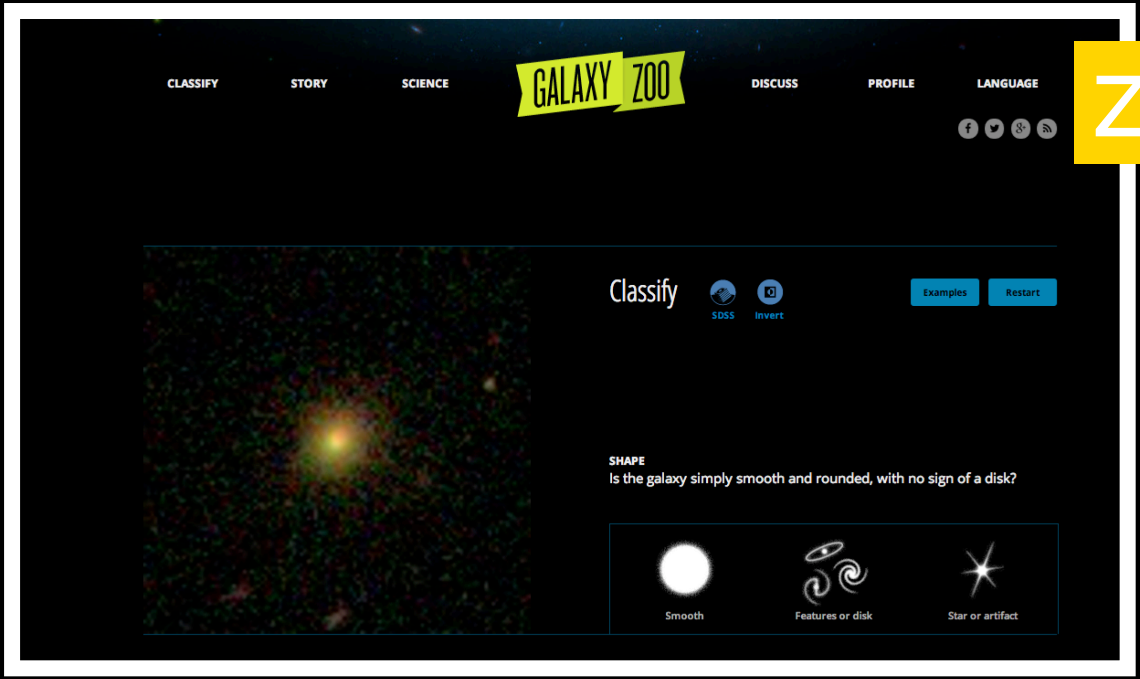
DATA PROCESSING



SETI@Home

CITIZEN SCIENCE

HUMAN VISION & PROBLEM SOLVING



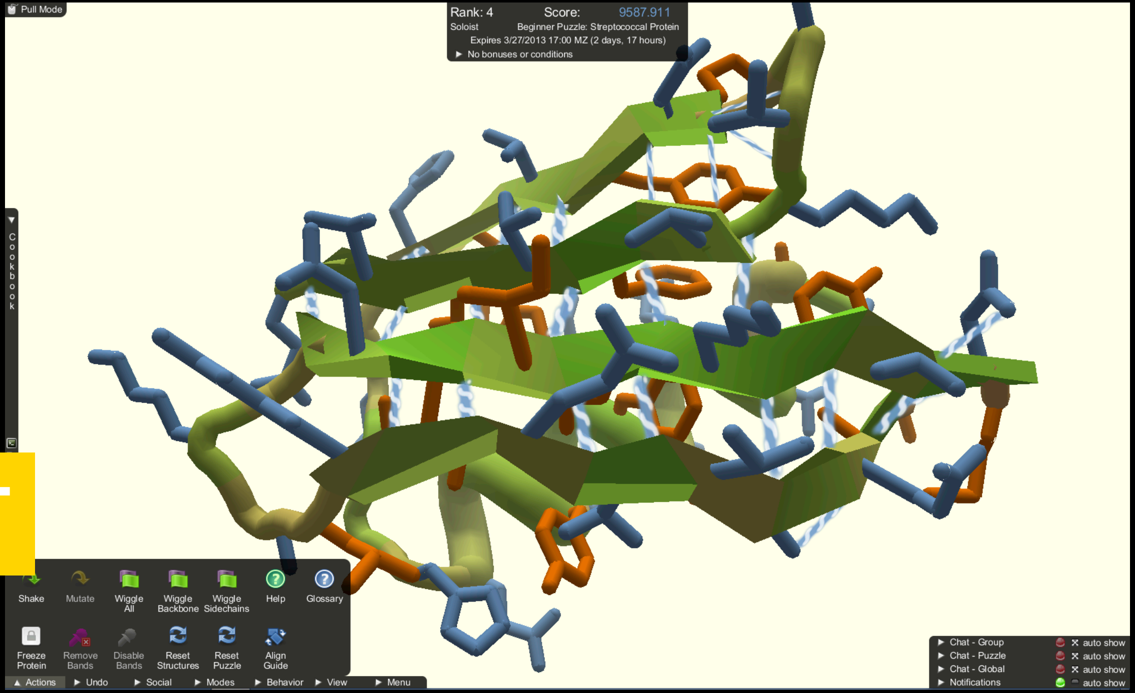
SHAPE

Is the galaxy simply smooth and rounded, with no sign of a disk?

ZOONIVERSE

CITIZEN SCIENCE

HUMAN VISION & PROBLEM SOLVING



Rank: 4 Score: 9587.911

Soloist Beginner Puzzle: Streptococcal Protein

Expires 3/27/2013 17:00 MZ (2 days, 17 hours)

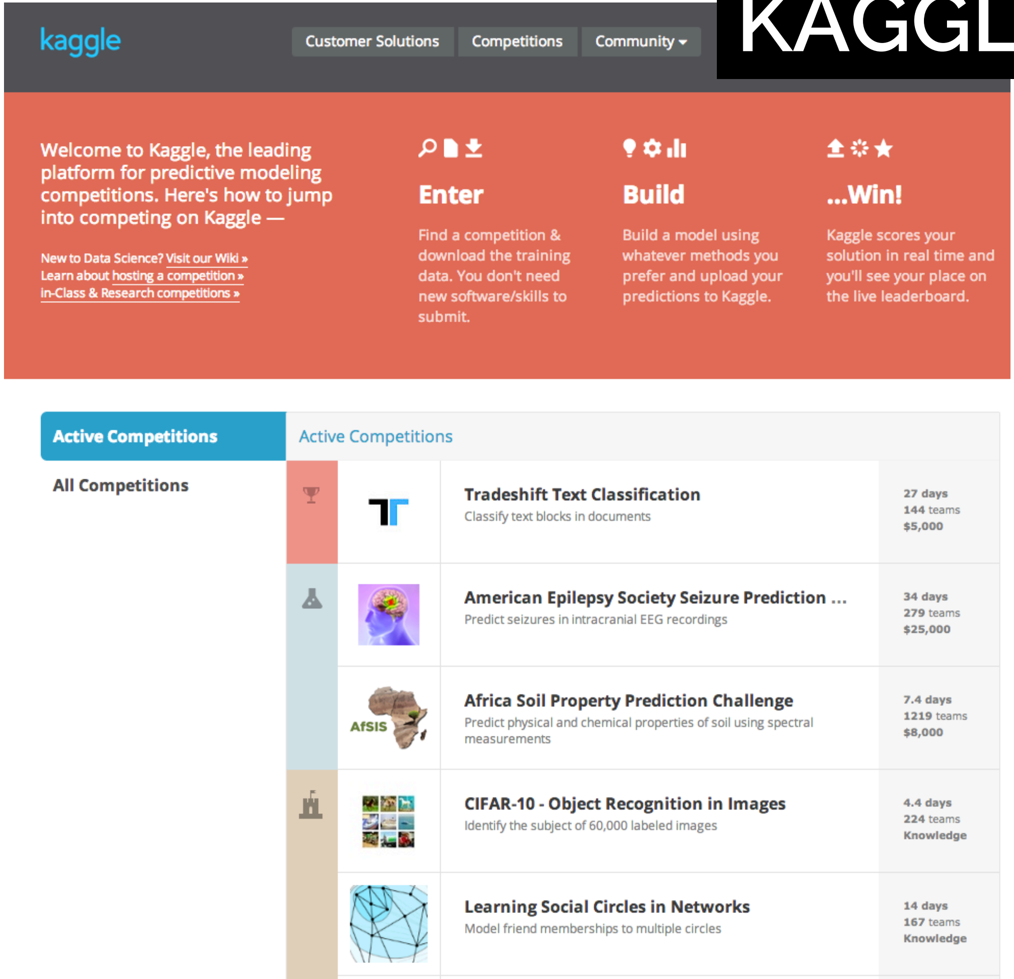
▶ No bonuses or conditions

FOLD.IT

ANALYSIS COMPETITIONS

NETFLIX PRIZE

KAGGLE



Welcome to Kaggle, the leading platform for predictive modeling competitions. Here's how to jump into competing on Kaggle —

New to Data Science? Visit our Wiki »
Learn about hosting a competition »
In-Class & Research competitions »

Enter

Find a competition & download the training data. You don't need new software/skills to submit.

Build

Build a model using whatever methods you prefer and upload your predictions to Kaggle.

...Win!

Kaggle scores your solution in real time and you'll see your place on the live leaderboard.

Active Competitions

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

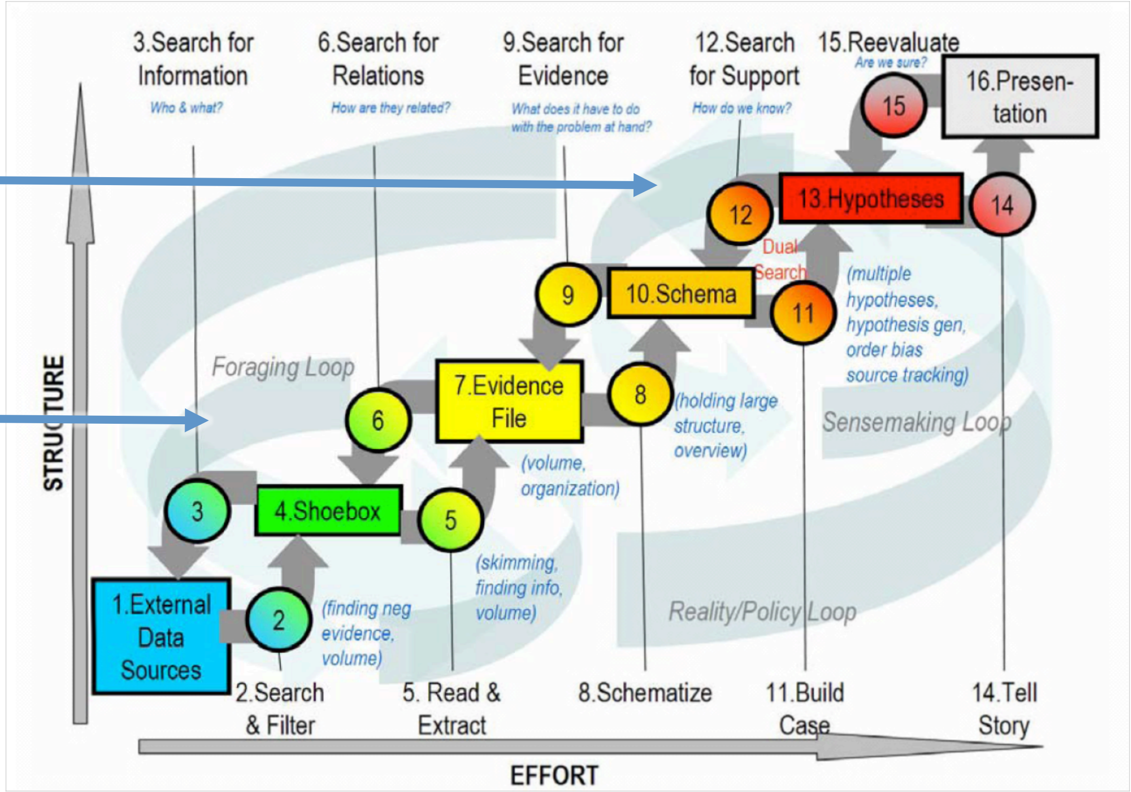
DATA CLEANING

SENTIMENT ANALYSIS

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

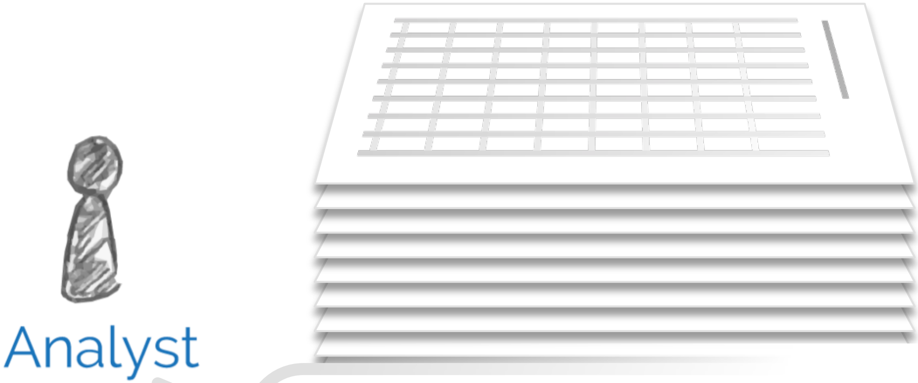
Sensemaking
Loop

Foraging
Loop



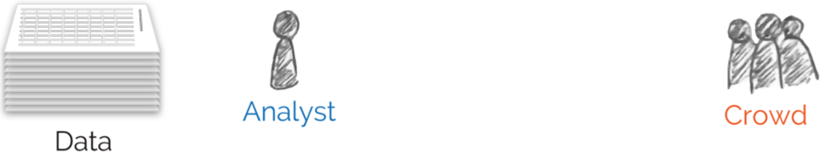
[Pirolli & Card 2005]

CROWDSOURCING HIGHER-LEVEL ANALYSIS TASKS



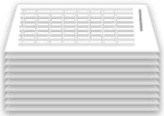
"Can I screen this dataset to **quickly** find the **most interesting** parts?

A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



[Willett et al. CHI 2012, VAST 2013]

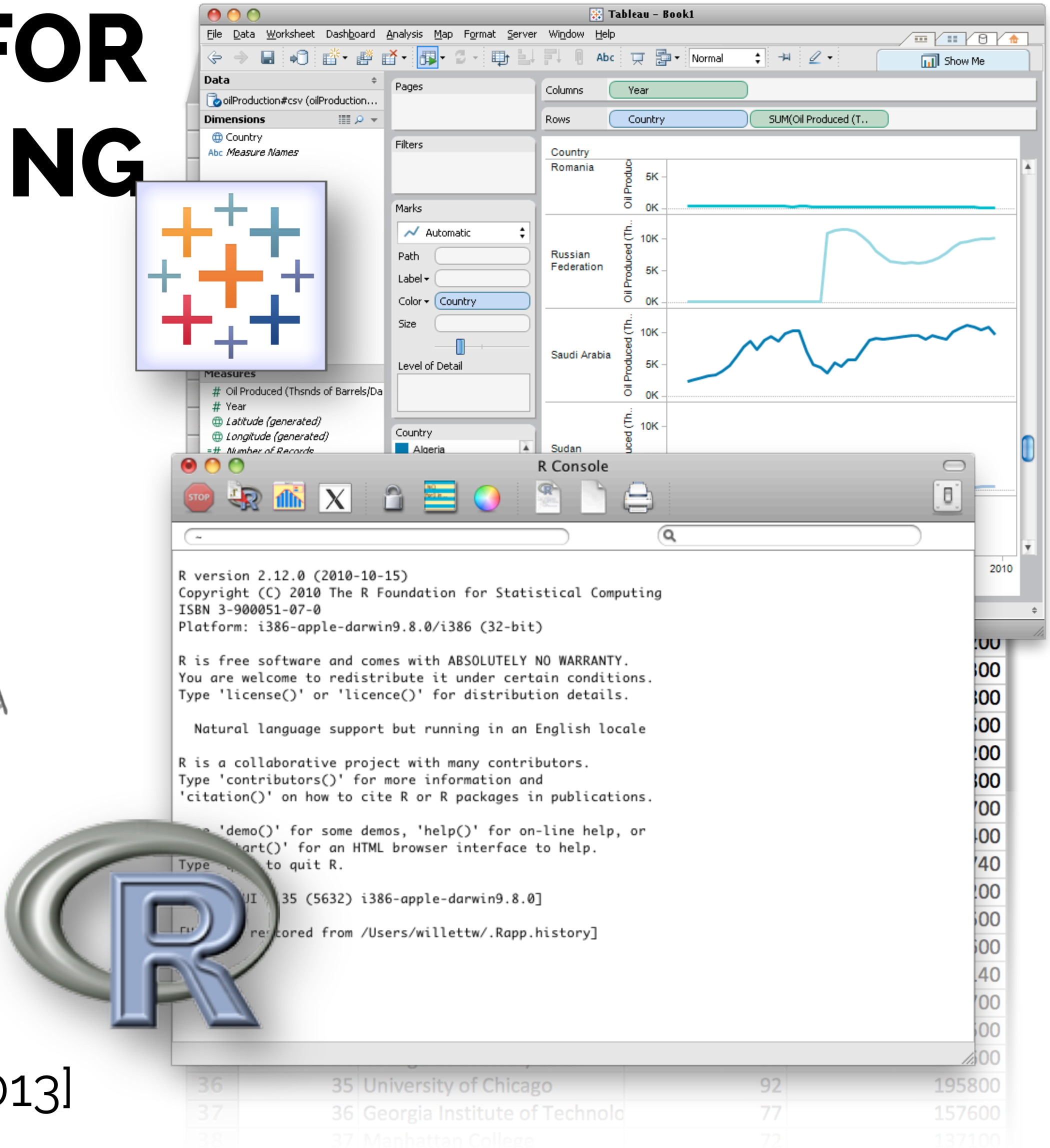
A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



Data

Analyst

Crowd

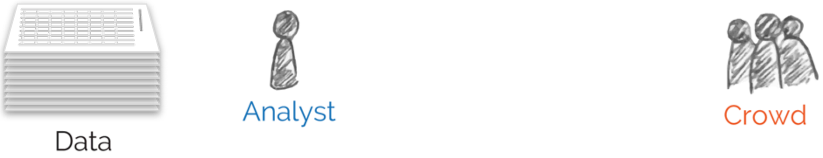


```
R version 2.12.0 (2010-10-15)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

  e 'demo()' for some demos, 'help()' for on-line help, or
    art()' for an HTML browser interface to help.
Type      to quit R.

      UI  35 (5632) i386-apple-darwin9.8.0]

[H    restored from /Users/willettw/.Rapp.history]
```

36	35 University of Chicago	92	195800
37	36 Georgia Institute of Technolo	77	157600

[Willett et al. CHI 2012, VAST 2013]

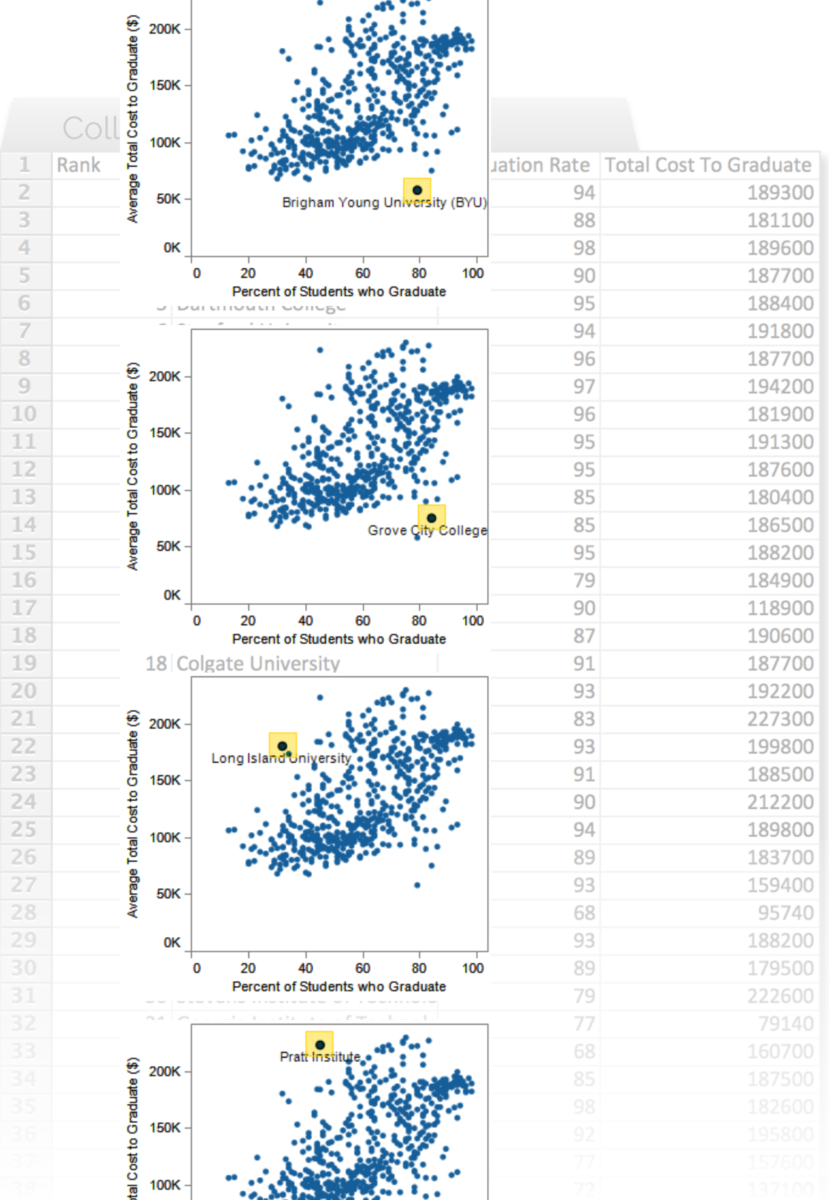
A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



Data

Analyst

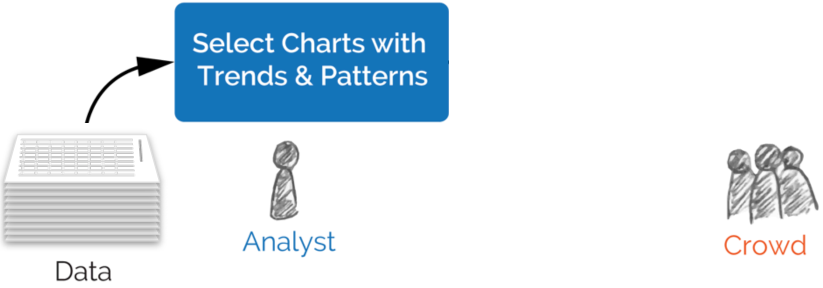
Crowd



	Rank	…uation Rate	Total Cost To Graduate
2		94	189300
3		88	181100
4		98	189600
5		90	187700
6		95	188400
7		94	191800
8		96	187700
9		97	194200
10		96	181900
11		95	191300
12		95	187600
13		85	180400
14		85	186500
15		95	188200
16		79	184900
17		90	118900
18		87	190600
19		91	187700
20		93	192200
21		83	227300
22		93	199800
23		91	188500
24		90	212200
25		94	189800
26		89	183700
27		93	159400
28		68	95740
29		93	188200
30		89	179500
31		79	222600
32		77	79140
33		68	160700
34		85	187500
35		98	182600
36		92	195800
37		77	157600
38		72	137100

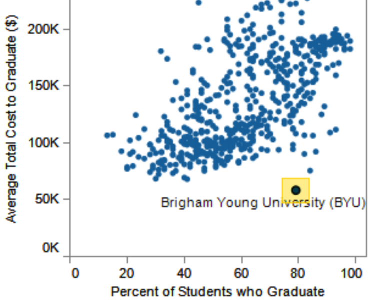
[Willett et al. CHI 2012, VAST 2013]

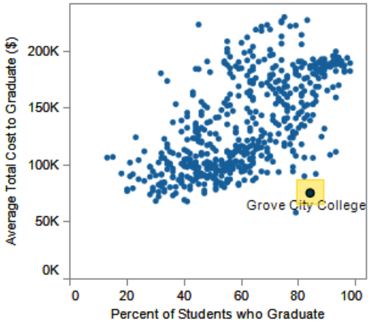
A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS

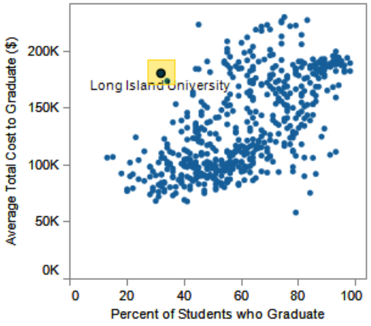


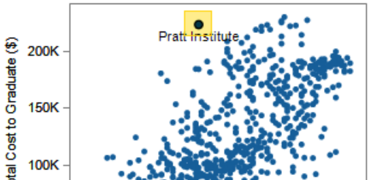
Select Charts with Trends & Patterns

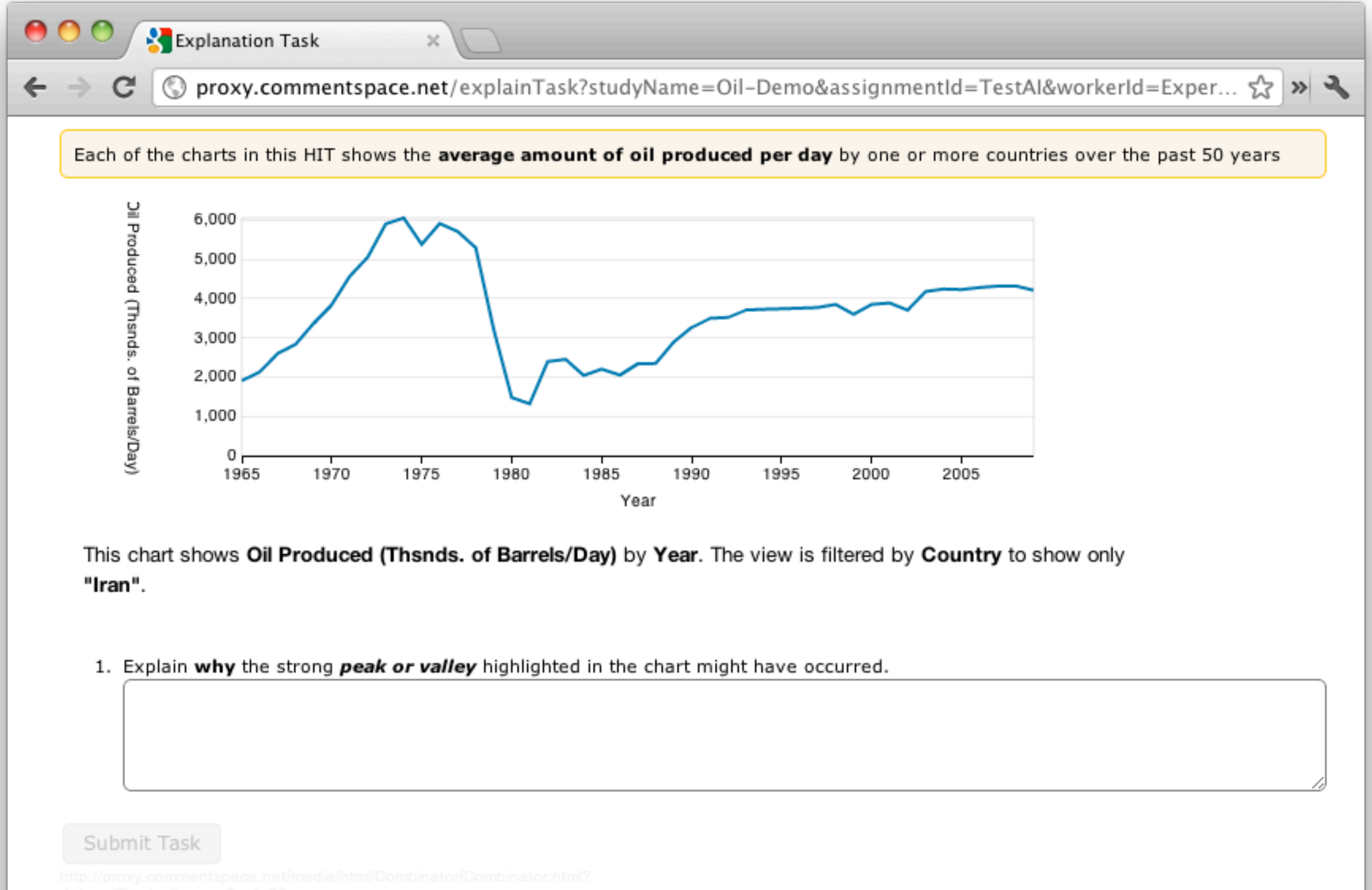
[Willett et al. CHI 2012, VAST 2013]











Explanation Task
proxy.commentspace.net/explainTask?studyName=Oil-Demo&assignmentId=TestAI&workerId=Exper…

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

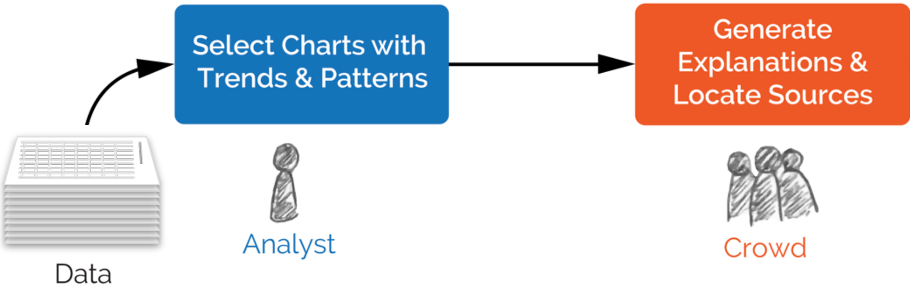
Oil Produced (Thsnds. of Barrels/Day)
6,000
5,000
4,000
3,000
2,000
1,000
0
1965
1970
1975
1980
1985
1990
1995
2000
2005
Year

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
http://proxy.commentspace.net/media/html/Combinator/Combinator.html?

A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS



Select Charts with Trends & Patterns

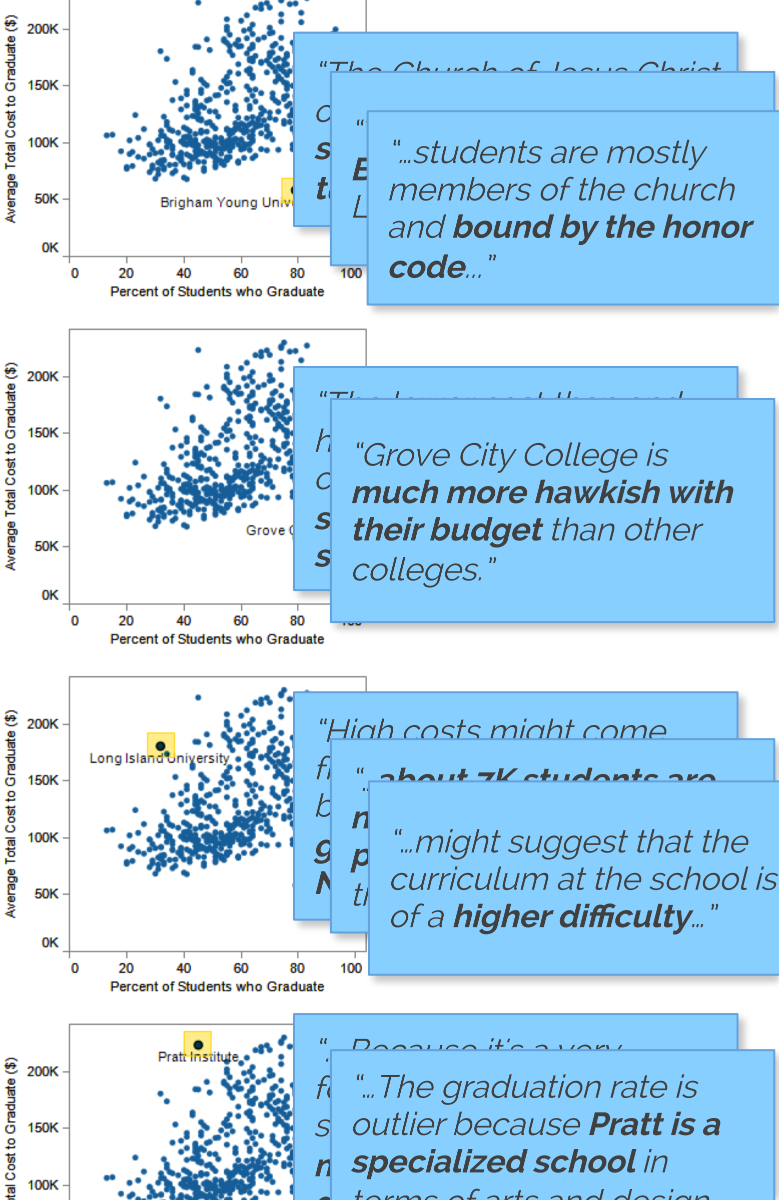
Generate Explanations & Locate Sources

Data

Analyst

Crowd

[Willett et al. CHI 2012, VAST 2013]



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

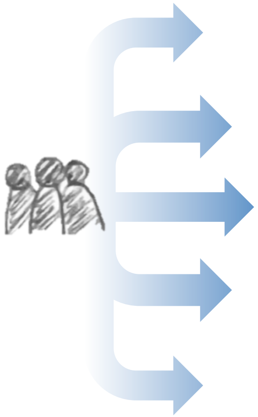
"Grove City College is **much more hawkish with their budget** than other colleges."

"…might suggest that the curriculum at the school is of a **higher difficulty**…"

"…The graduation rate is outlier because **Pratt is a specialized school** in terms of arts and design

“COULD THIS CREATE MORE WORK FOR THE ANALYST?”

“COULD THIS CREATE MORE WORK FOR THE ANALYST?”



“High costs might come

“...about 7K students are

“...might suggest that the curriculum at the school is of a **higher difficulty**...”

“The lower cost than and

“Grove City College is **much more hawkish with their budget** than other colleges.”

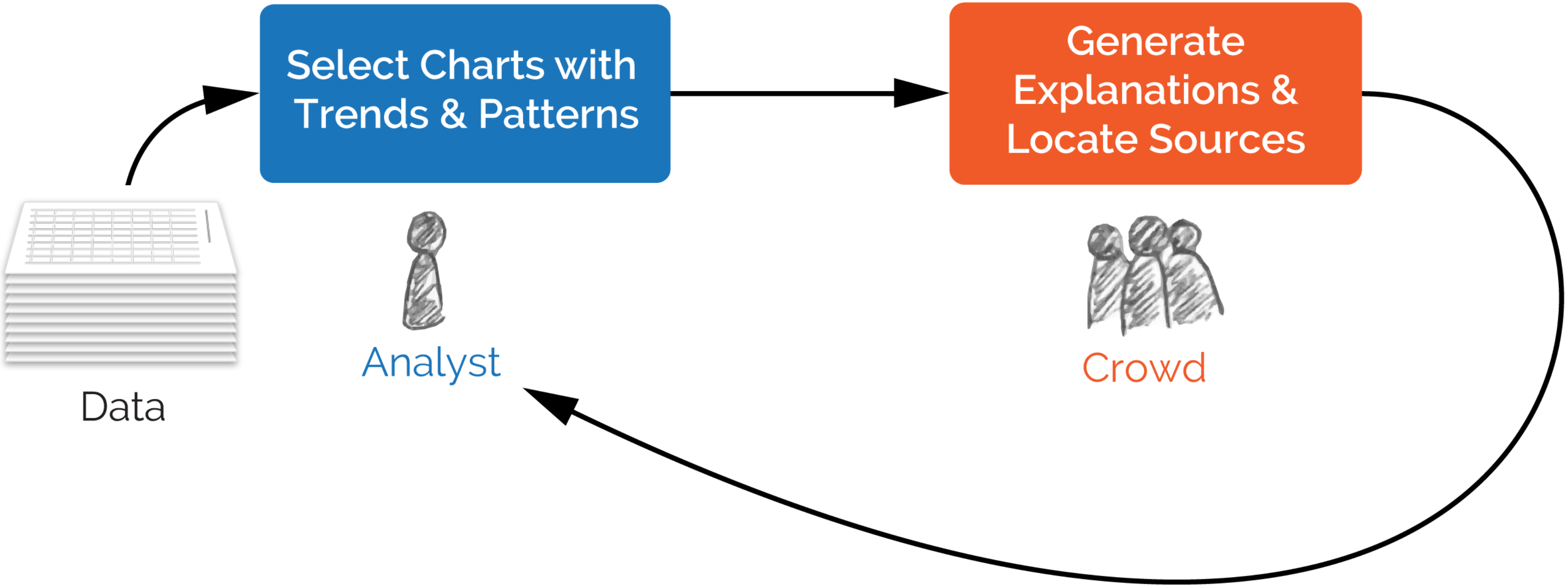
“The Church of Jesus Christ

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

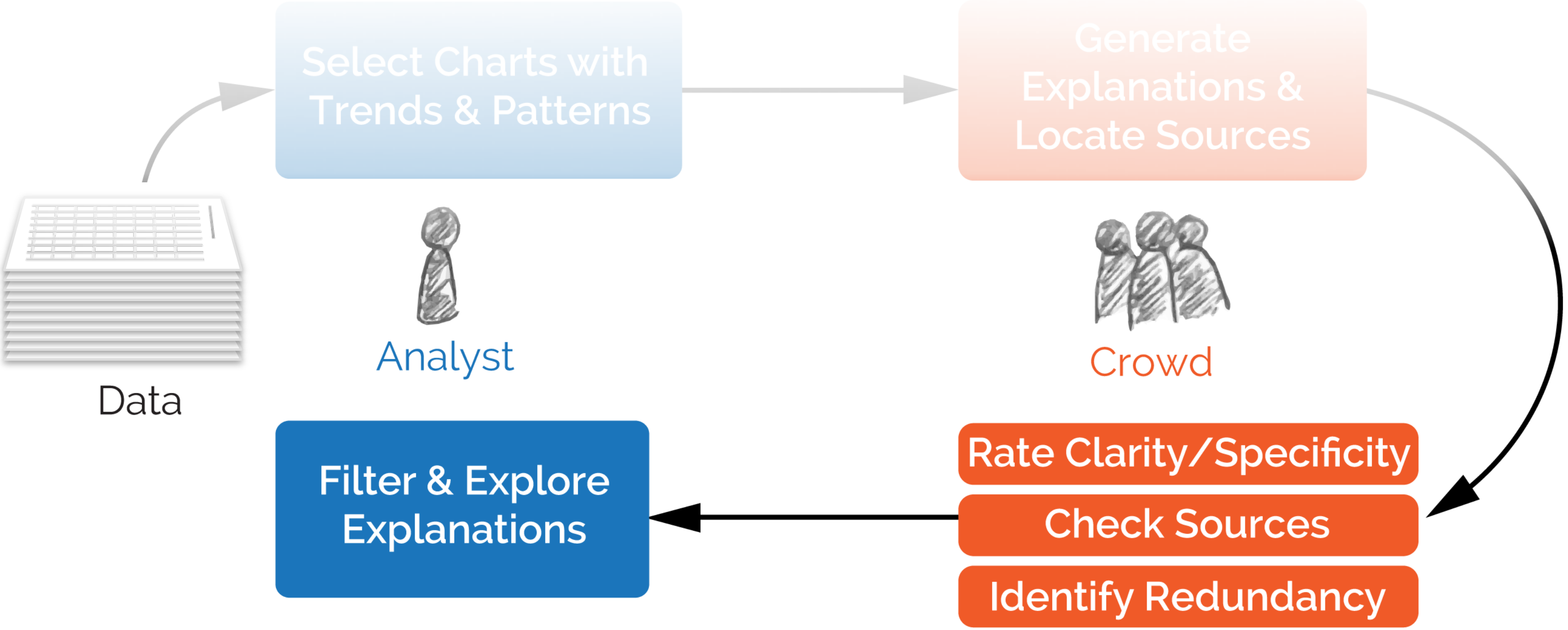
“...Because it’s a very

“...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


Select Charts with Trends & Patterns
Generate Explanations & Locate Sources
Data
Analyst
Crowd

CROWD-ENABLED EXTENSIONS FOR PROCESSING AND MANAGING RESULTS


Select Charts with Trends & Patterns
Generate Explanations & Locate Sources
Data
Analyst
Crowd
Rate Clarity/Specificity
Check Sources
Identify Redundancy
Filter & Explore Explanations

THREE CRITERIA FOR PLAUSIBLE EXPLANATIONS

CLARITY AND SPECIFICITY

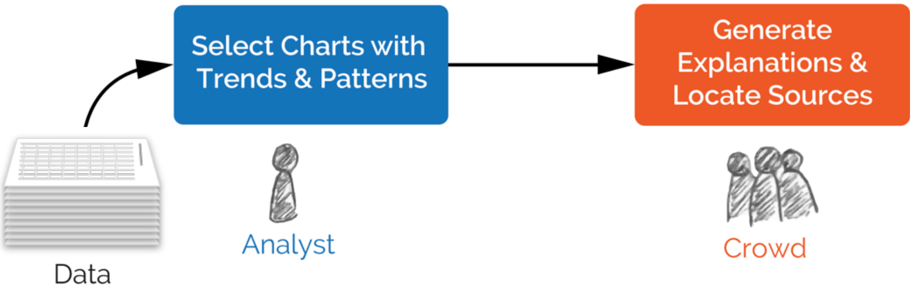
PROVENANCE

REDUNDANCY

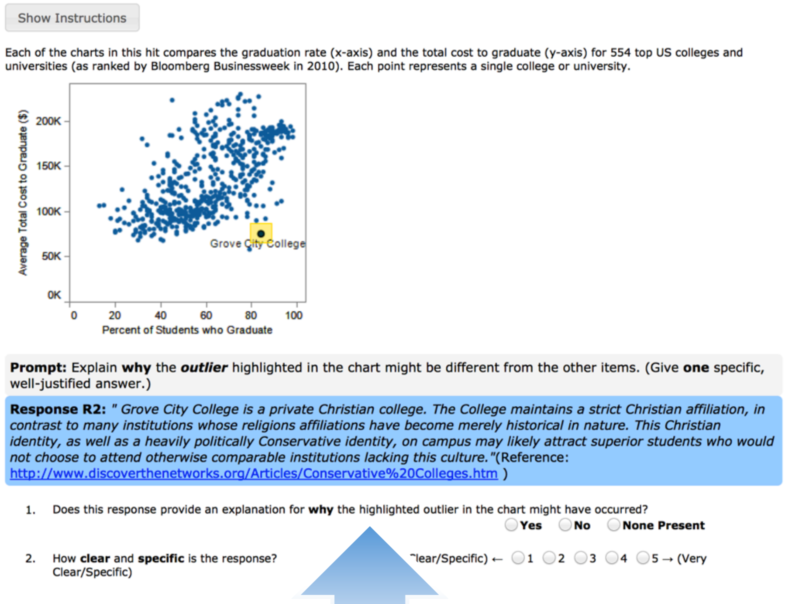
+ AN INTERFACE FOR MANAGING CROWDSOURCED EXPLANATIONS

CLARITY & SPECIFICITY

CLARITY AND SPECIFICITY


Select Charts with Trends & Patterns
Generate Explanations & Locate Sources
Data
Analyst
Crowd

Rating Task

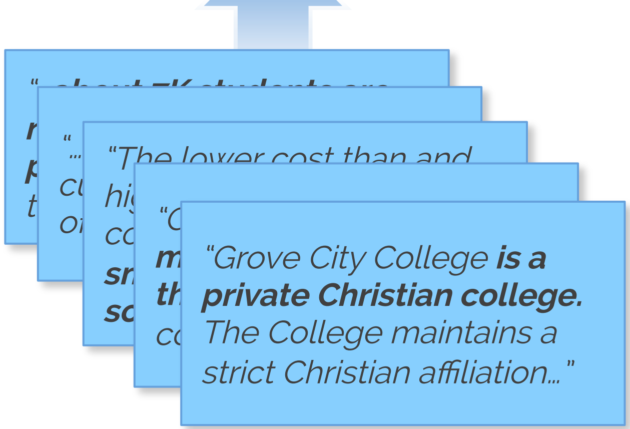

Show Instructions
200K
150K
100K
50K
0K
0
20
40
60
80
100
Average Total Cost to Graduate ($)
Percent of Students who Graduate
Grove City College

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.)

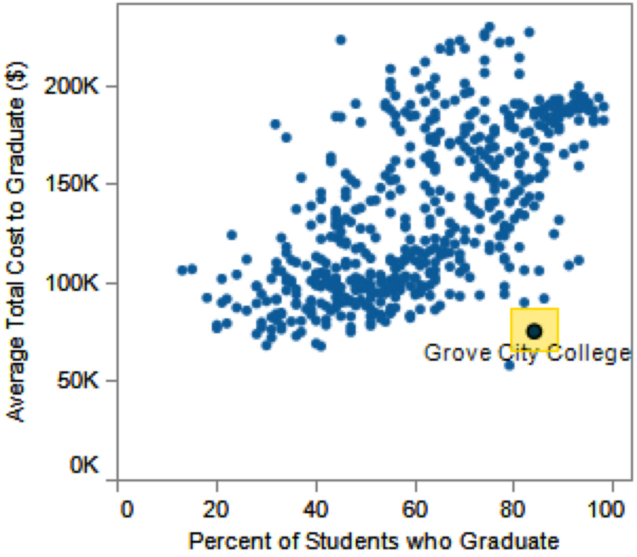
Response R2: *" Grove City College is a private Christian college. The College maintains a strict Christian affiliation, in contrast to many institutions whose religions affiliations have become merely historical in nature. This Christian identity, as well as a heavily politically Conservative identity, 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? lear/Specific) ← ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 → (Very Clear/Specific)



*“Grove City College **is a private Christian college.** The College maintains a strict Christian affiliation…”*

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.)

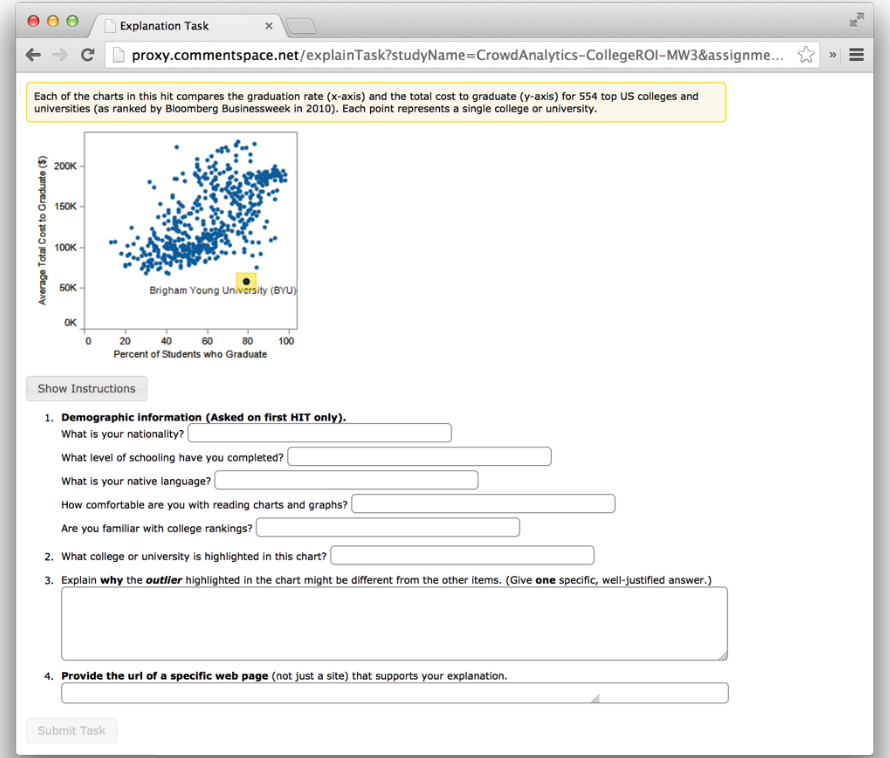
Response R2: *" Grove City College is a private Christian college. The College maintains a strict Christian affiliation, in contrast to many institutions whose religions affiliations have become merely historical in nature. This Christian identity, as well as a heavily politically Conservative identity, 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...
Average Total Cost to Graduate ($)
200K
150K
100K
50K
0K
0
20
40
60
80
100
Percent of Students who Graduate
Brigham Young University (BYU)
Show Instructions
Submit Task

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.

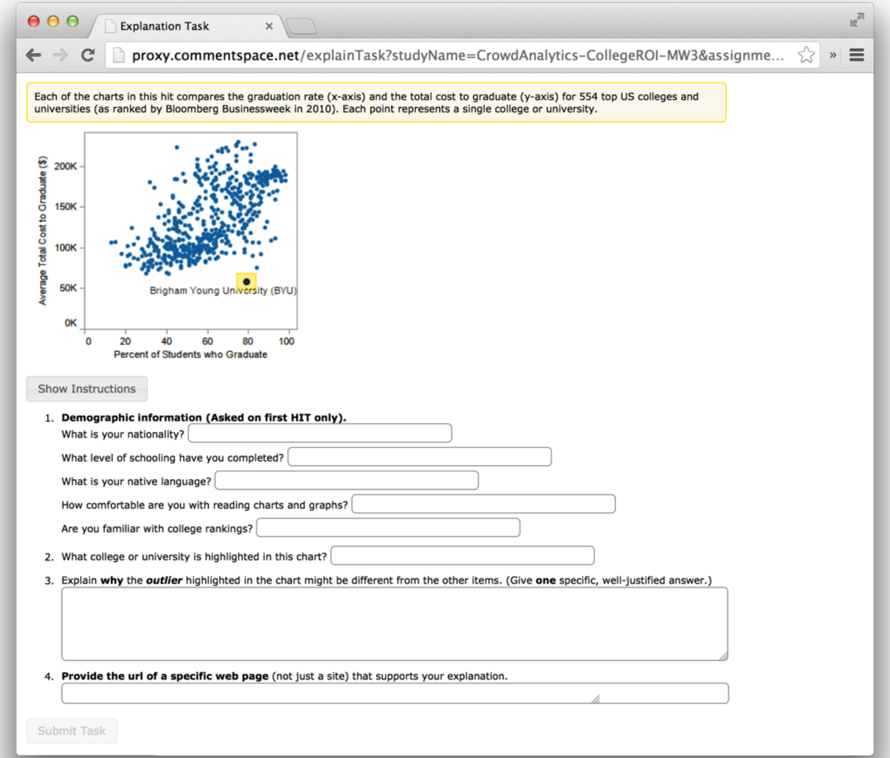
1. **Demographic information (Asked on first HIT only).**
   What is your nationality?
   What level of schooling have you completed?
   What is your native language?
   How comfortable are you with reading charts and graphs?
   Are you familiar with college rankings?
2. What college or university is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

Explanation Task

What are our

workers doing?

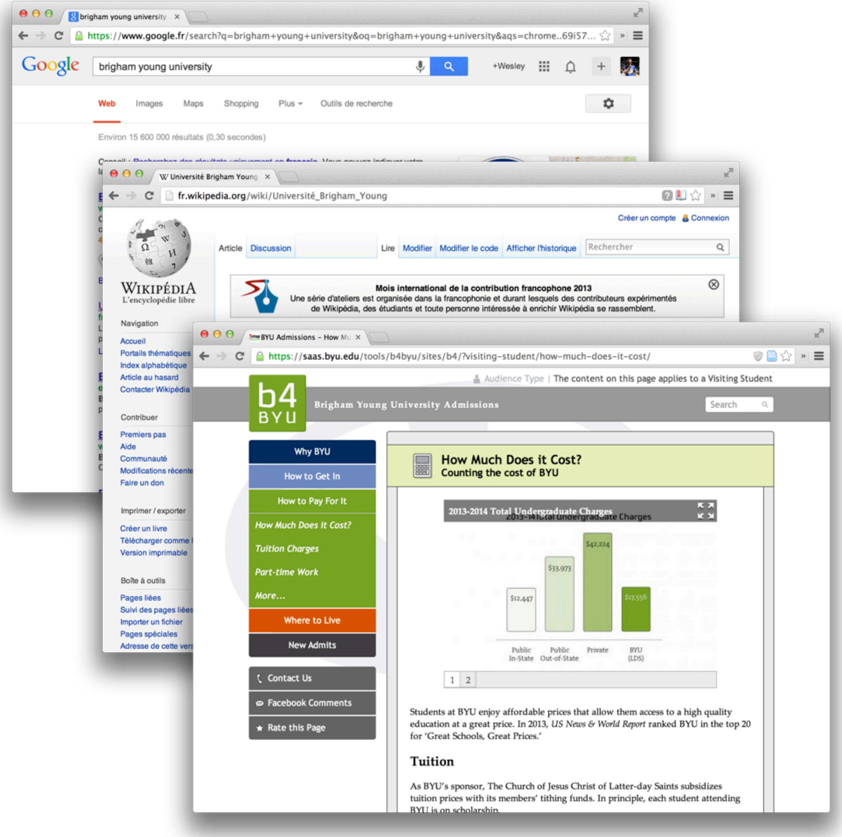
PROVENANCE


Explanation Task
proxy.commentspace.net/explainTask?studyName=CrowdAnalytics-CollegeROI-MW3&assignme...
Average Total Cost to Graduate ($)
200K
150K
100K
50K
0K
0
20
40
60
80
100
Percent of Students who Graduate
Brigham Young University (BYU)
Show Instructions
Submit Task

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.

1. **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?
2. What college or university is highlighted in this chart?
3. Explain **why** the ***outlier*** highlighted in the chart might be different from the other items. (Give **one** specific, well-justified answer.)
4. **Provide the url of a specific web page** (not just a site) that supports your explanation.

Explanation Task



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.

The content on this page applies to a Visiting Student

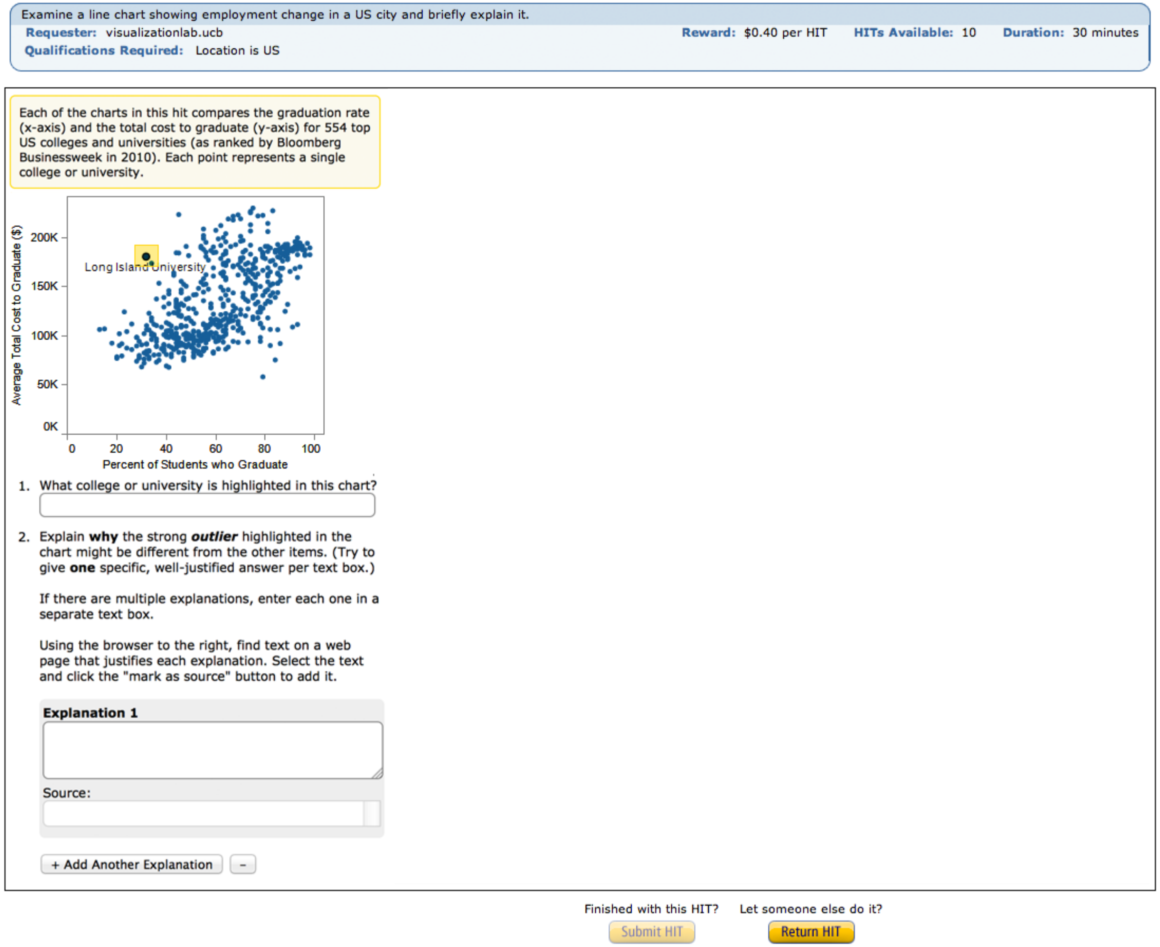
How Much Does it Cost?
Counting the cost of BYU

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 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 ($)
200K
150K
100K
50K
0K
Long Island University
0
20
40
60
80
100
Percent of Students who Graduate

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.)

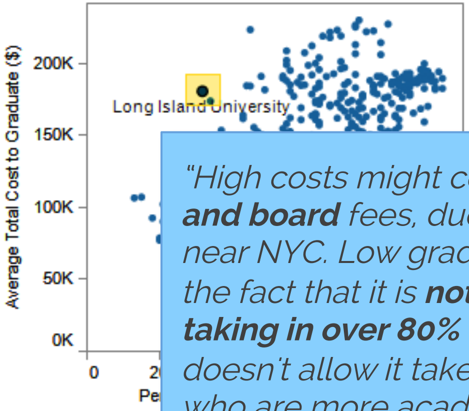
   If there are multiple explanations, enter each one in a separate text box.

   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

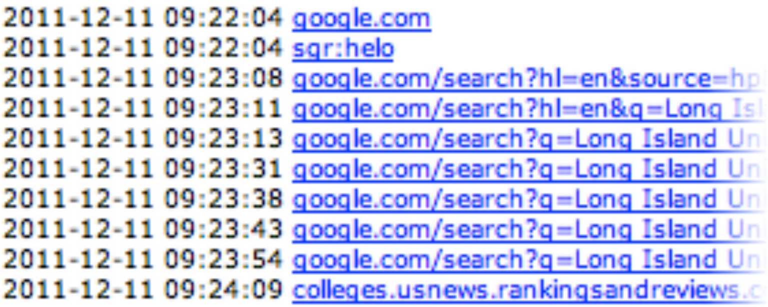
Source:

PROVENANCE



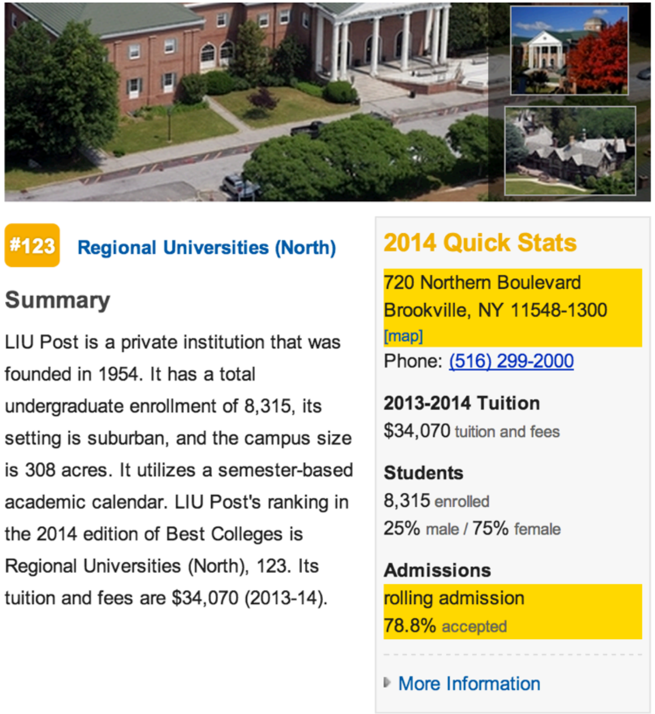
*"High costs might come from it's **high room and board** fees, due to its geographic location near NYC. Low graduation rates come from the fact that it is **not a very selective school, taking in over 80% of applicants**, which doesn't allow it take many top ranked students who are more academically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

Visitation logs



```
2011-12-11 09:22:04 google.com
2011-12-11 09:22:04 sqr:helo
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 Isl
2011-12-11 09:23:13 google.com/search?q=Long Island Un
2011-12-11 09:23:31 google.com/search?q=Long Island Un
2011-12-11 09:23:38 google.com/search?q=Long Island Un
2011-12-11 09:23:43 google.com/search?q=Long Island Un
2011-12-11 09:23:54 google.com/search?q=Long Island Un
2011-12-11 09:24:09 colleges.usnews.rankingsandreviews.c
```

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

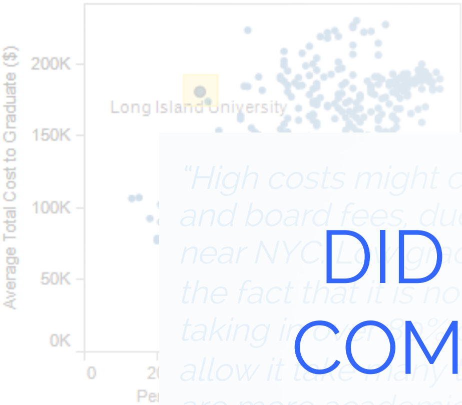
Students
8,315 enrolled
25% male / 75% female

Admissions
rolling admission
78.8% accepted

More Information

PROVENANCE

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

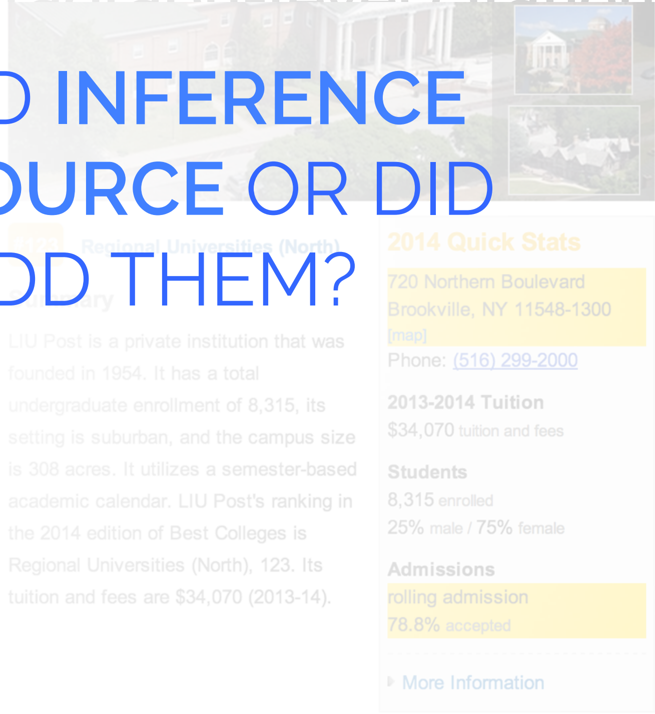


> *"High costs might come from it's high room and board fees, due to its geographic location near NYC. Low graduation rates come from the fact that it is not a very selective school, taking in over 80% of applicants, which doesn't allow it take many top ranked students who are more academically motivated."*

Visitation logs

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

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

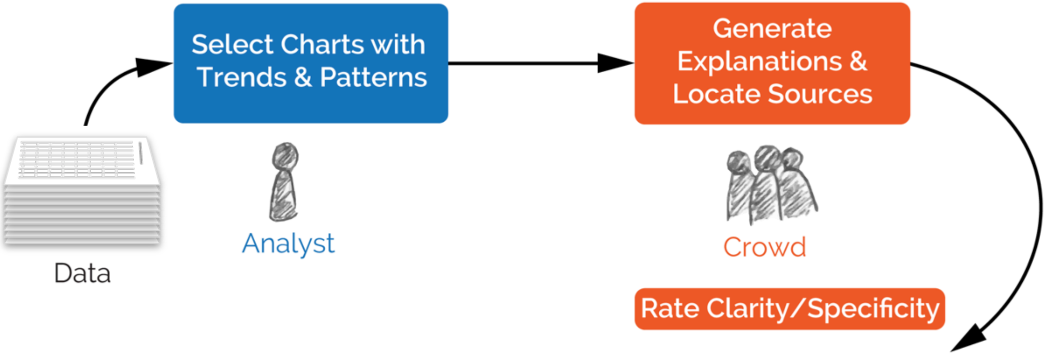
2013-2014 Tuition
$34,070 tuition and fees

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

Admissions
rolling admission
78.8% accepted

More Information

SOURCE-CHECKING MICROTASKS


Select Charts with Trends & Patterns
Generate Explanations & Locate Sources
Data
Analyst
Crowd
Rate Clarity/Specificity

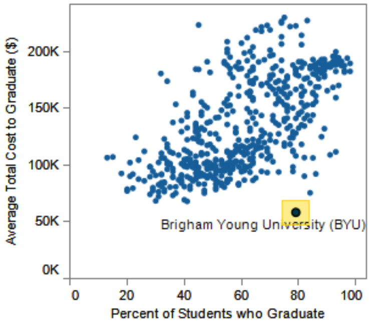
A second group of workers **verifies links** and **attributes explanations** to the **source** or the **worker.**

(75% accurate in our preliminary tests)

REDUNDANCY

REDUNDANCY

Many explanations provided by workers are redundant.



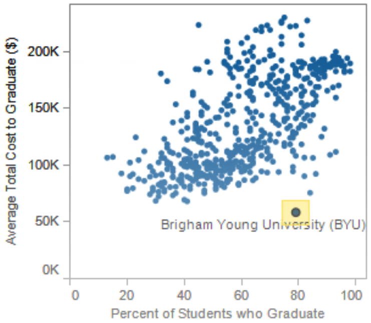
"The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs…"

"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."

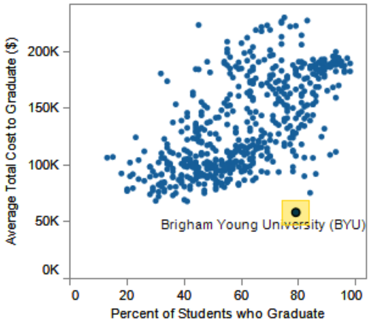
"98% of their students are members of LDS and they have lowered tuition.."

− 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.

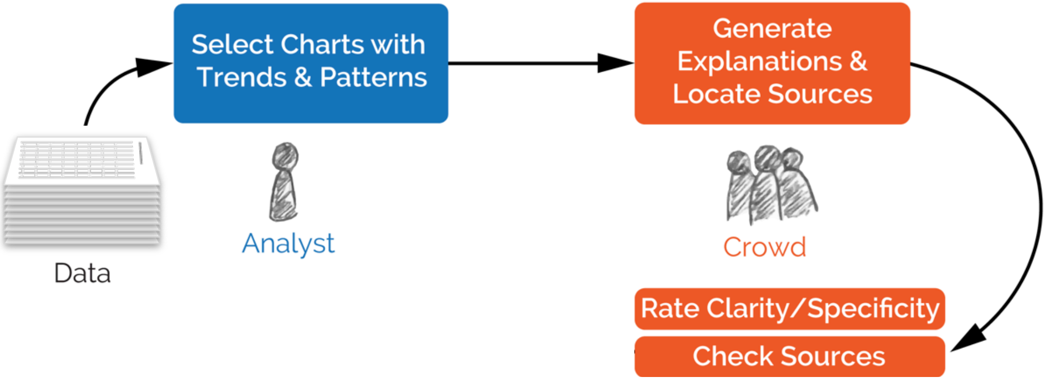


"The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs..."

"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..."

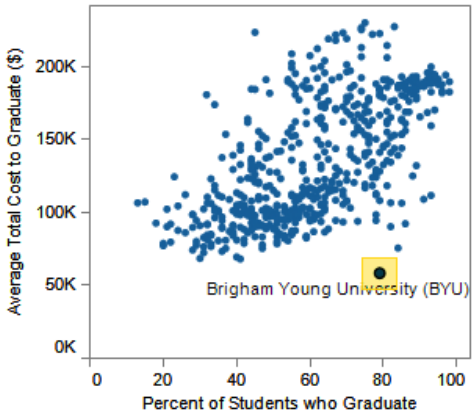
REDUNDANCY



Select Charts with Trends & Patterns
Generate Explanations & Locate Sources
Data
Analyst
Crowd
Rate Clarity/Specificity
Check Sources

Can we crowdsource redundancy detection?

CLUSTERING VIA DISTRIBUTED COMPARISON

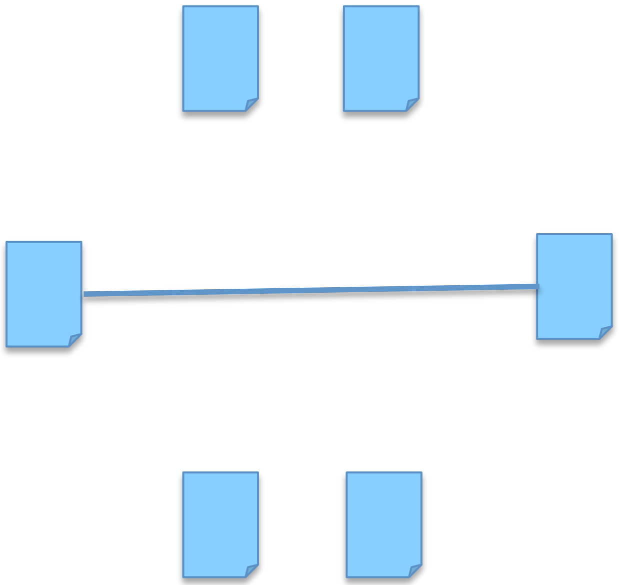


“98% of their students are members of LDS and they have lowered tuition…”

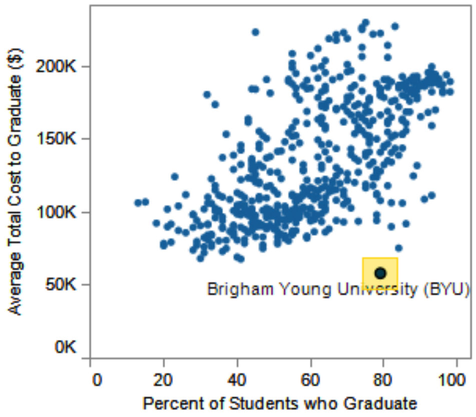
“The cost of attendance at BYU is subsidized by the LDS church.”

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

“The Church of Jesus Christ of Latter Day Saints

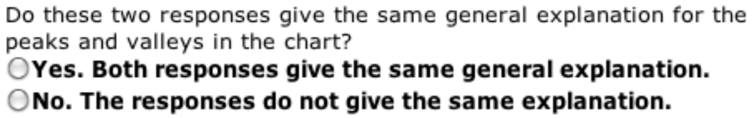


CLUSTERING VIA DISTRIBUTED COMPARISON



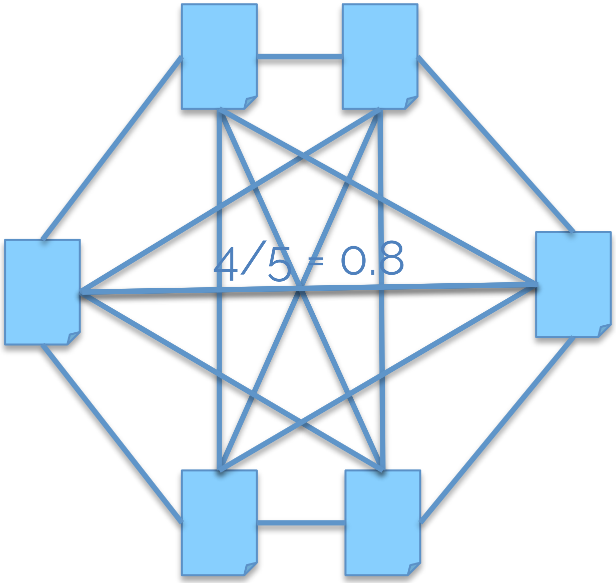
"98% of their students are members of LDS and they have lowered tuition…"

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

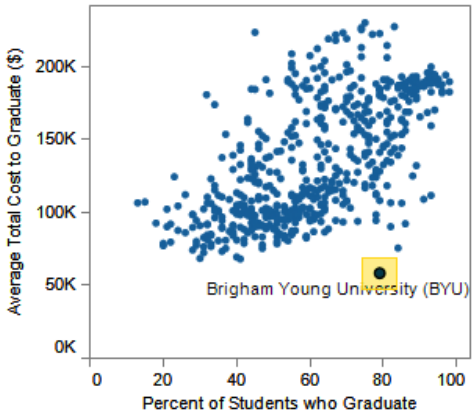

Do these two responses give the same general explanation for the peaks and valleys in the chart?

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

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

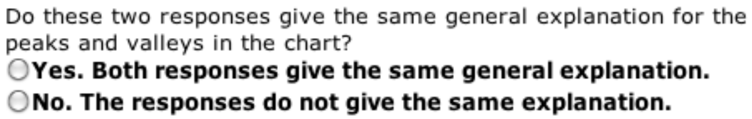


CLUSTERING VIA DISTRIBUTED COMPARISON



"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.**

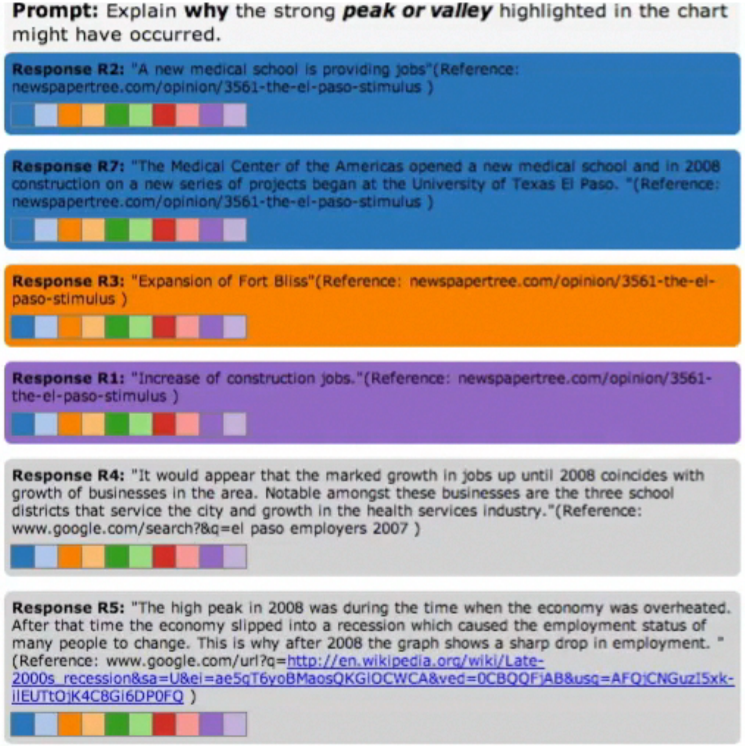
+ Simple tasks for workers

− Scales p rly O)

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=ae5qT6yoBMaosQKGIOCWCA&ved=0CBQQFjAB&usg=AFQjCNGuzI5xk-iIEUTtOjK4C8Gi6DP0FQ)

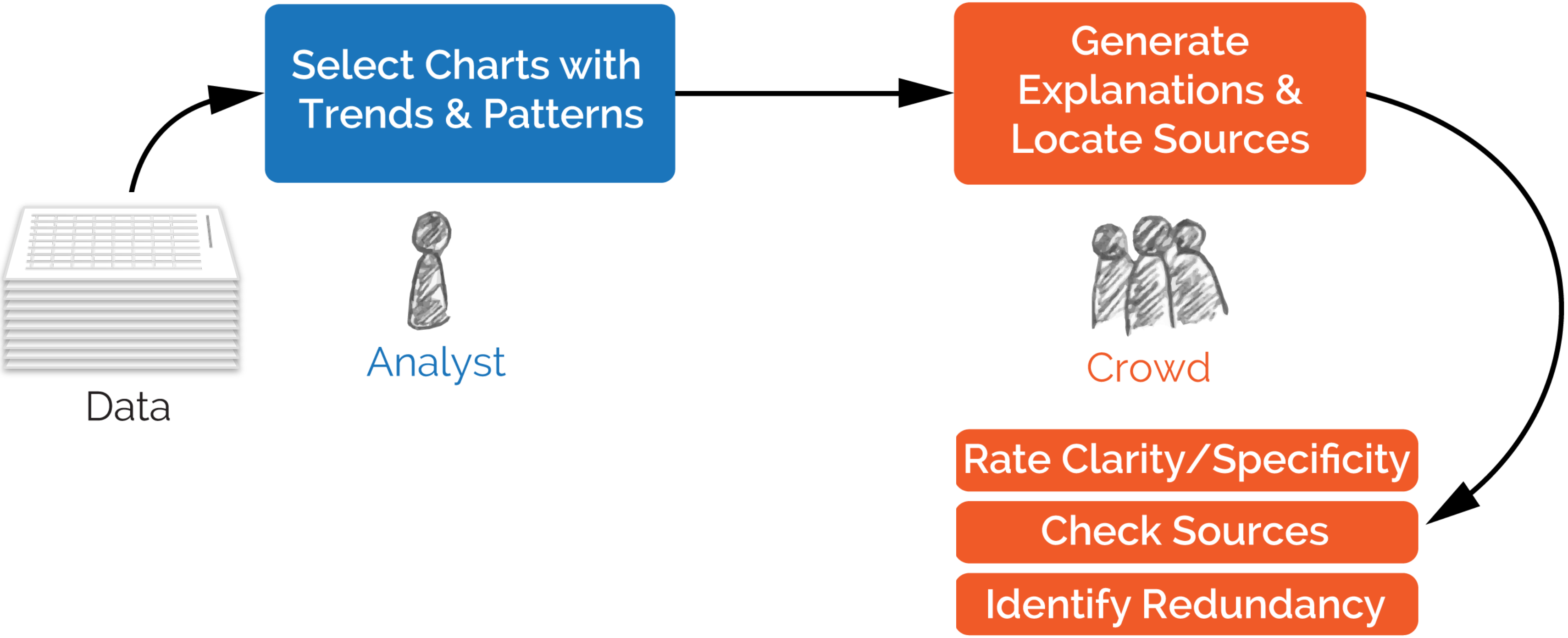
MULTIPLE WORKERS **INDEPENDENTLY CLUSTER** THE WHOLE SET.

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

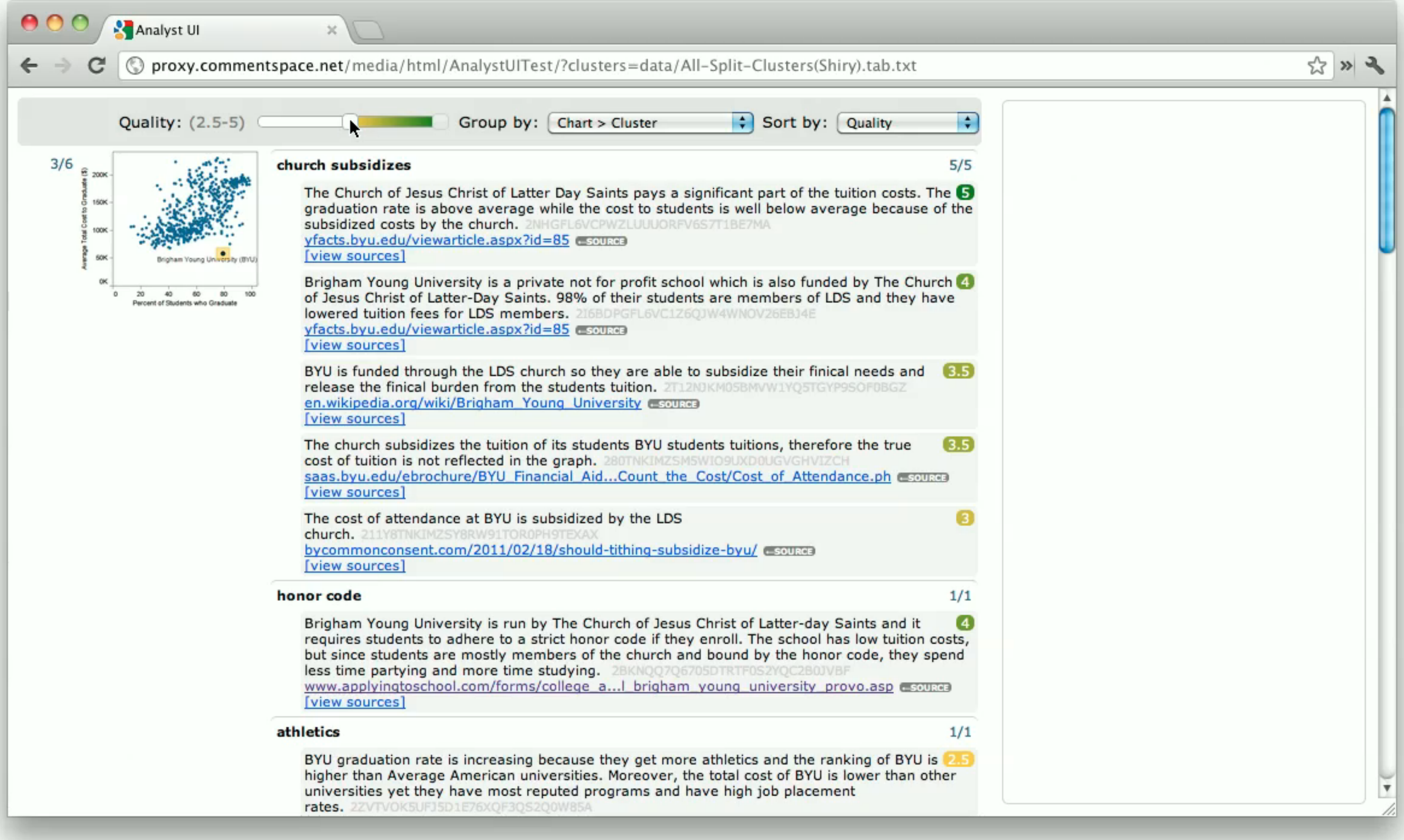
FINDING THE RIGHT BALANCE OF HUMAN AND AUTOMATED EFFORT

MANAGING THE CROWD'S WORK

MANAGING THE CROWD'S WORK


Select Charts with Trends & Patterns
Generate Explanations & Locate Sources
Data
Analyst
Crowd
Rate Clarity/Specificity
Check Sources
Identify Redundancy

EXPLANATION MANAGEMENT INTERFACE


Analyst UI
Average Total Cost to Graduate ($)
200K
150K
100K
50K
0K
0
20
40
60
80
100
Percent of Students who Graduate
Brigham Young University (BYU)

proxy.commentspace.net/media/html/AnalystUITest/?clusters=data/All-Split-Clusters(Shiry).tab.txt

Quality: (2.5-5) Group by: Chart > Cluster Sort by: Quality

3/6

church subsidizes 5/5

The Church of Jesus Christ of Latter Day Saints pays a significant part of the tuition costs. The graduation rate is above average while the cost to students is well below average because of the subsidized costs by the church. 2NHGFL6VCPWZLUUUORFV6S7T1BE7MA 5
yfacts.byu.edu/viewarticle.aspx?id=85 SOURCE
[view sources]

Brigham Young University is a private not for profit school which is also funded by The Church of Jesus Christ of Latter-Day Saints. 98% of their students are members of LDS and they have lowered tuition fees for LDS members. 2I6BDPGFL6VC1Z6QJW4WNOV26EBJ4E 4
yfacts.byu.edu/viewarticle.aspx?id=85 SOURCE
[view sources]

BYU is funded through the LDS church so they are able to subsidize their finical needs and release the finical burden from the students tuition. 2T12NJKM05BMVW1YQ5TGYP9SOF0BGZ 3.5
en.wikipedia.org/wiki/Brigham_Young_University SOURCE
[view sources]

The church subsidizes the tuition of its students BYU students tuitions, therefore the true cost of tuition is not reflected in the graph. 280TNKIMZSM5WIO9UXD0UGVGHVIZCH 3.5
saas.byu.edu/ebrochure/BYU_Financial_Aid...Count_the_Cost/Cost_of_Attendance.ph SOURCE
[view sources]

The cost of attendance at BYU is subsidized by the LDS church. 211Y8TNKIMZSY8RW91TOR0PH9TEXAX 3
bycommonconsent.com/2011/02/18/should-tithing-subsidize-byu/ SOURCE
[view sources]

honor code 1/1

Brigham Young University is run by The Church of Jesus Christ of Latter-day Saints and it requires students to adhere to a strict honor code if they enroll. The school has low tuition costs, but since students are mostly members of the church and bound by the honor code, they spend less time partying and more time studying. 2BKNQQ7Q6705DTRTF0S2YQC2B0JVBF 4
www.applyingtoschool.com/forms/college_a...l_brigham_young_university_provo.asp SOURCE
[view sources]

athletics 1/1

BYU graduation rate is increasing because they get more athletics and the ranking of BYU is higher than Average American universities. Moreover, the total cost of BYU is lower than other universities yet they have most reputed programs and have high job placement rates. 2ZVTVOK5UFJ5D1E76XQF3QS2Q0W85A 2.5

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
APPLICATION AREAS (PETRA)

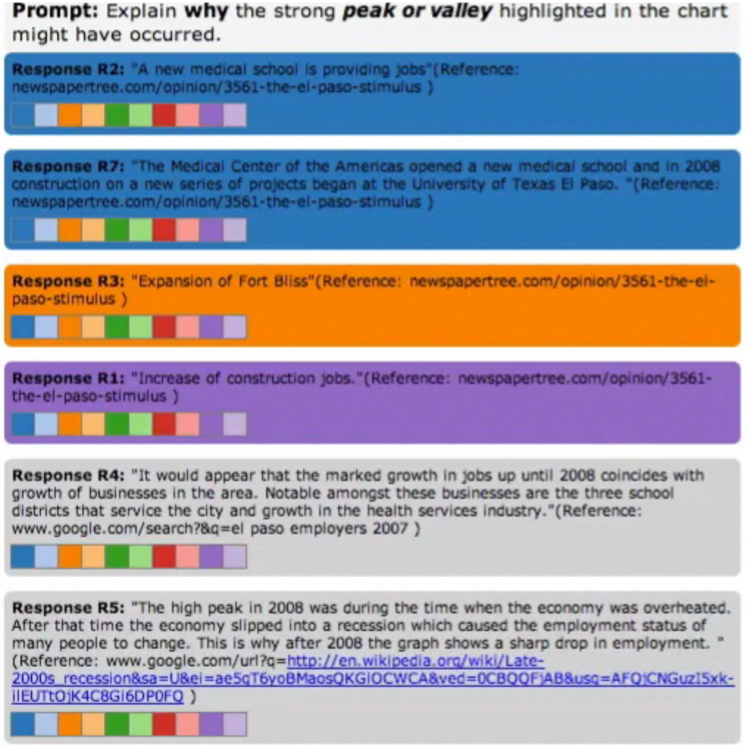
THIS AFTERNOON
FINAL KRONOS ASSIGNMENT
(OPEN LAB)

DECEMBER 8th-19th
INFORMATION VISUALIZATION LECTURES
AT UNIVERSITÉ PARIS SUD

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=ae5qT6yoBMaosQKGIOCWCA&ved=0CBQQFjAB&usg=AFQjCNGuzI5xk-ilEUTtOjK4C8Gi6DP0FQ)

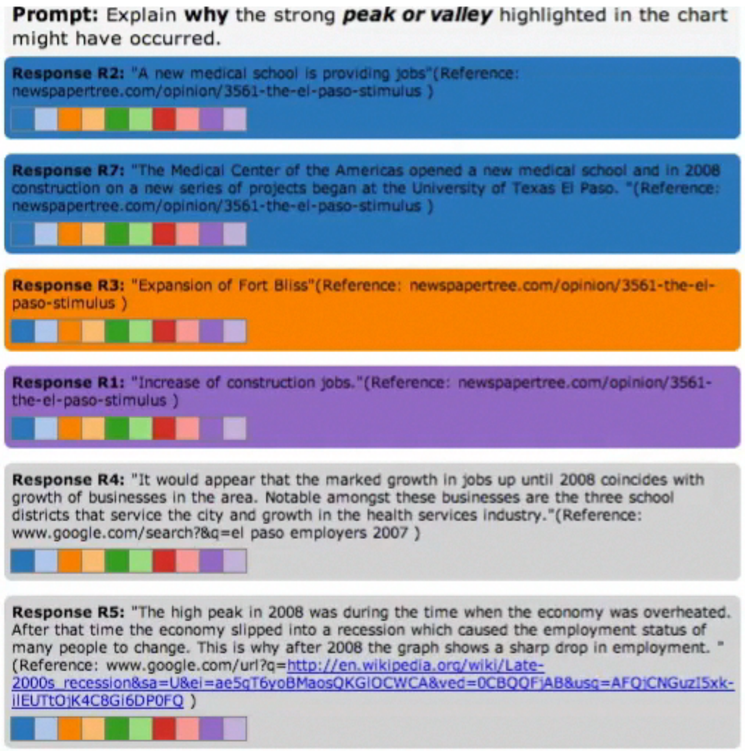
Individual workers cluster the whole set.

+ 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)

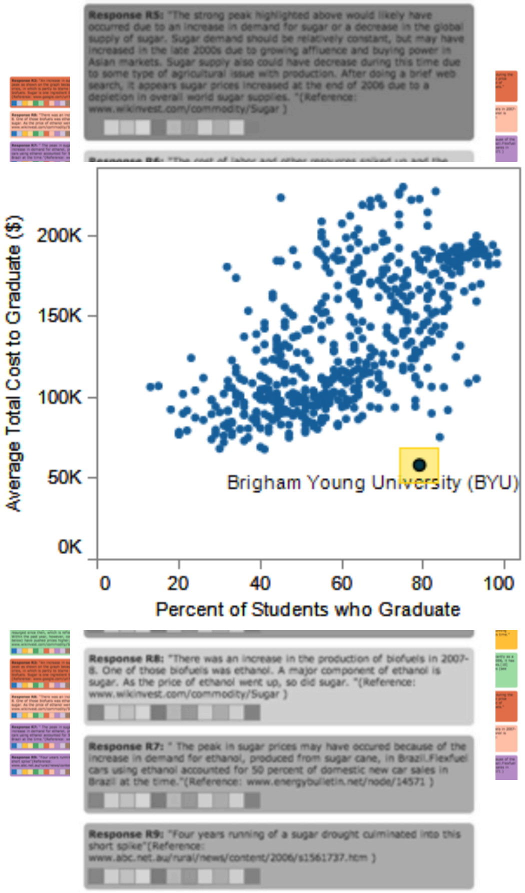
Response R1: "Increase of construction jobs."(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus)

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=ae5qT6yoBMaosQKGIOCWCA&ved=0CBQQFjAB&usg=AFQjCNGuzI5xk-iIEUTtOjK4C8Gi6DP0FQ)

- 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?



Response R5: "The strong peak highlighted above would likely have occurred due to an increase in demand for sugar or a decrease in the global supply of sugar. Sugar demand should be relatively constant, but may have increased in the late 2000s due to growing affluence and buying power in Asian markets. Sugar supply also could have decrease during this time due to some type of agricultural issue with production. After doing a brief web search, it appears sugar prices increased at the end of 2006 due to a depletion in overall world sugar supplies. "(Reference: www.wikinvest.com/commodity/Sugar)

Response R6: "The cost of labor and other resources spiked up and the

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.wikinvest.com/commodity/Sugar)

Response R7: " The peak in sugar prices may have occured because of the increase in demand for ethanol, produced from sugar cane, in Brazil.Flexfuel cars using ethanol accounted for 50 percent of domestic new car sales in Brazil at the time."(Reference: www.energybulletin.net/node/14571)

Response R9: "Four years running of a sugar drought culminated into this short spike"(Reference: www.abc.net.au/rural/news/content/2006/s1561737.htm)

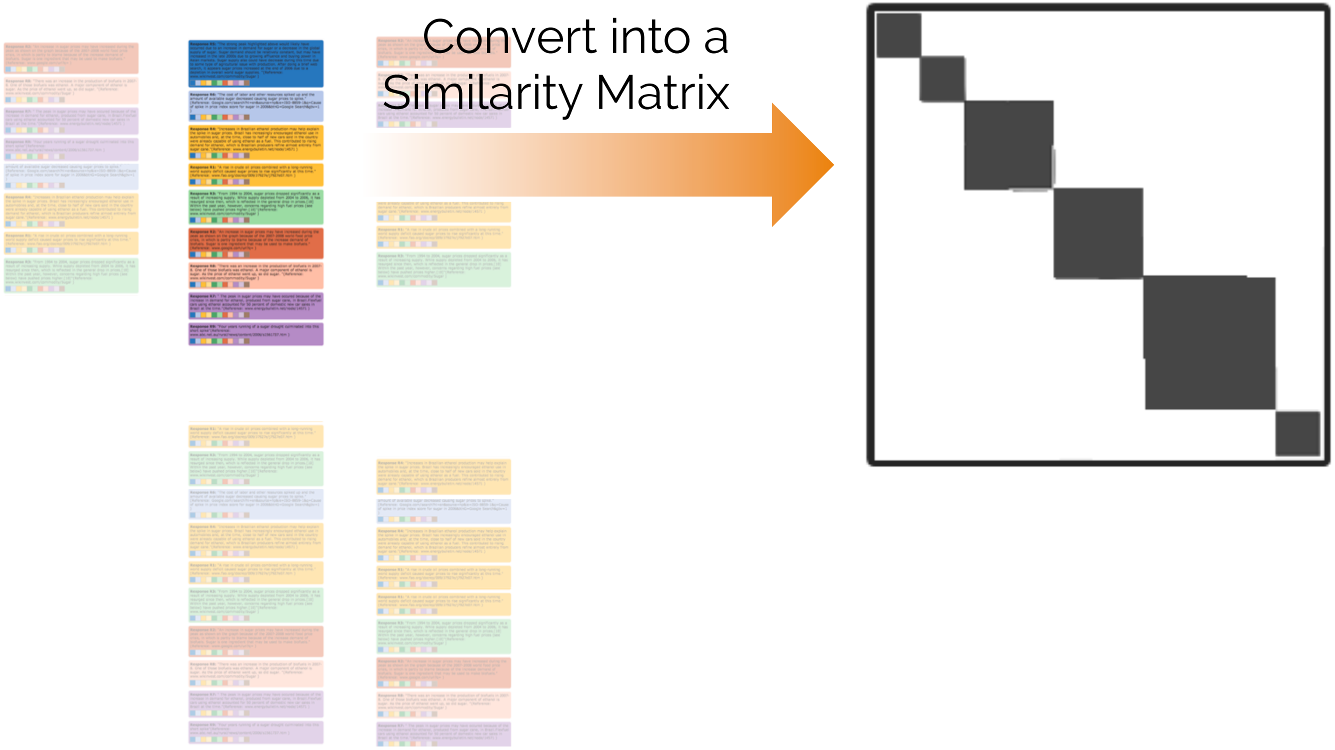
Average Total Cost to Graduate ($)
200K
150K
100K
50K
0K
0
20
40
60
80
100
Percent of Students who Graduate
Brigham Young University (BYU)

Selecting the

Most-Representative Clustering

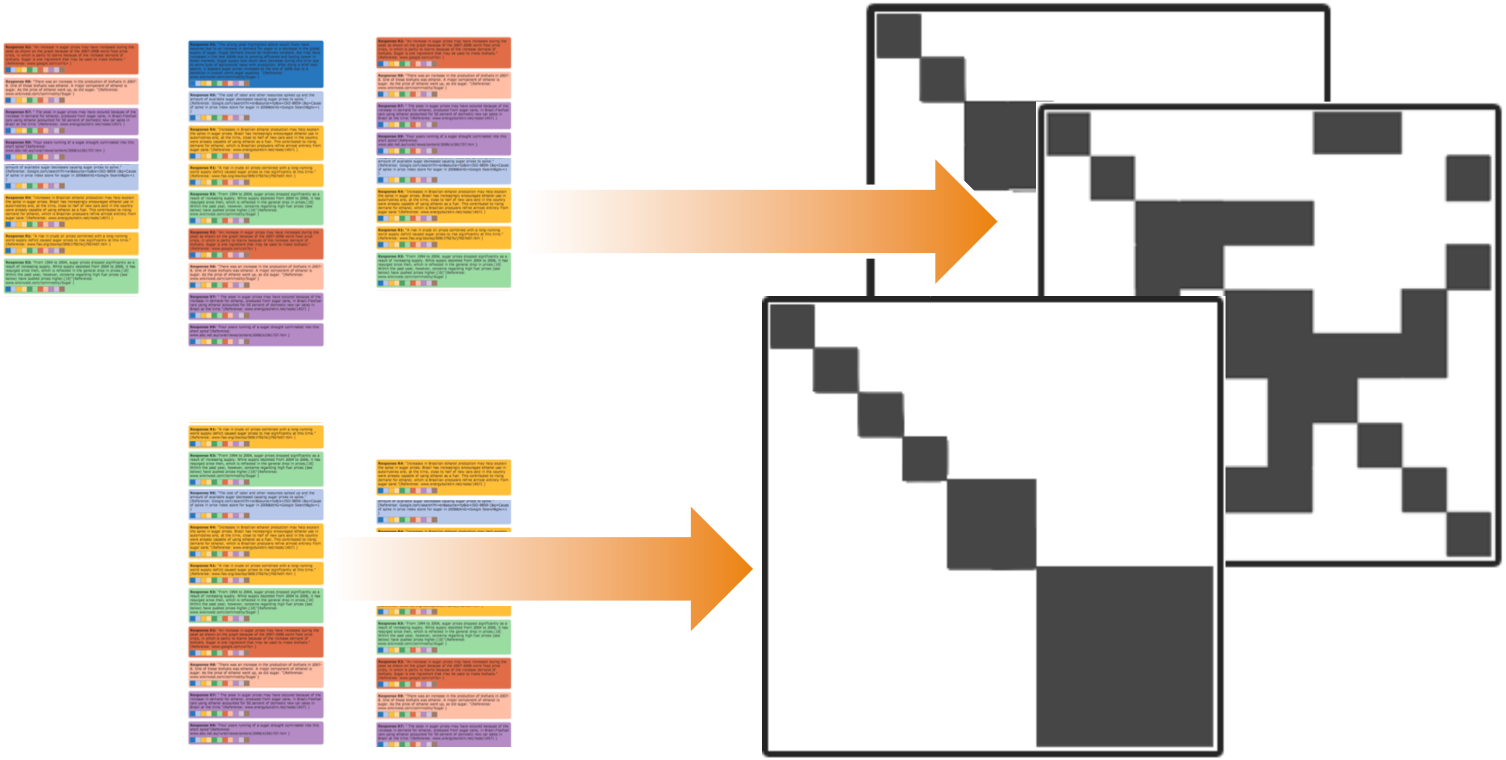
HOW TO INTEGRATE COLOR-CLUSTERINGS?

SELECTING THE MOST-REPRESENTATIVE CLUSTERING

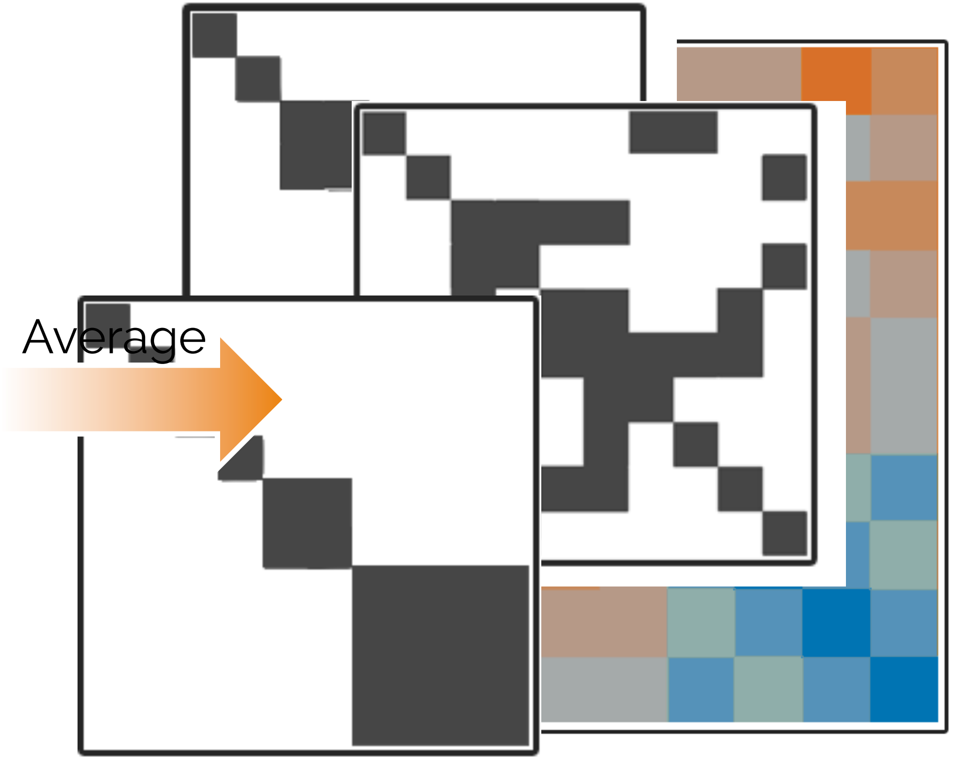


Convert into a Similarity Matrix

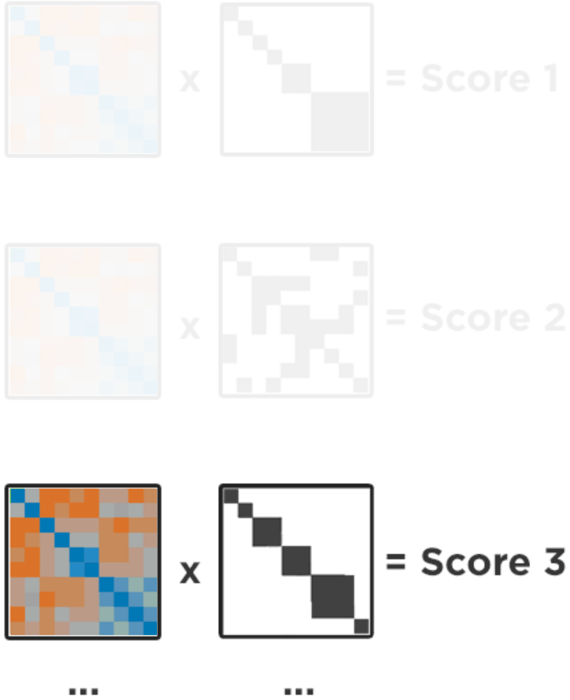
SELECTING THE MOST-REPRESENTATIVE CLUSTERING



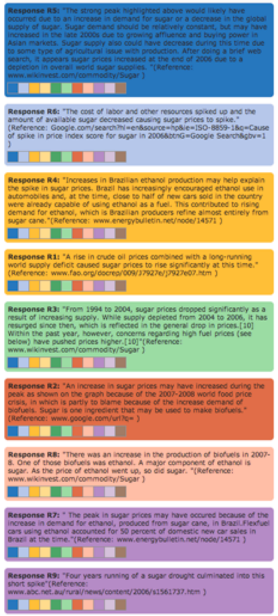
SELECTING THE MOST-REPRESENTATIVE CLUSTERING



SELECTING THE MOST-REPRESENTATIVE CLUSTERING



Select
Highest
Scoring



Response R5: "The strong peak highlighted above would likely have occurred due to an increase in demand for sugar or a decrease in the global supply of sugar. Sugar demand should be relatively constant, but may have increased in the late 2000s due to growing affluence and buying power in Asian markets. Sugar supply also could have decrease during this time due to some type of agricultural issue with production. After doing a brief web search, it appears sugar prices increased at the end of 2006 due to a depletion in overall world sugar supplies. "(Reference: www.wikinvest.com/commodity/Sugar)

Response R6: "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?hl=en&source=hp&ie=ISO-8859-1&q=Cause of spike in price index score for sugar in 2006&btnG=Google Search&gbv=1)

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.energybulletin.net/node/14571)

Response R1: "A rise in crude oil prices combined with a long-running world supply deficit caused sugar prices to rise significantly at this time." (Reference: www.fao.org/docrep/009/J7927e/j7927e07.htm)

Response R3: "From 1994 to 2004, sugar prices dropped significantly as a result of increasing supply. While supply depleted from 2004 to 2006, it has resurged since then, which is reflected in the general drop in prices.[10] Within the past year, however, concerns regarding high fuel prices (see below) have pushed prices higher.[10]"(Reference: www.wikinvest.com/commodity/Sugar)

Response R2: "An increase in sugar prices may have increased during the peak as shown on the graph because of the 2007-2008 world food price crisis, in which is partly to blame because of the increase demand of biofuels. Sugar is one ingredient that may be used to make biofuels." (Reference: 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.wikinvest.com/commodity/Sugar)

Response R7: " The peak in sugar prices may have occured because of the increase in demand for ethanol, produced from sugar cane, in Brazil.Flexfuel cars using ethanol accounted for 50 percent of domestic new car sales in Brazil at the time."(Reference: www.energybulletin.net/node/14571)

Response R9: "Four years running of a sugar drought culminated into this short spike"(Reference: www.abc.net.au/rural/news/content/2006/s1561737.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 ⟷ 1] (identical)

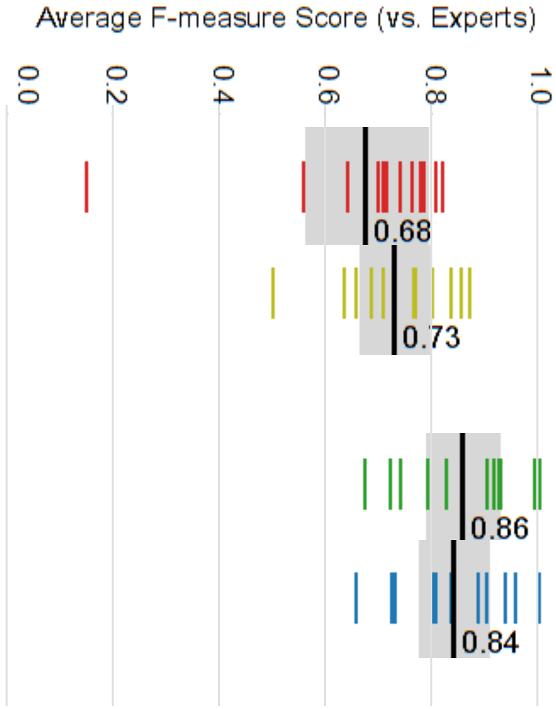
EVALUATING REDUNDANCY-DETECTION

Unclustered Results F=0.68

Color Clustering F=0.73

Most-Representative Selection F=0.86

Experts vs. One Another F=0.84



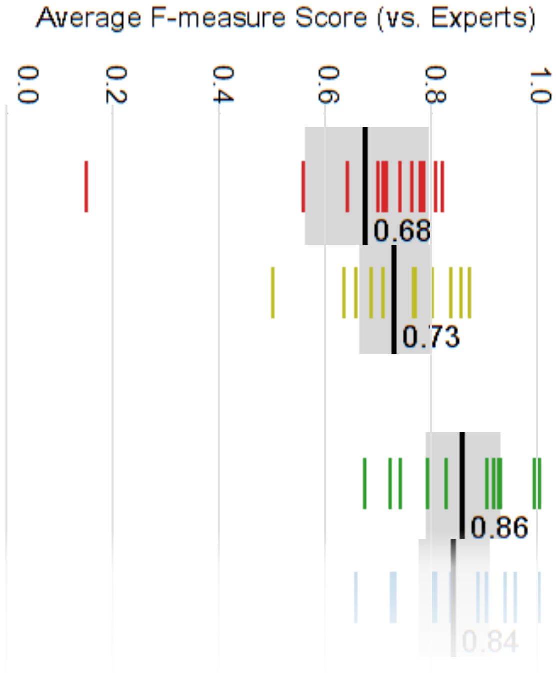
EVALUATING REDUNDANCY-DETECTION

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)