HUMAN-POWERED DATA MANAGEMENT

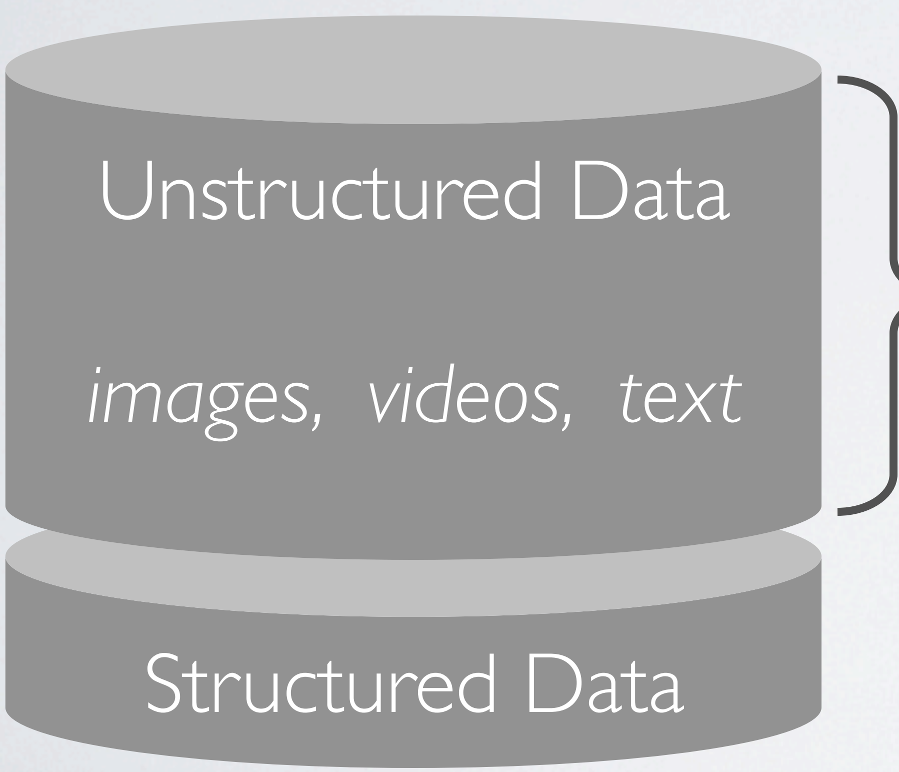
Aditya Parameswaran

with H. Garcia-Molina,
J. Widom, A. Polyzotis, M. Teh

Massachusetts Institute of Technology

Why should we (DM/DB folks) care?

Reason 1: Most data is unstructured



Automated processing: not yet solved

Incorporate 

Why should we (DM/DB folks) care?

Reason 2: S/ware companies use crowds at scale

We undertook a survey of industry crowdsourcing users

Microsoft

use crowds!

Often 10s+ of Millions of $ / yr. / company
(on crowds + supervisors)

Plenty of startups too!

Why should we (DM/DB folks) care?

Reason 3: Marketplaces are growing rapidly

Crowdsourcing Marketplaces

20+ marketplaces

Big companies have internal ones

Size of these marketplaces have doubled in 2011 – 2013

Why should we (DM/DB folks) care?

Reason 1: Most data is unstructured

Reason 2: Software companies use crowds at scale

Reason 3: Marketplaces are growing rapidly

What is Human-Powered Data Management?

Data Processing Algorithms

Data Processing Systems

where humans act as "data processors"
e.g., *compare*, *label*, *extract*

Learning accuracies — Machine Learning

Interfaces Patterns — HCI

Incentives — Economics

Efficient Data Processing Algorithms & Systems

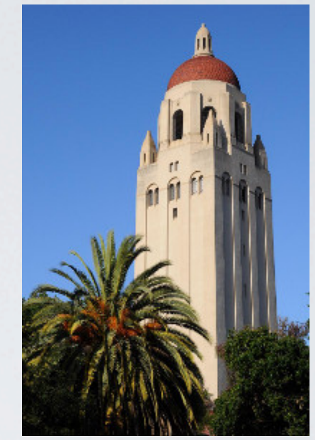
Data Processing Algorithms	Filter [SIGMOD12, VLDB14] Max [SIGMOD12] Clean [KDD12, TKDD13] Categorize [VLDB11] Search [ICDE14] Debugging [NIPS12]
Data Processing Systems	Deco [CIKM12, VLDB12, TR12, SIGMOD Record 12] DataSift [HCOMP13, SIGMOD14] HQuery [CIDR11]
Auxiliary Plugins: Quality, Pricing	Confidence [KDD13, TR14] Eviction [TR12] Pricing [VLDB15] Quality [HCOMP14]

i.stanford.edu/~adityagp/scoop.html

Data Proc. Sys.: Crowd-Powered Search

Can your search engine handle this?

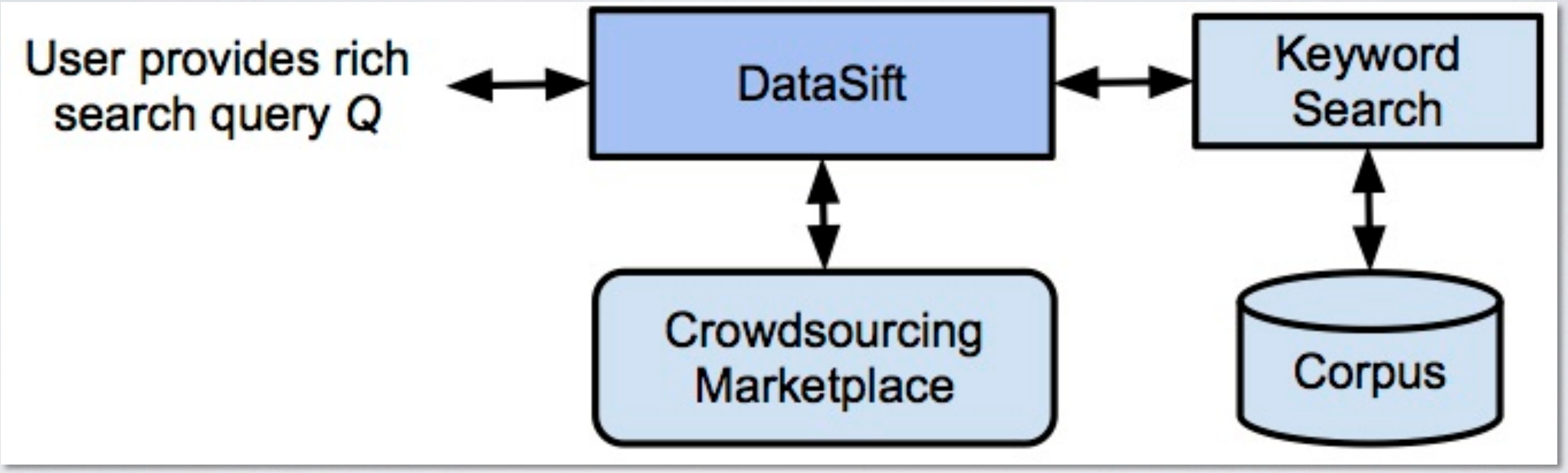
buildings in the vicinity of



type of cable that connects to

apartments in a good school district near Urbana, with a bus stop near by

DataSift: Crowd-Powered Search



- *Non-textual content:*
  "cables that plug into <img>"
  "funny pictures of cats with hats with captions"

- *Time-consuming:*
  "find noise canceling headphones where the battery lasts 13 hrs"
  "apartments in a nice area around urbana"

searched for **type of cable that connects to**  using Amazon Products

DataSift Rank	Thumbnail	Product Details
1		**Mediabridge Hi-Speed USB 2.0 Cable - (6 Feet)** Product page: http://www.amazon.co/dp/B001MXLD4G Price: USD 4.99
2		**AmazonBasics USB 2.0 A-Male to B-Male Cable with Lighted Ends - Braided (6 Feet/1.8 Meters)** Product page: http://www.amazon.co/dp/B003ES5ZQE Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 6.99
3		**Epson Stylus USB Printer Cord NEW !! 2.0 A - B Cable 6'** Product page: http://www.amazon.co/dp/B0032GO0SW Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 2.88
4		**USB Printer Cable for HP DeskJet 1000 with Life Time Warranty** Product page: http://www.amazon.co/dp/B004PRXM2C Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 4.95
5		**Mediabridge Hi-Speed USB 2.0 Cable - (10 Feet)** Product page: http://www.amazon.co/dp/B001MSU1HG Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 5.49
6		**Mediabridge Hi-Speed USB 2.0 Cable - (16 Feet)** Product page: http://www.amazon.co/dp/B001MSZBNA Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 7.49

Building DataSift: Challenges

Gather — Ask for text reformulations for query

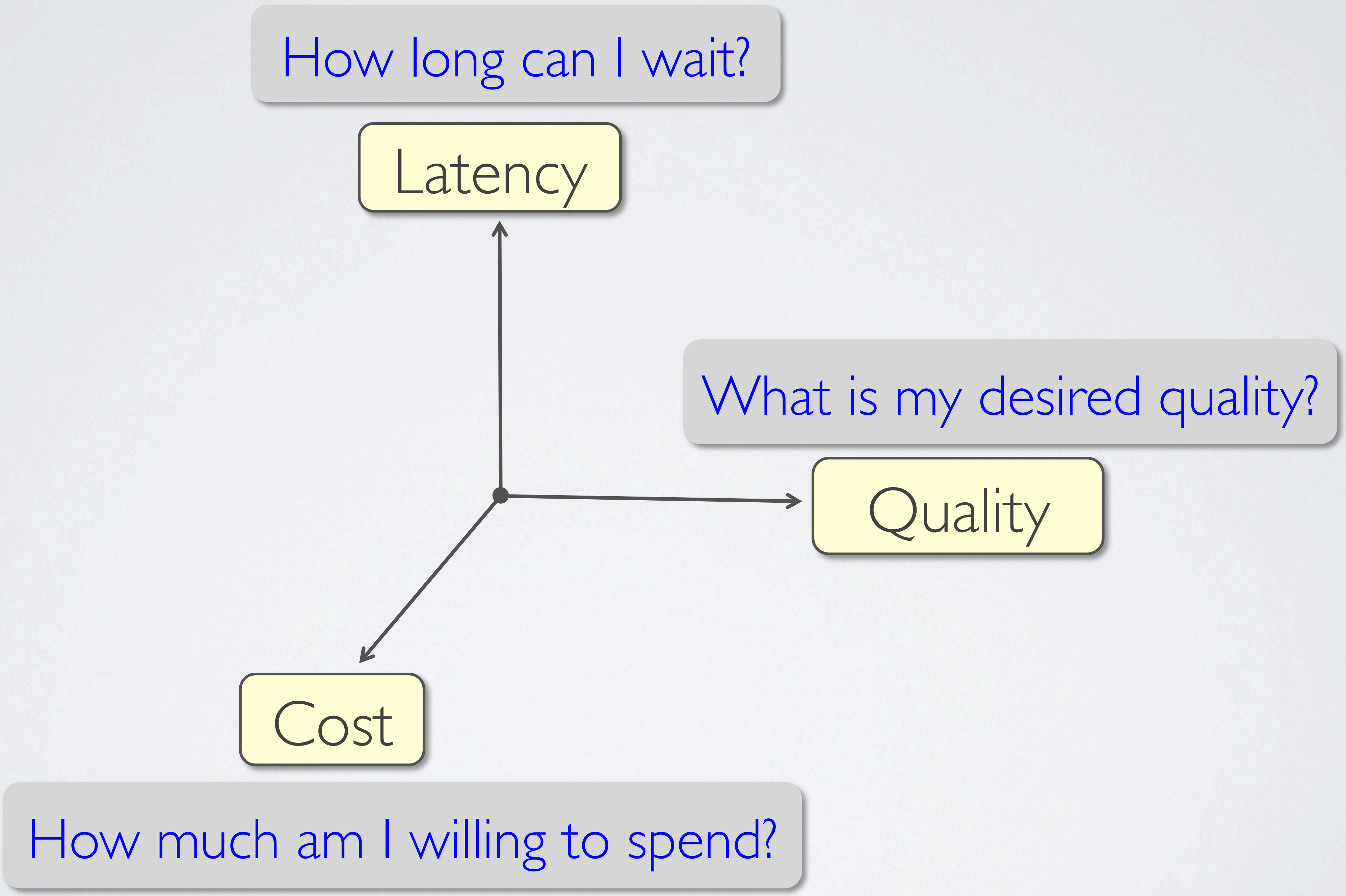
Filter — Check if item satisfies query

Gather → Retrieve → Filter

Gather → Retrieve → Filter → Retrieve → Filter

- *How many reformulations should we gather?*
- *How many items should we retrieve at each step?*
- *How do we filter items? How many people do we ask?*
- *How do we optimize the workflow?*
- *How do we guarantee correctness?*

Fundamental Tradeoffs

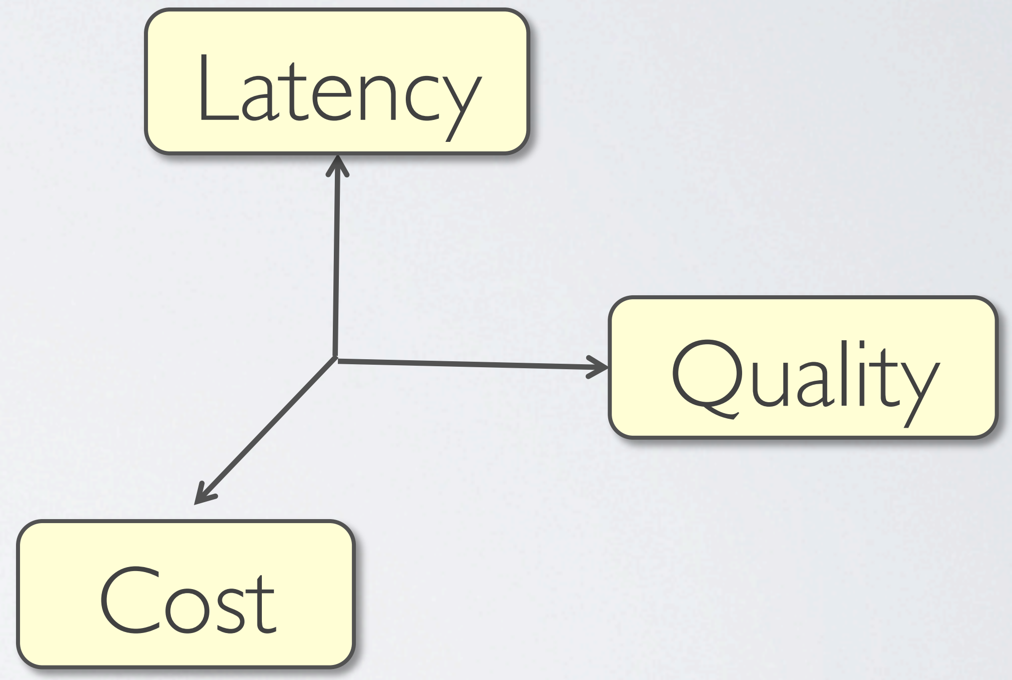


DataSift Summary

Sample applications:

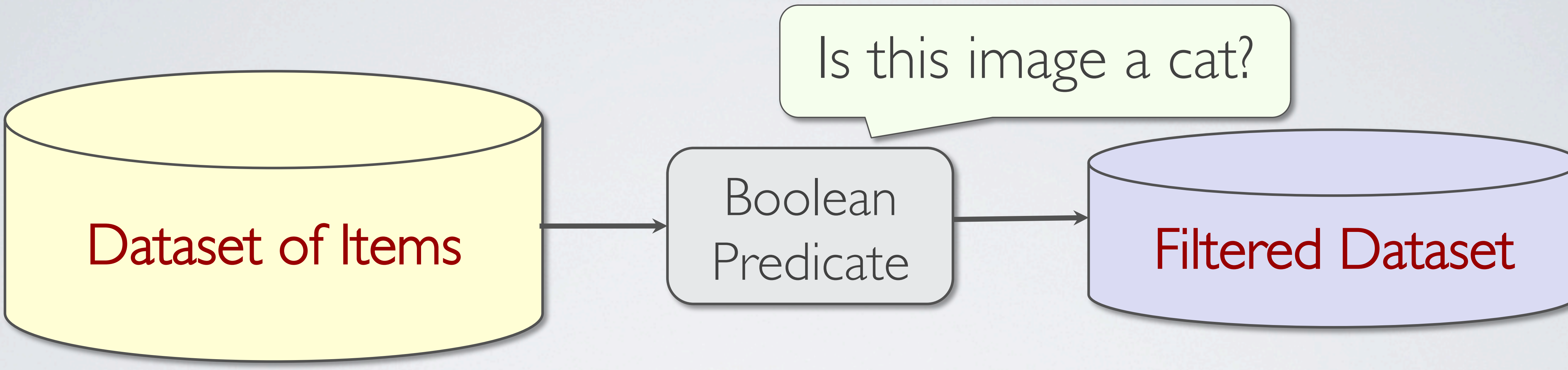
education, social media, commerce, journalism, …

Gather Retrieve Filter

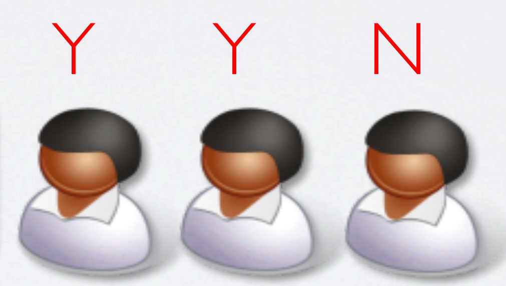


[SIGMOD14] DataSift: A Crowd-Powered Search Toolkit (demo)
[HCOMP13] An expressive and accurate crowd powered search

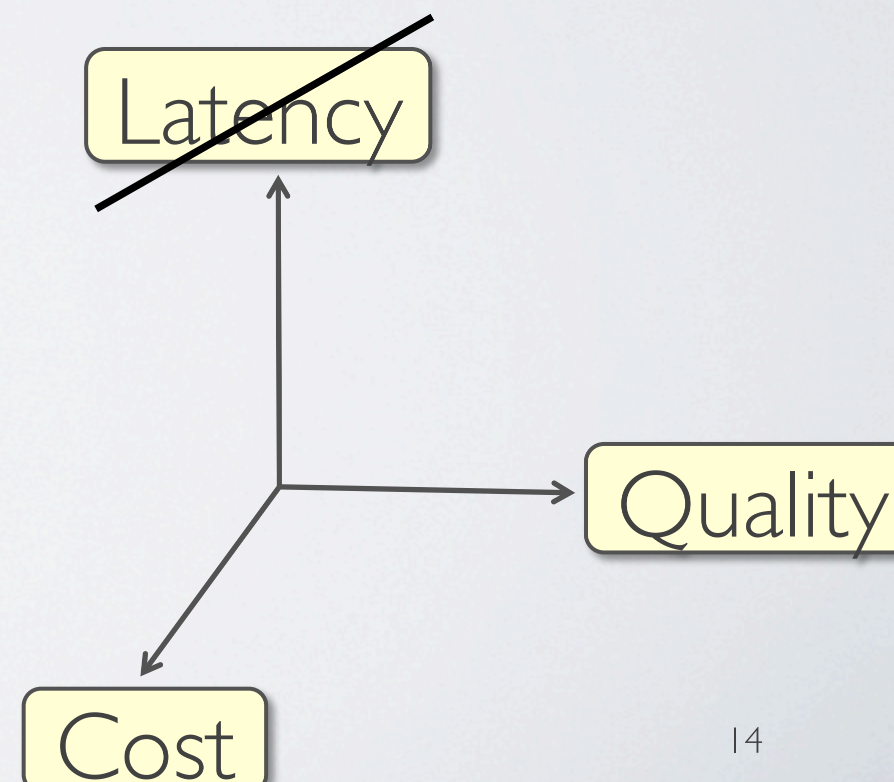
Filtering: The Simplest Version



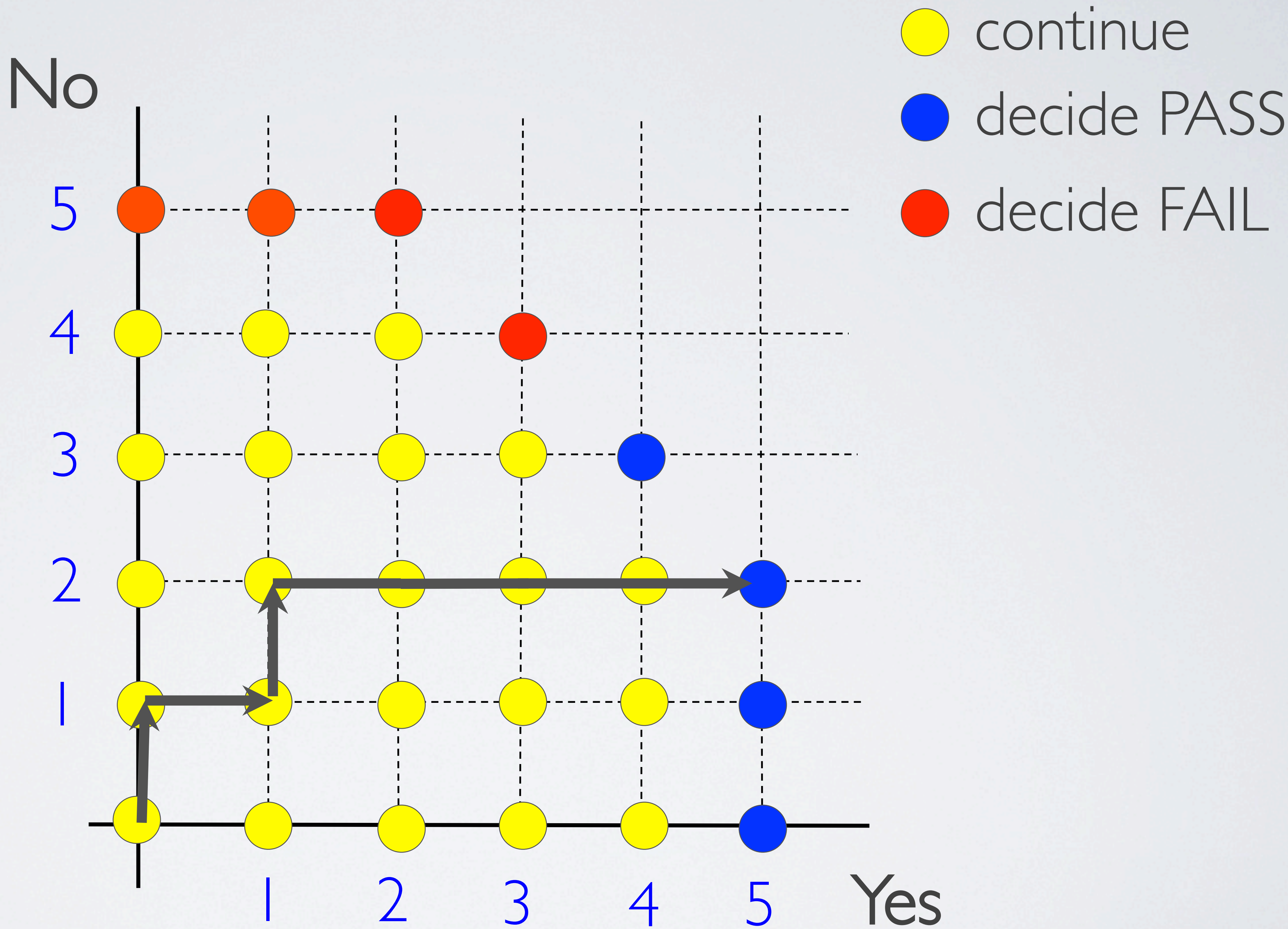
Does X satisfy predicate?



For now, all humans have same error rates

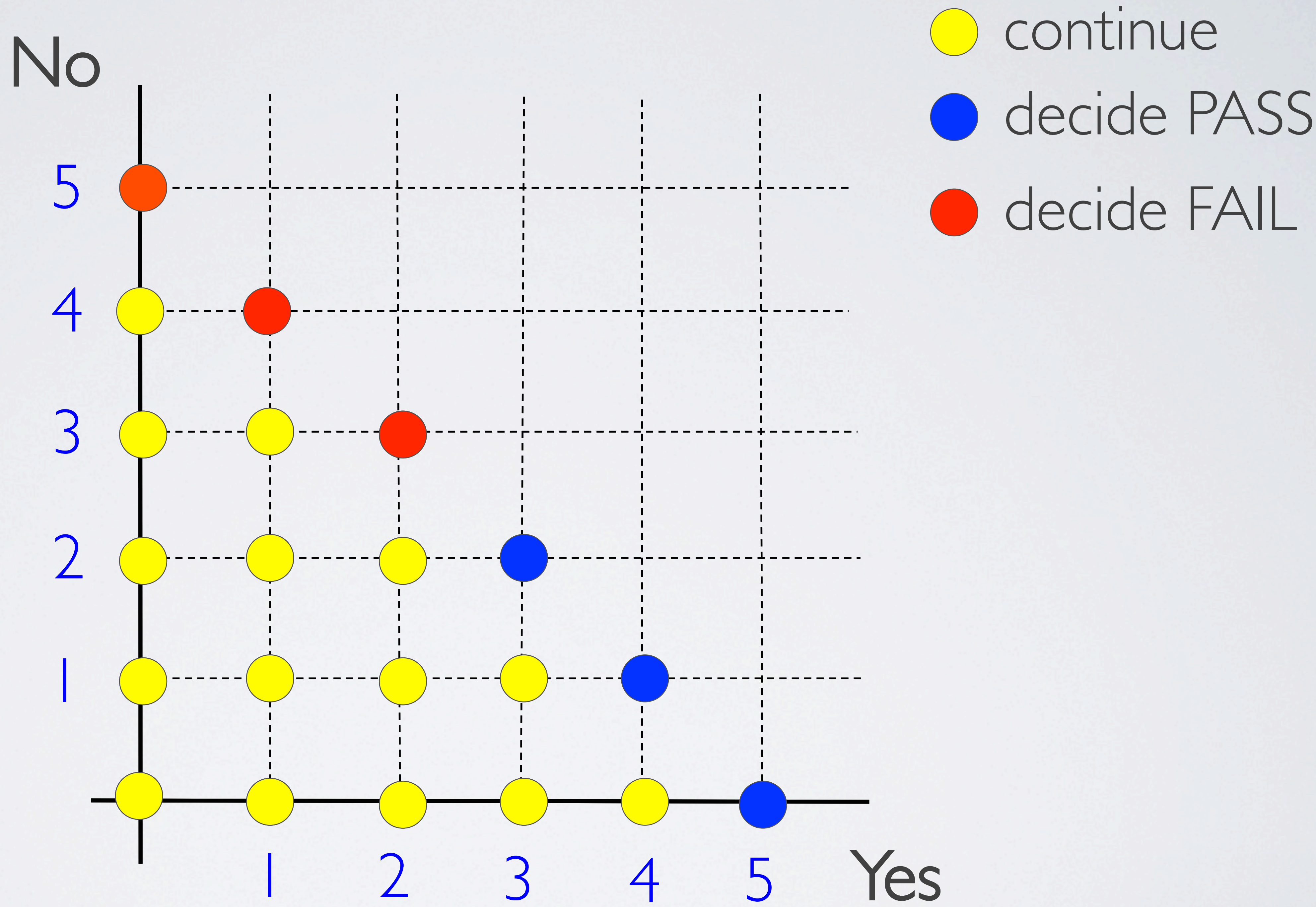


Our Visualization of Strategies



Markov Decision Process

Strategy Examples



Simplest Version

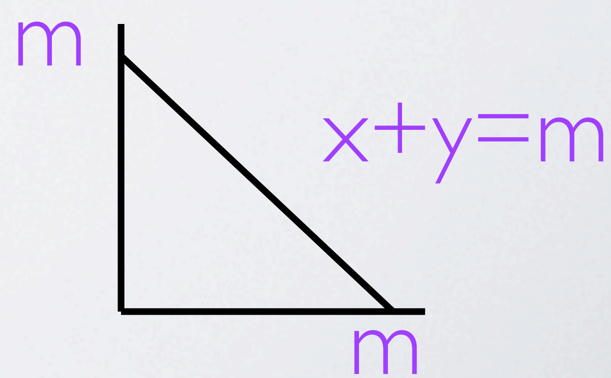
Given:

- Human error probability (FP/FN)
  - Pr [Yes | 0]; Pr [No | 1]
- A-priori probability
  - Pr [0]; Pr[1]

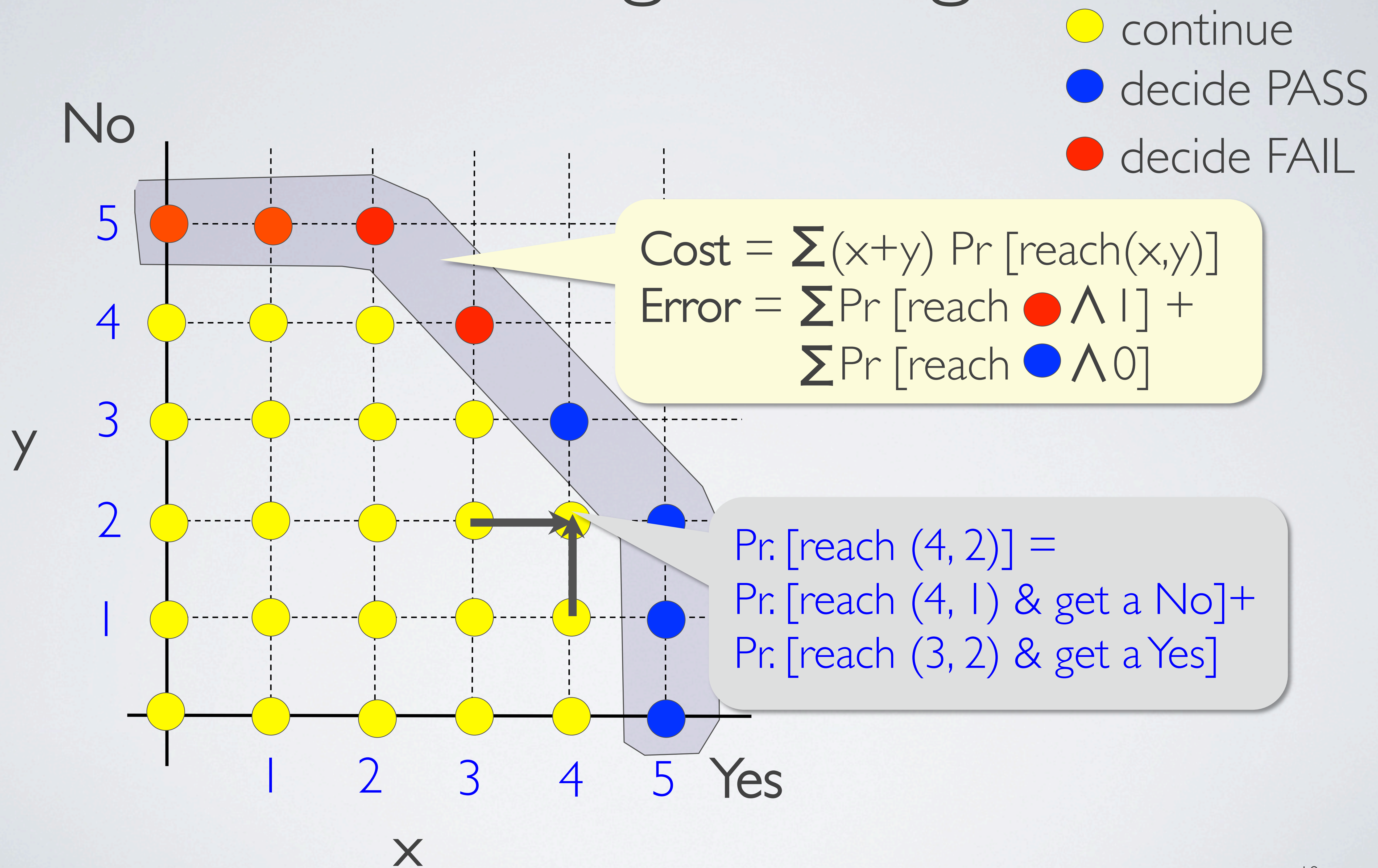
Via sampling, prior history, or gold standard

Find **strategy** with minimum expected cost (# of questions)

- Expected error < t (say, 5%)
- Cost per item < m (say, 20 questions)



Evaluating Strategies



Naïve Approach

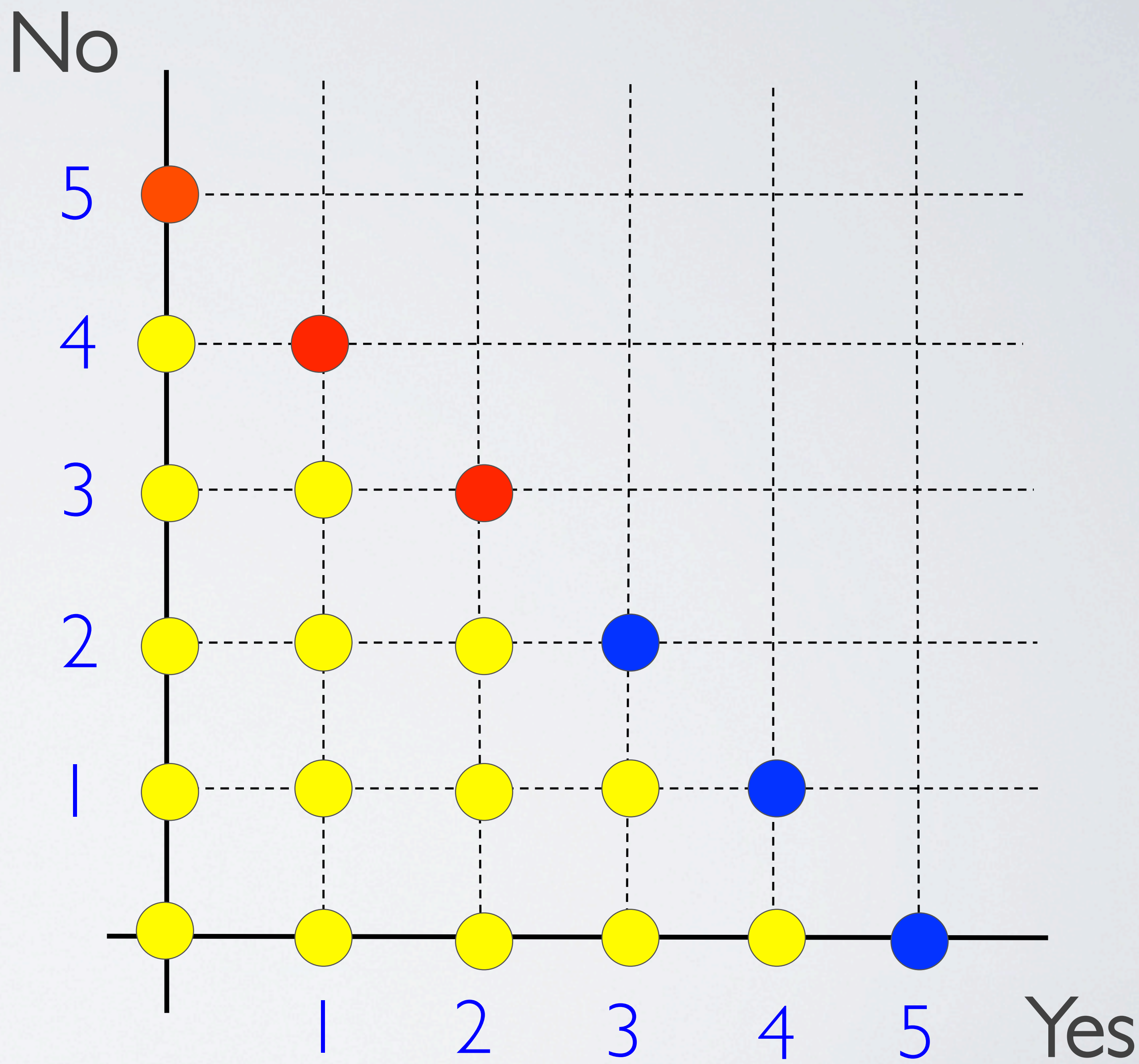
For each grid point

Assign , or

For all strategies:

- Evaluate cost & error

Return the best



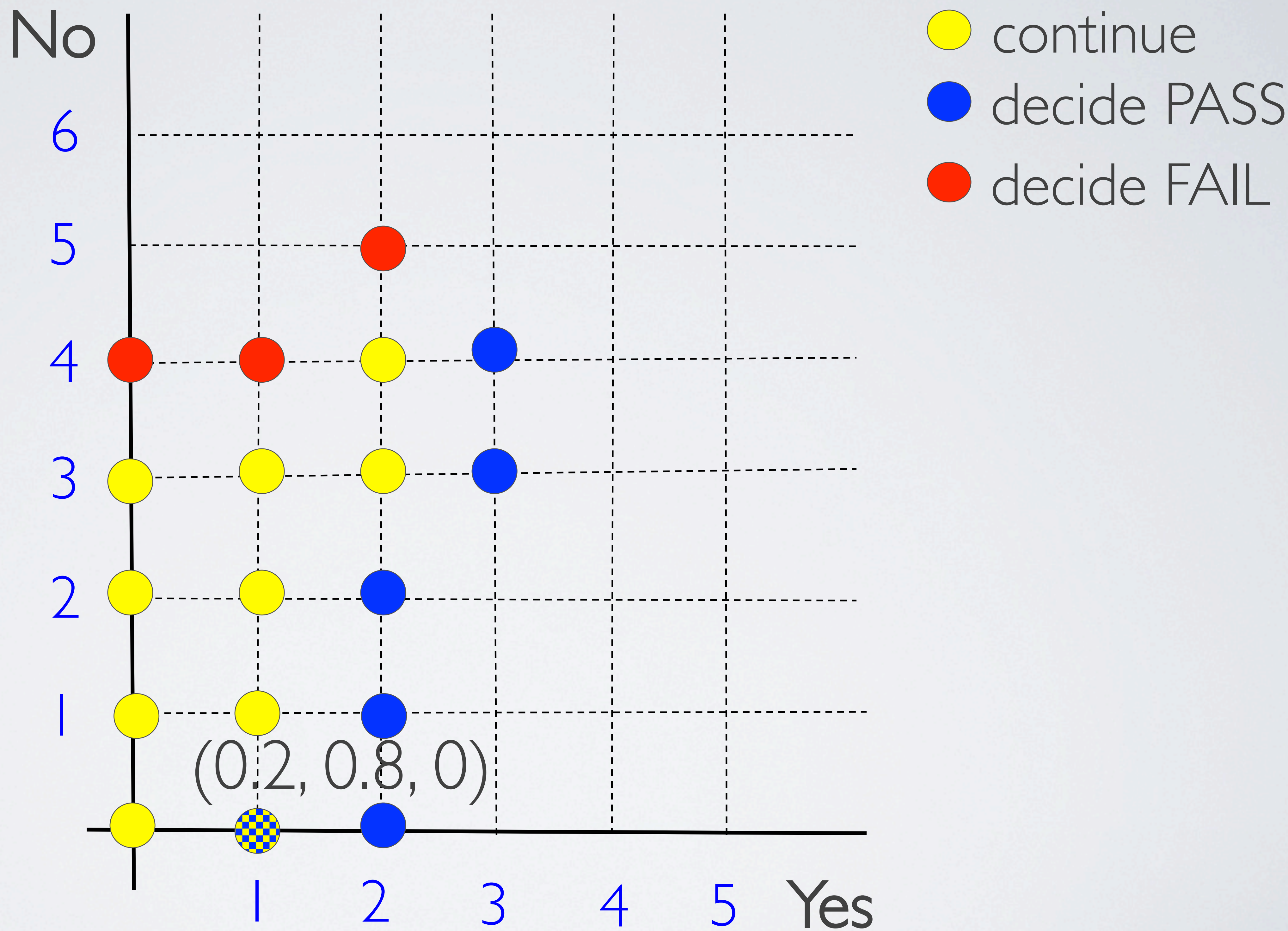
$O(3^g)$, $g = O(m^2)$

If m= 5, g = 21

Comparison

	Computing Strategy	Money
Naïve deterministic	Not feasible	$$
Our best deterministic	Exponential; feasible	$$$

Probabilistic Strategy Example



Comparison

	Computing Strategy	Money
Naïve deterministic	Exponential; not feasible	$$
Our best deterministic	Exponential; feasible	$$$
~~The best probabilistic~~ THE BEST	Polynomial(m)	$

Finding the Optimal Strategy

Simple: Use Linear Programming

- variables: “probabilistic decision per grid point”
- constraints:
  - probability conservation
  - boundary conditions

[SIGMOD12] Crowdscreen: Algorithms for filtering data with humans

Generalizations

- Multiple answers (ratings, categories)
- Multiple independent filters
- Difficulty
- Different penalty functions
- Latency
- Different worker abilities
- Different worker probes
- A-priori scores

Doable

Hard!

Generalization: Worker Abilities

	Item 1	Item 2	Item 3
Actual	0	1	0
W_1	0	1	0
W_2	1	1	1
W_3	1	0	1

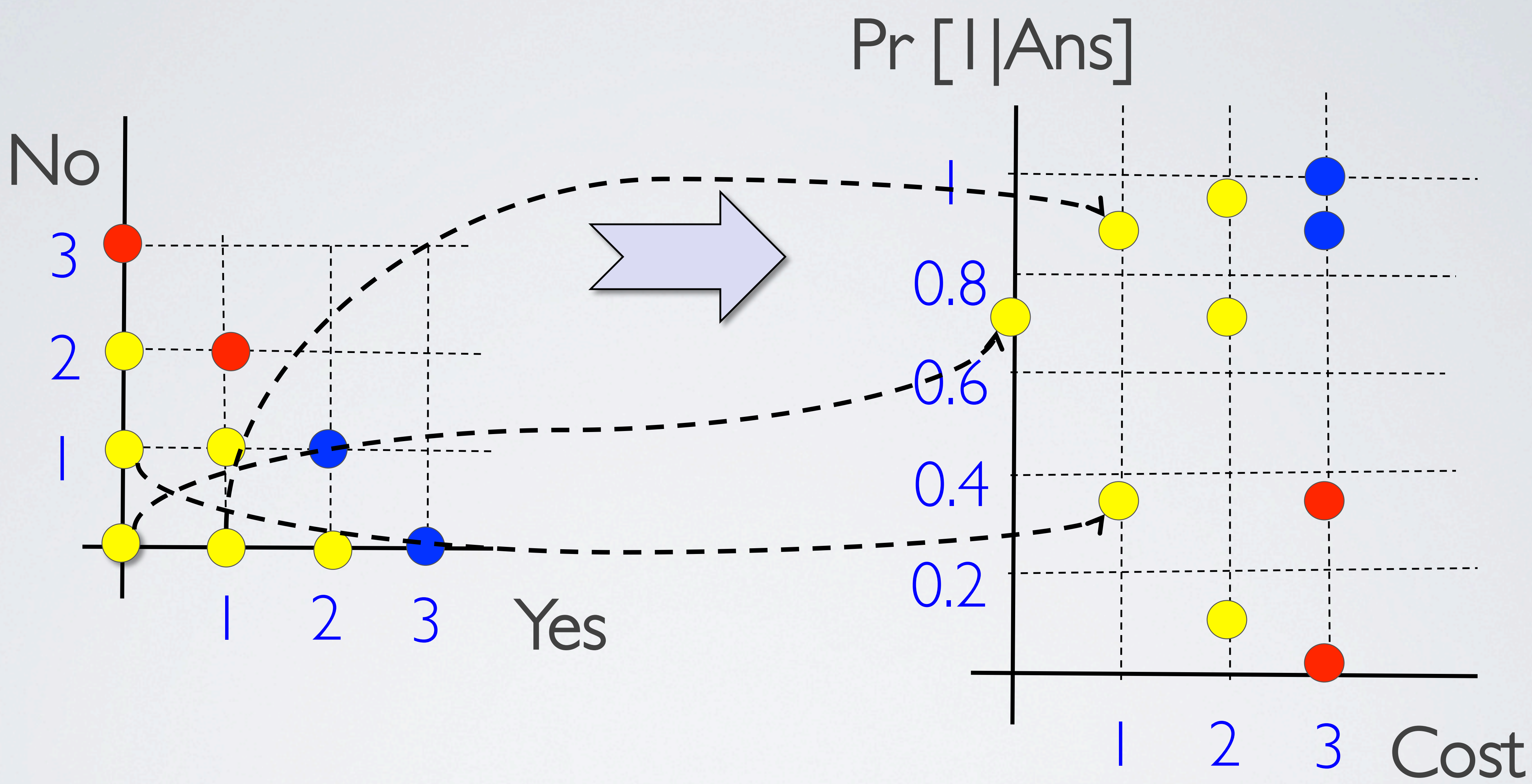
$(W_1\,\text{Yes},\ W_1\,\text{No},\ \ldots,\ W_n\,\text{Yes},\ W_n\,\text{No})$

$O(m^{2n})$ points

$n \approx 1000$

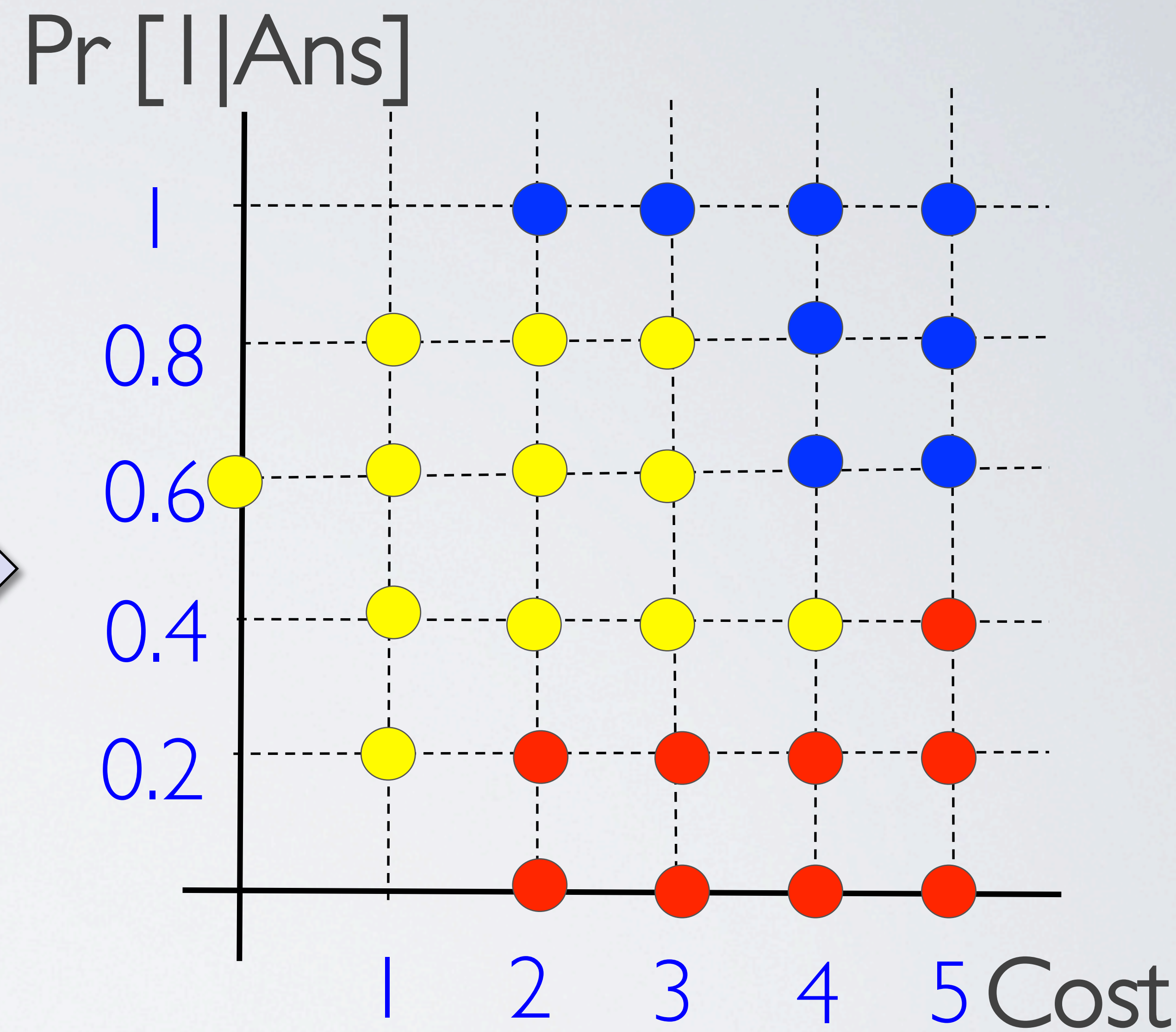
Explosion of state!

A Different Representation



Worker Abilities: Sufficiency

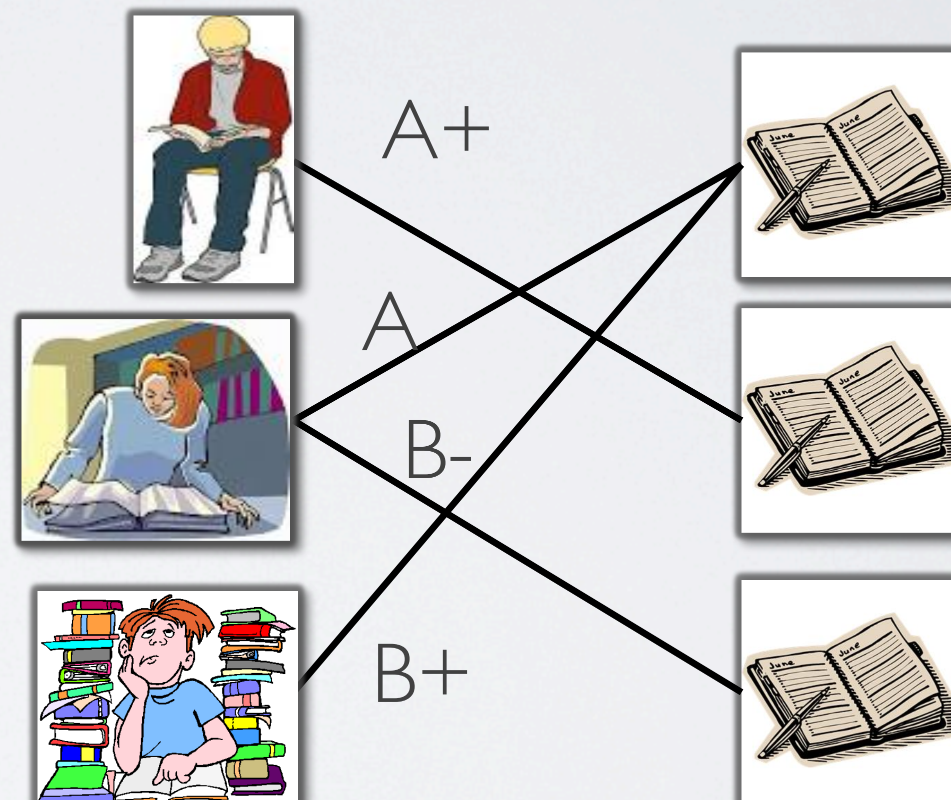
(W_1 Yes, W_1 No,
W_2 Yes, W_2 No,
…,
W_n Yes, W_n No)



Recording Pr[1|Ans] is sufficient:
Strategy → Optimal

MOOCs: Application of Filtering

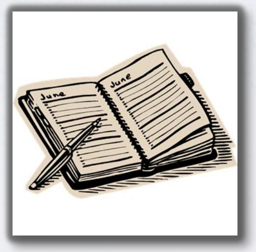
Peer Evaluation ≈ Crowdsourcing Required



Generalization of boolean filtering to scoring [1-5]

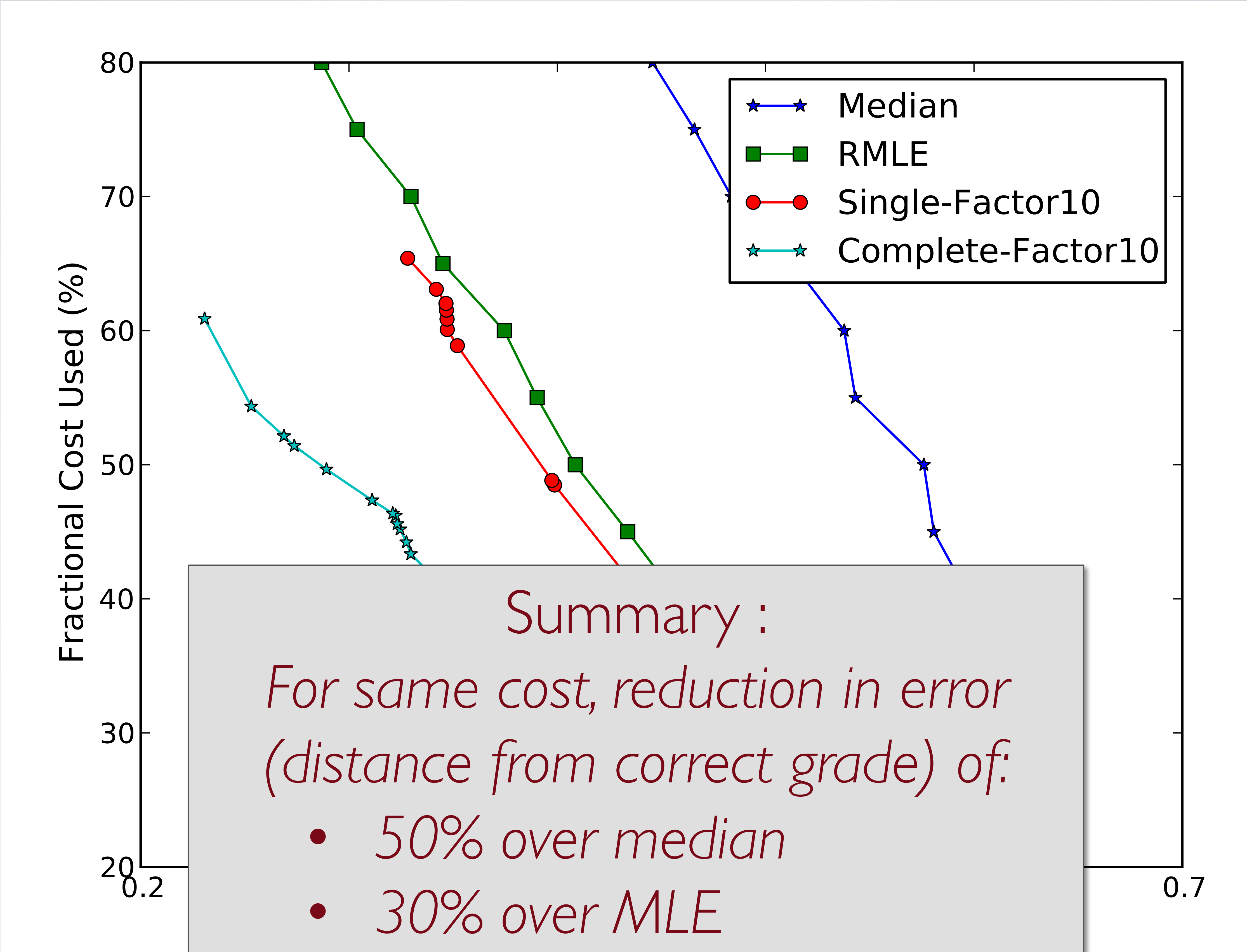
Experiments on MOOCs

Stanford HCI Course

1000  x 5  x 5 Parts = 25000 Parts

Graded by random peers with known error rates

To study: how much we can reduce error for fixed cost



Summary :

For same cost, reduction in error (distance from correct grade) of:

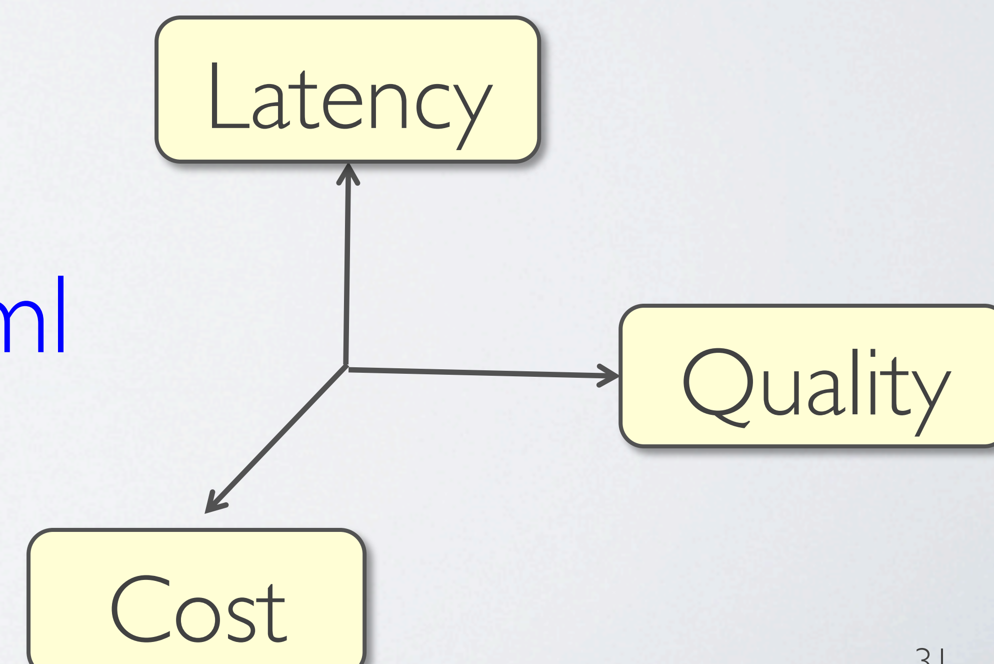
- *50% over median*
- *30% over MLE*
- *10-20% over same accuracy*

[VLDB14] Optimal Crowd-Powered Rating and Filtering Algorithms

Efficient Data Processing Algorithms & Systems

Data Processing Algorithms	**Filter [SIGMOD12,VLDB14]** Max [SIGMOD12] Clean [KDD12,TKDD13] Categorize [VLDB11] Search [ICDE14] Debugging [NIPS12]
Data Processing Systems	Deco [CIKM12, VLDB12, TR12, SIGMOD Record 12] **DataSift [HCOMP13, SIGMOD14]** HQuery [CIDR11]
Auxiliary Plugins: Quality, Pricing	Confidence [KDD13,TR14] Eviction [TR12] Pricing [VLDB15] Quality [HCOMP14]

i.stanford.edu/~adityagp/scoop.html



VISUAL DATA MANAGEMENT with SeeDB

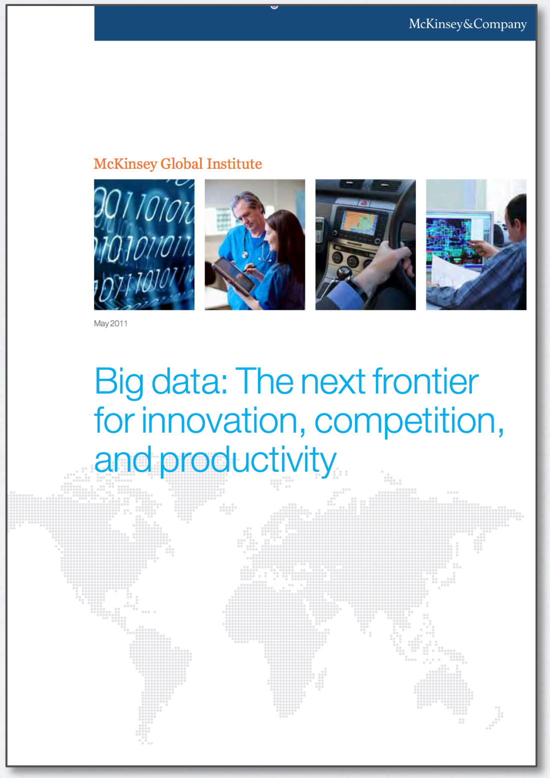
Aditya Parameswaran

with:
Hector Garcia Molina, Sam Madden,
Alkis Polyzotis, Manasi Vartak

Massachusetts Institute of Technology

Simplifying Data Analytics

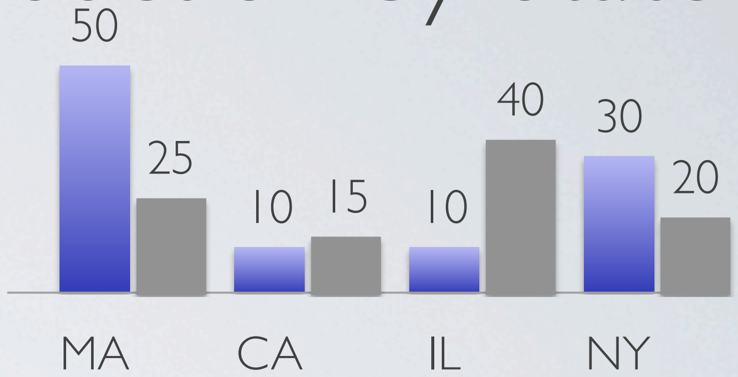
Up to a million additional analysts will be needed to address data analytics needs in 2018 in the US alone.
--- *McKinsey Big Data Report, 2013*



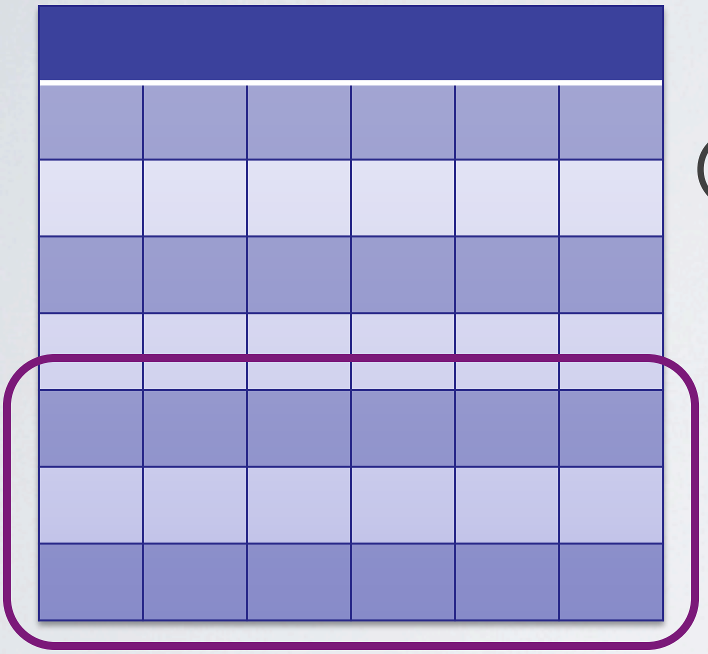
How do we make it easier for novice data analysts to get insights from data?

Data Analytics Workflow

"Production by State"

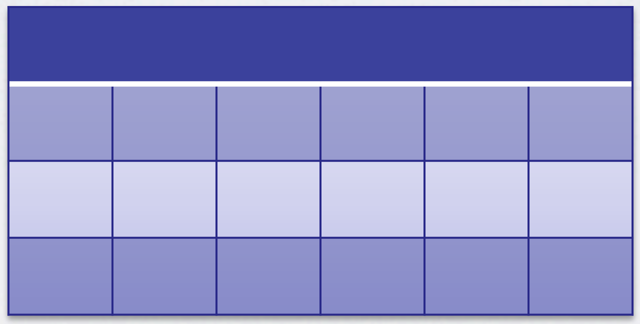


All Products

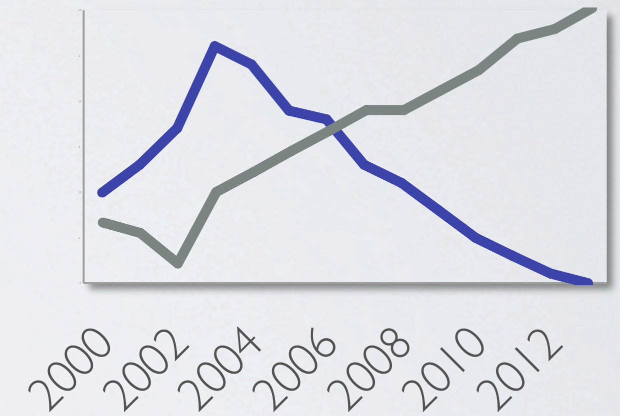


Query

"Staplers"



Views "Sales by Year"



Laborious and Tiresome!

Can we automate this?

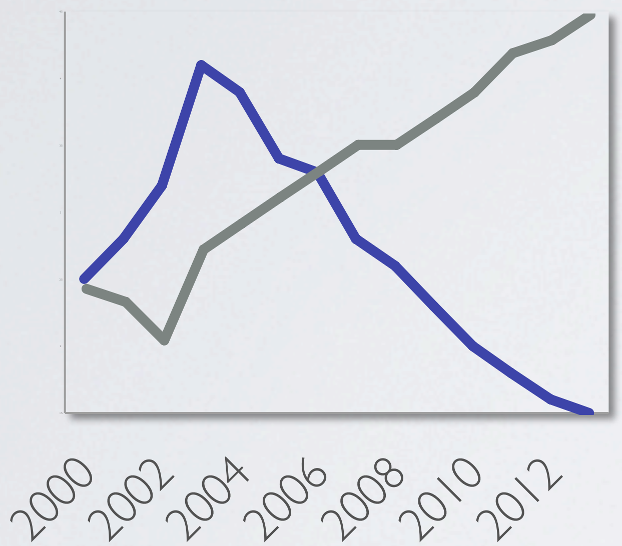
'Production by Year"

Similar issues with

Tableau, ShowMe, Profiler, Spotfire

Potentially Interesting Views(Visualizations)

"Sales by Year"

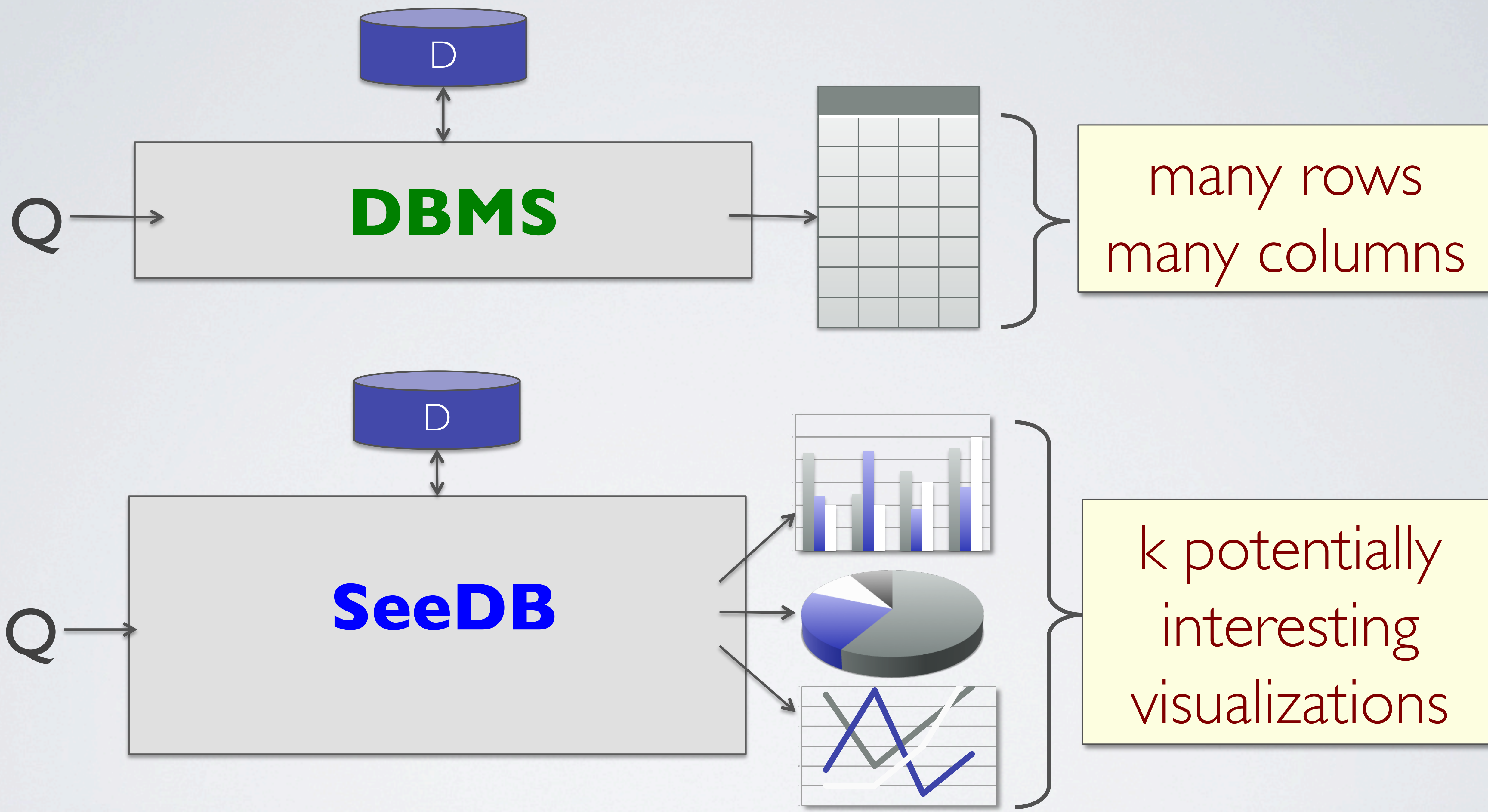


"Potentially interesting": trend in subset that is not in overall data

Can we automatically highlight potentially interesting views?

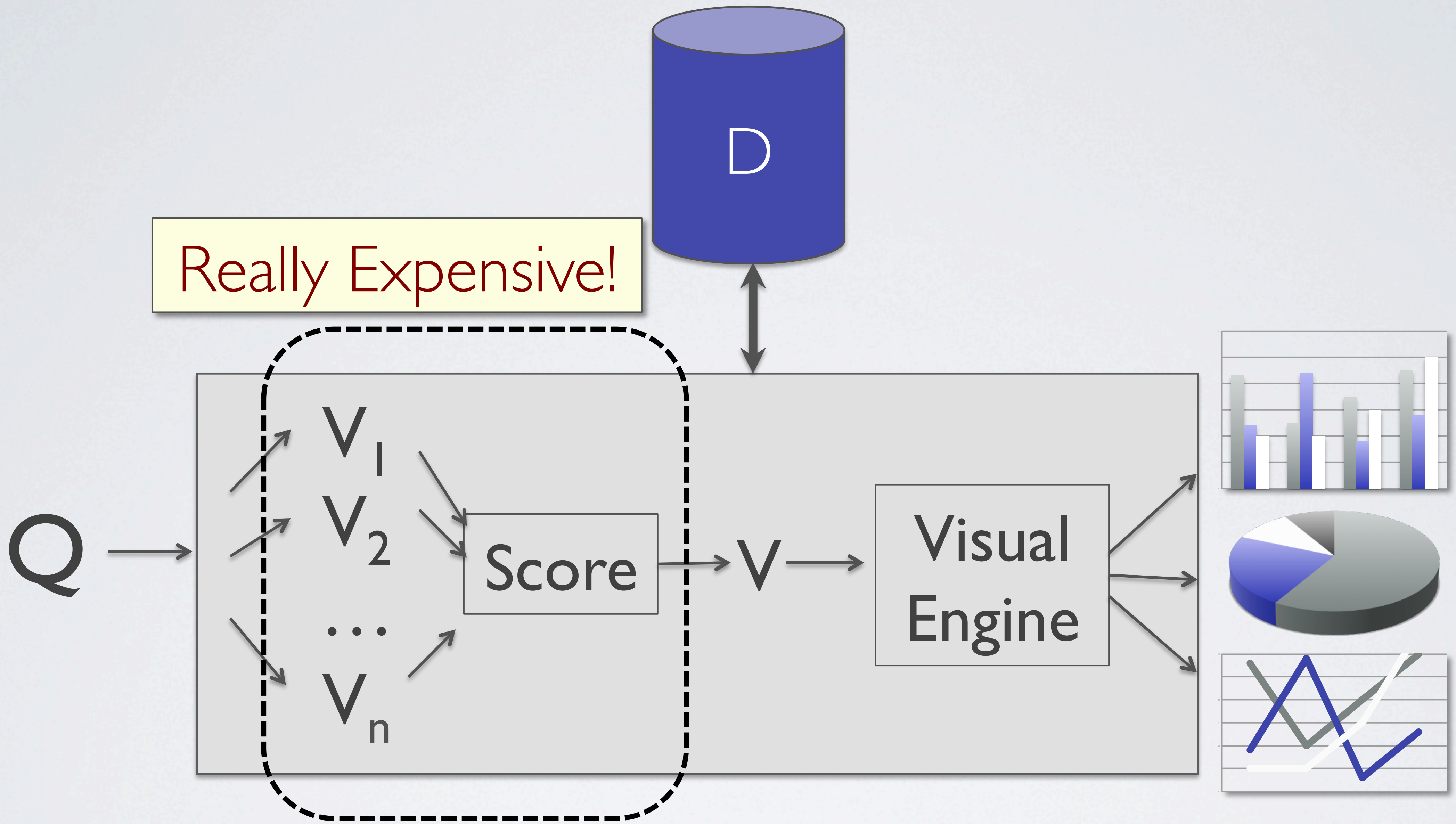
Saving: stepping through all views now only potentially interesting ones!

Our Proposed System: SeeDB



[VLDB14] SeeDB: Visualizing Database Queries Efficiently (Vision)
[VLDB14] Automatically Generating Query Visualizations (Demo)

SeeDB: Conceptual Workflow

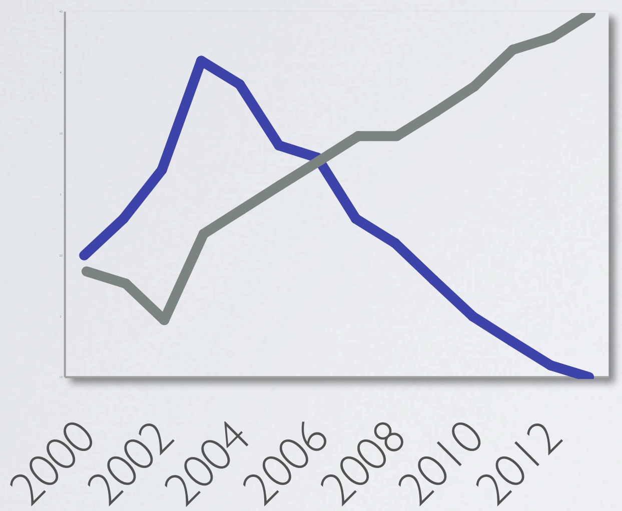


Objective : find k-best scoring views (or visualizations)

How do we score views?

This is a hard, domain-specific question!

“Sales by Year”



We are pursuing ways to learn this scoring function using crowds.

For now, a proxy that is “good-enough”

differences in “distribution”

e.g., *EMD, euclidean, KL-divergence*

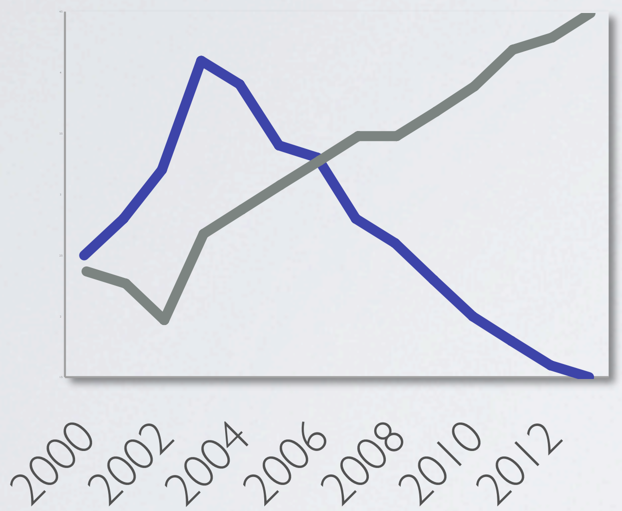
Difference(Distribution of Sales by year overall,
Distribution of Sales by year for Staplers)

our techniques work with any scoring metric

How many views to consider?

Star Schema; Histogram Visualizations

“Sales by Year”



M measure attributes

A dimension attributes

F aggregation measures

One-dimensional visualizations:

M x A x F

If we consider binning:

M x A x F x B

Building SeeDB: Concrete Directions

How do we minimize computation?

- Sharing computation
- Approximate visualizations
- Approximate scoring
- Visualization pruning

Technique 1: Sharing Computation

"Sales by Year" "Production by Year"

"Sales and Production by Year"

```
SELECT AGG(M1), AGG(M2), D,
FROM R
WHERE Prod = "Staplers"
GROUP BY D
```

Linear Speedup!

Technique 1: Sharing Computation

"Sales by Year" → "Sales by Year, Region"

"Sales by Region" → "Sales by Year, Region"

"Sales by Year, Region"

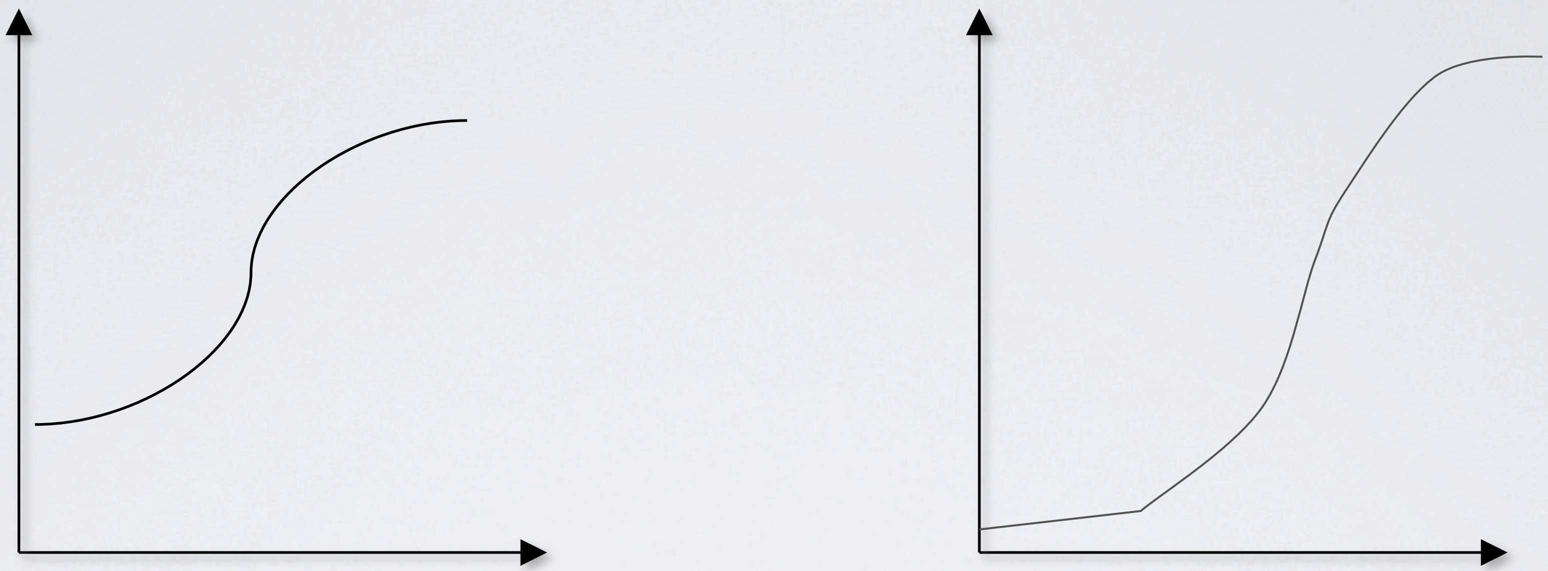
```
SELECT AGG(M), D1, D2
FROM R
WHERE Prod = "Staplers"
GROUP BY D1, D2
```

Problematic: # of aggregates grow rapidly

Intractable!

Technique 2: Approximate Visualizations

"Production by Year"



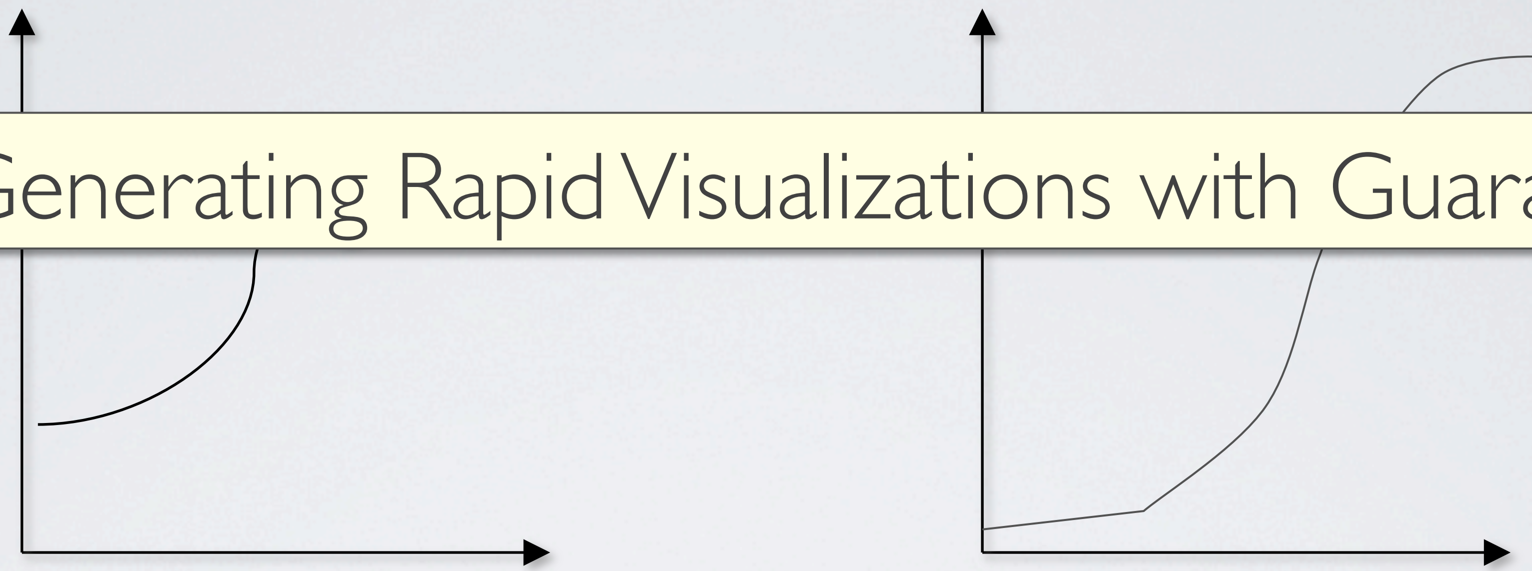
Analysts are only interested in trends, not absolutes

Limited also by resolution

*Can we provide visualizations that are **guaranteed** to look similar (e.g., similar order, similar differences) to actual ones, but at much lower cost?*

Technique 2: Approximate Visualizations

"Production by Year"



[TR14] Generating Rapid Visualizations with Guarantees

The answer is yes!

At a high-level, algorithm samples "more" from contentious areas

- Order of magnitudes saving compared to baselines
- Optimality guarantees
- Also of independent interest

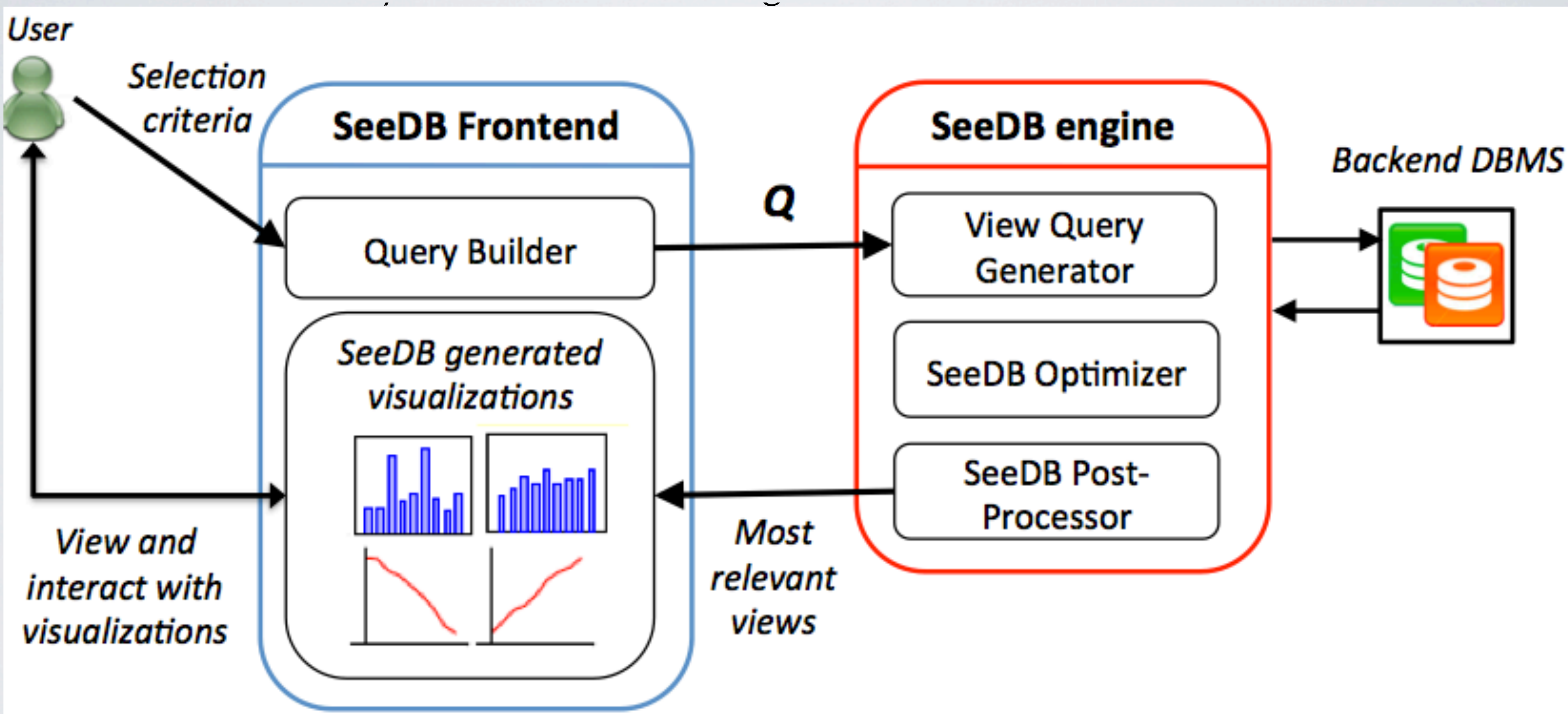
Building SeeDB: Concrete Directions

How do we minimize computation?

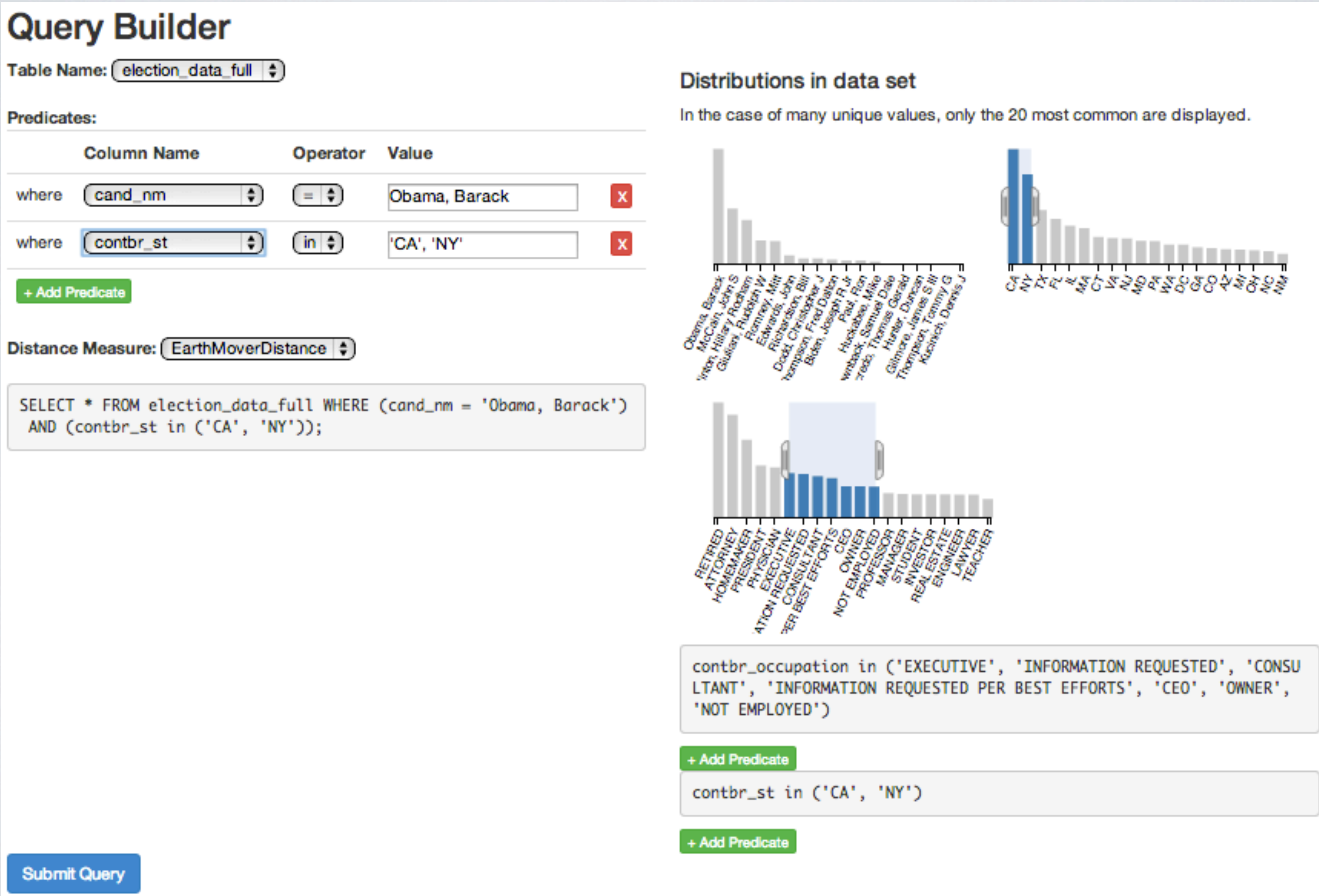
- ~~Sharing computation~~
- ~~Approximate visualizations~~
- Approximate utility computation
- Visualization pruning

Overall, a rich space of questions generalizable beyond SeeDB!

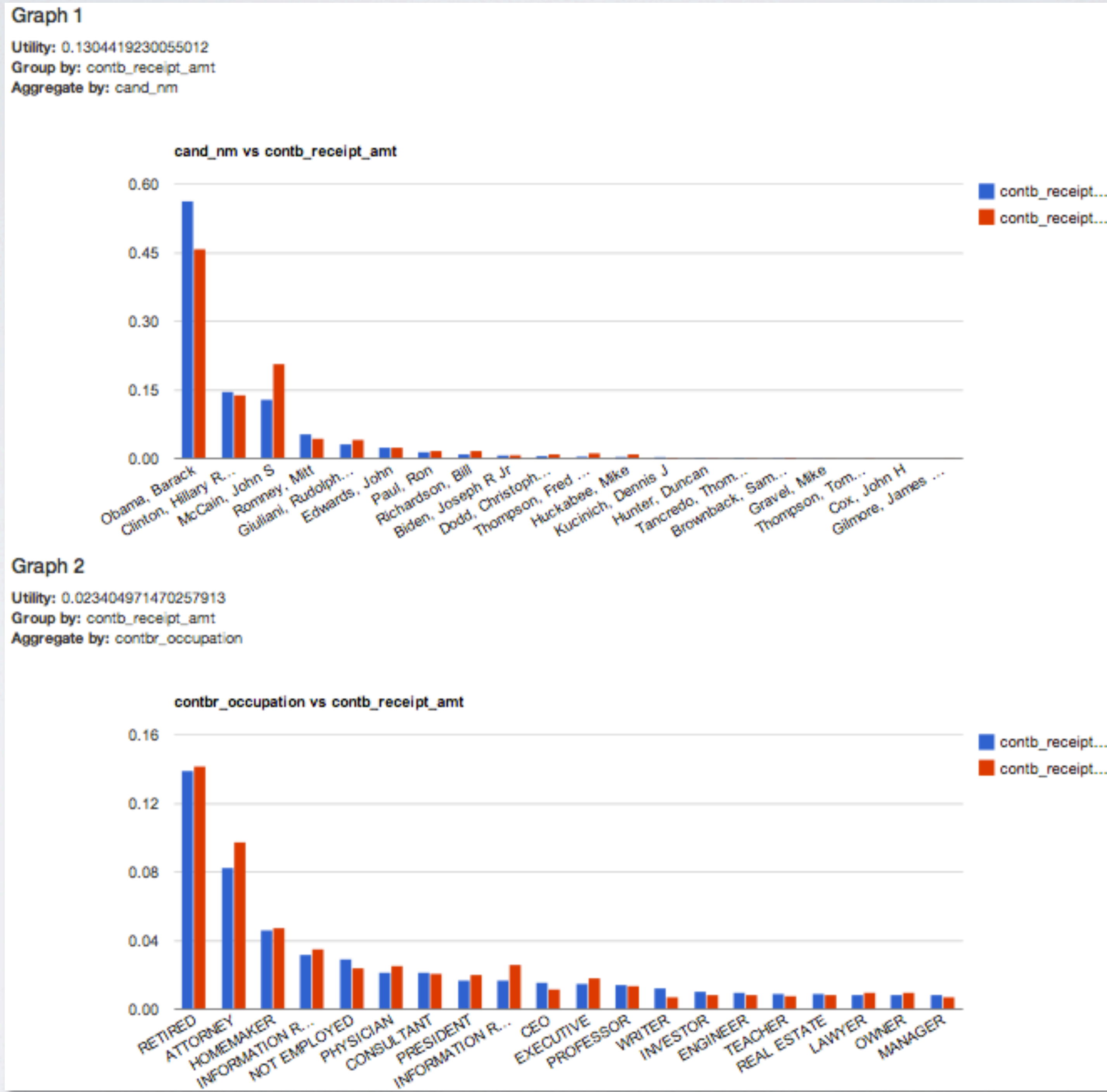
Our Current Design



Interactive Query Builder



Top-k Visualizations



To summarize…

SeeDB has some ambitious goals…

"show me all that's interesting about the query result"
i.e., the holy grail of exploratory visual data analysis

We've barely scratched the surface, yet!
… doesn't mean we can't build a useful tool