Information Management via CrowdSourcing

Hector Garcia-Molina
Stanford University

Crowdsourcing

Not to be confused with:

- Wisdom of the Crowd
- Cloud Computing :-)

Does figure show > 45 dots?

Question A

Does figure show > 45 dots?

Does figure show > 45 dots?

Report Results for Question A

Does figure show > 45 dots?

Question B

Does figure show > 45 dots?

Does figure show > 45 dots?

Report Results for Question B

Many Crowdsourcing Marketplaces!

Real World Examples

Categorizing Images

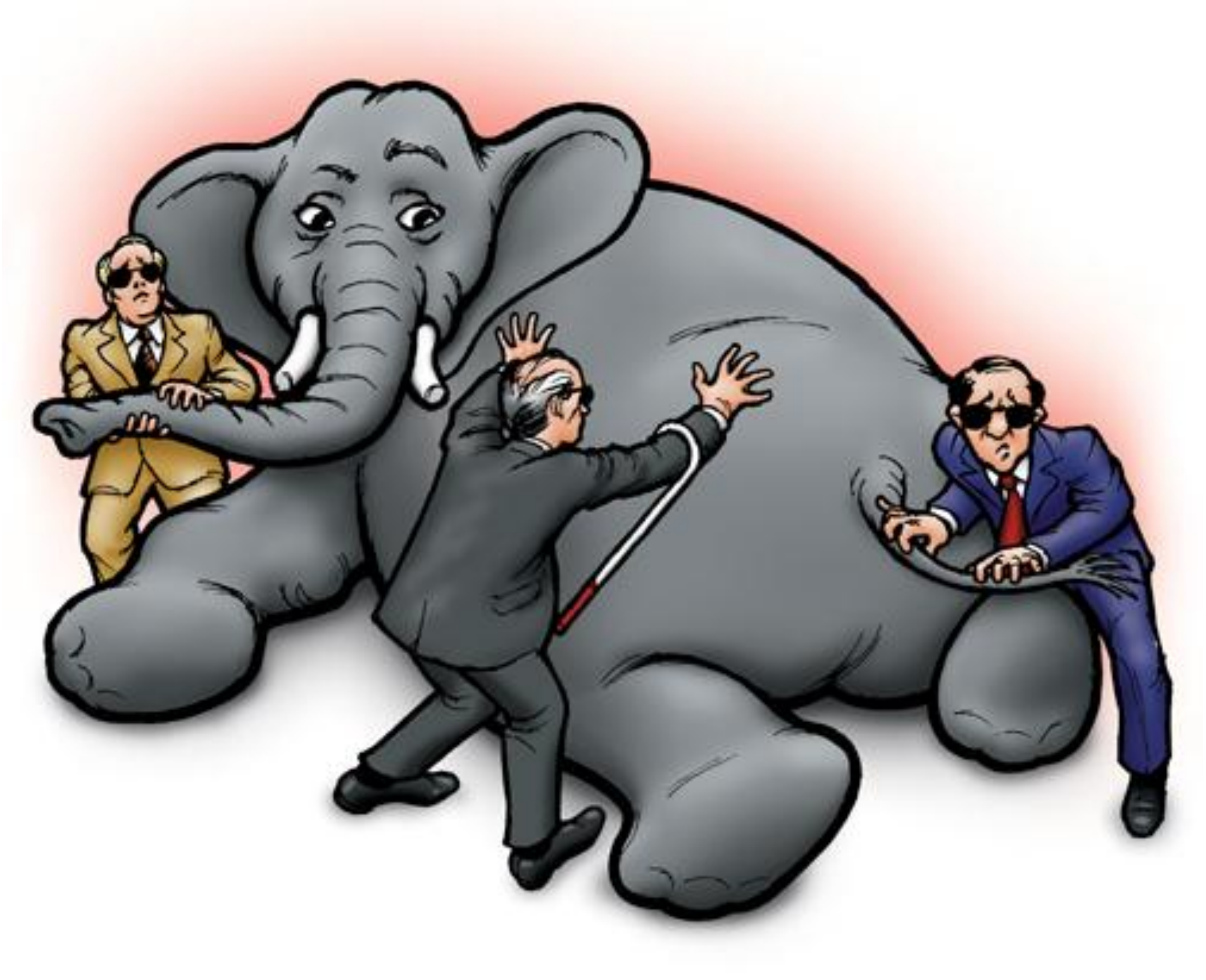
Search Relevance

Data Gathering

Image Matching
Translation

Many Research Projects!

The Many Faces of Crowdsourcing

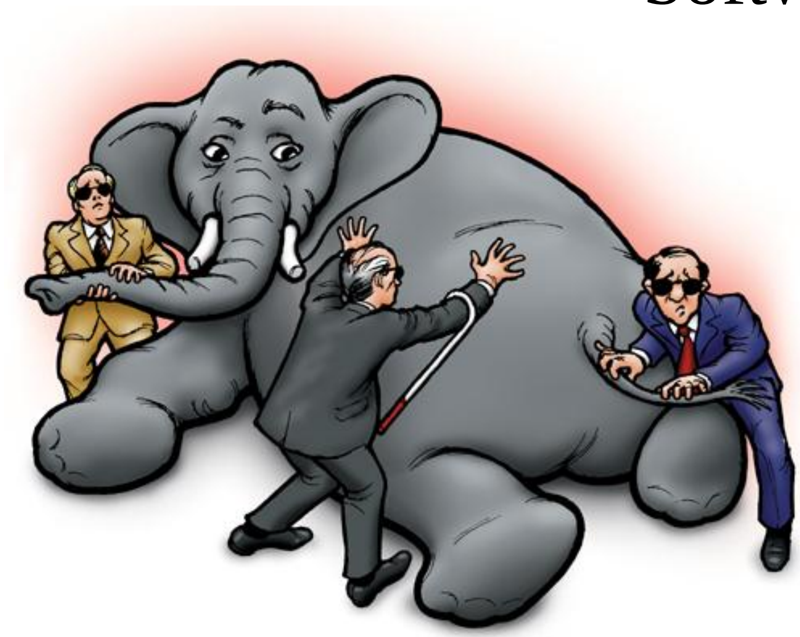


The Many Faces of Crowdsourcing

Human-Computer Interaction

Software Systems

Human Issues



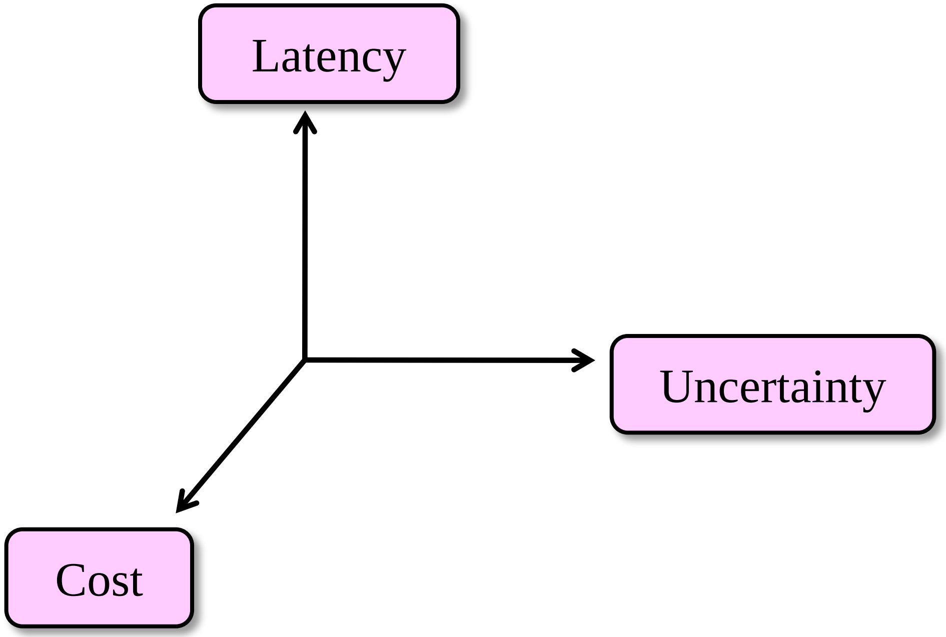
Machine Learning

Information Management

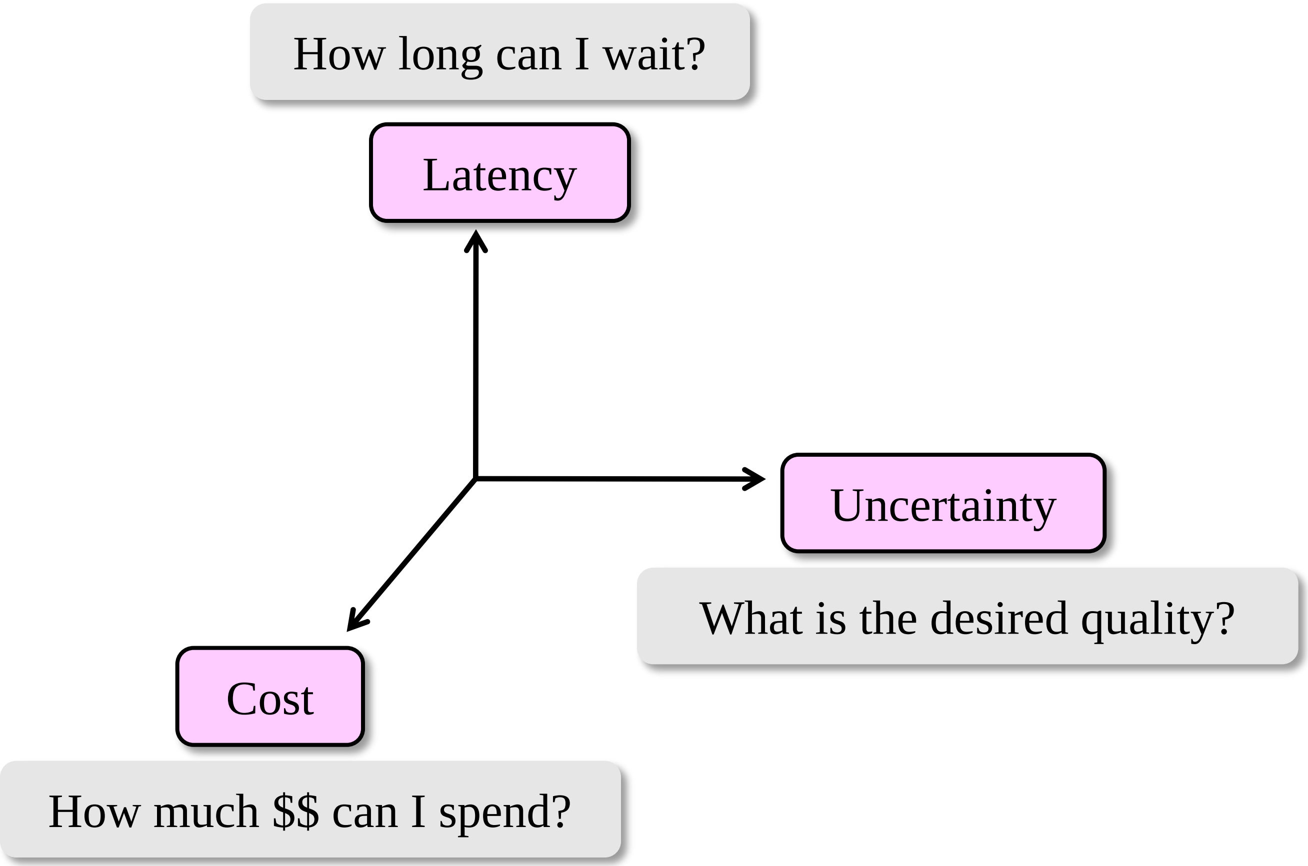
Crowd Information Management

- Two Aspects:
  - Crowd as Information Source
  - Crowd as Data Processor

Fundamental Tradeoffs

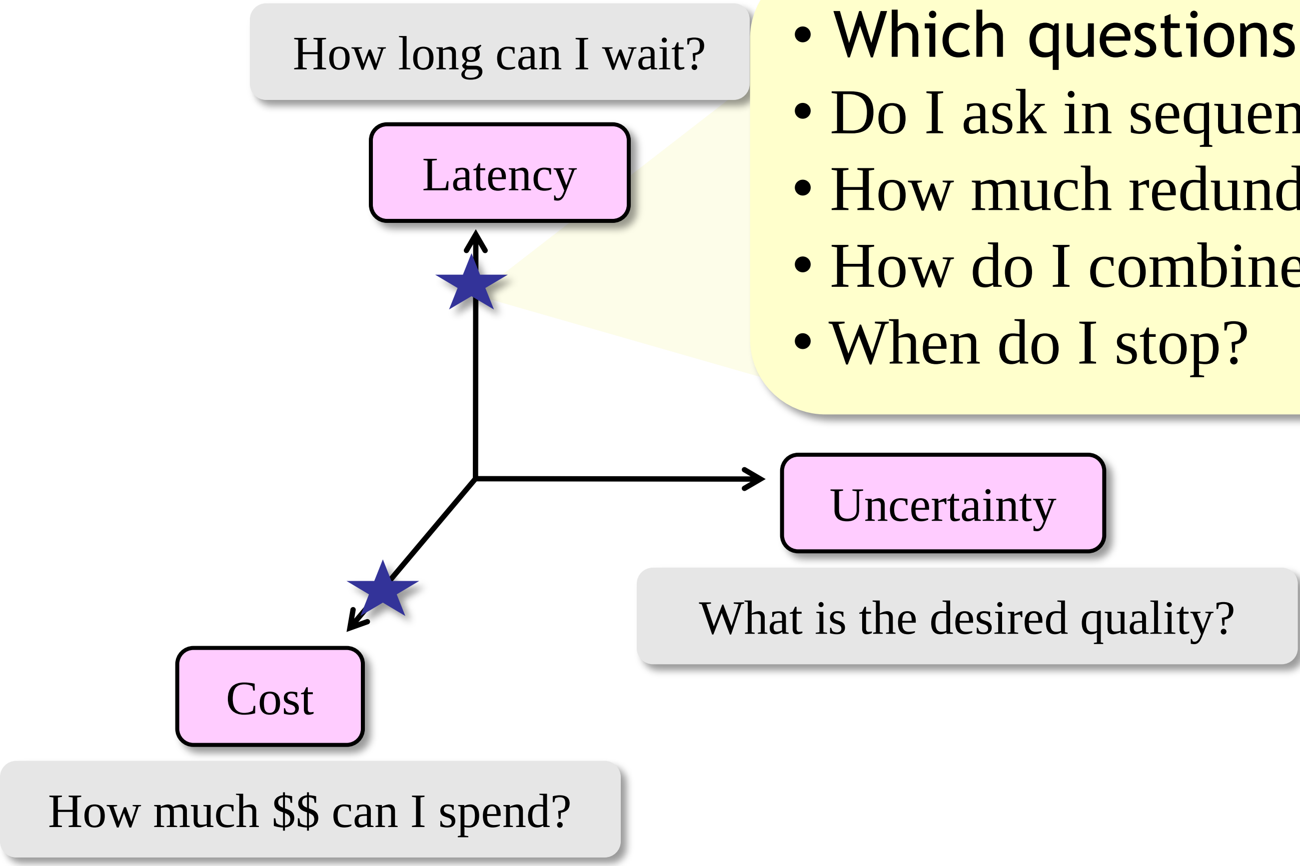


Efficiency: Fundamental Tradeoffs

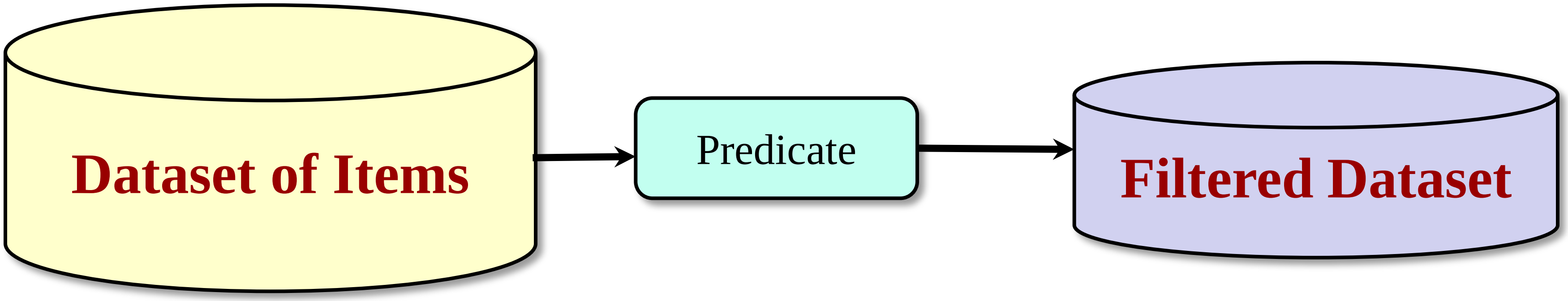


Efficiency: Fundamental Tradeoffs

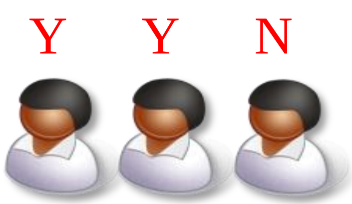
- Which questions do I ask humans?
- Do I ask in sequence or in parallel?
- How much redundancy in questions?
- How do I combine the answers?
- When do I stop?



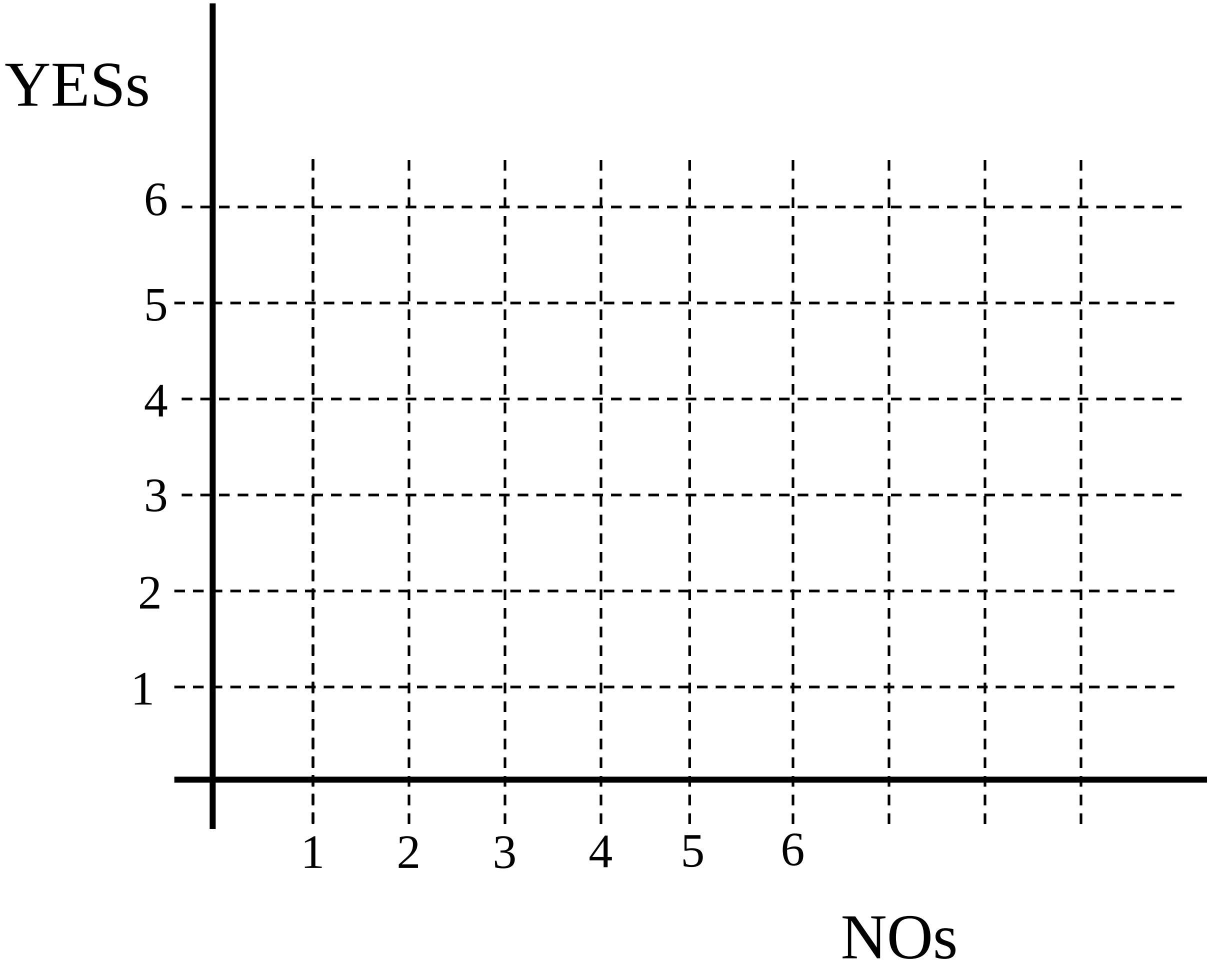
Example: CrowdScreen



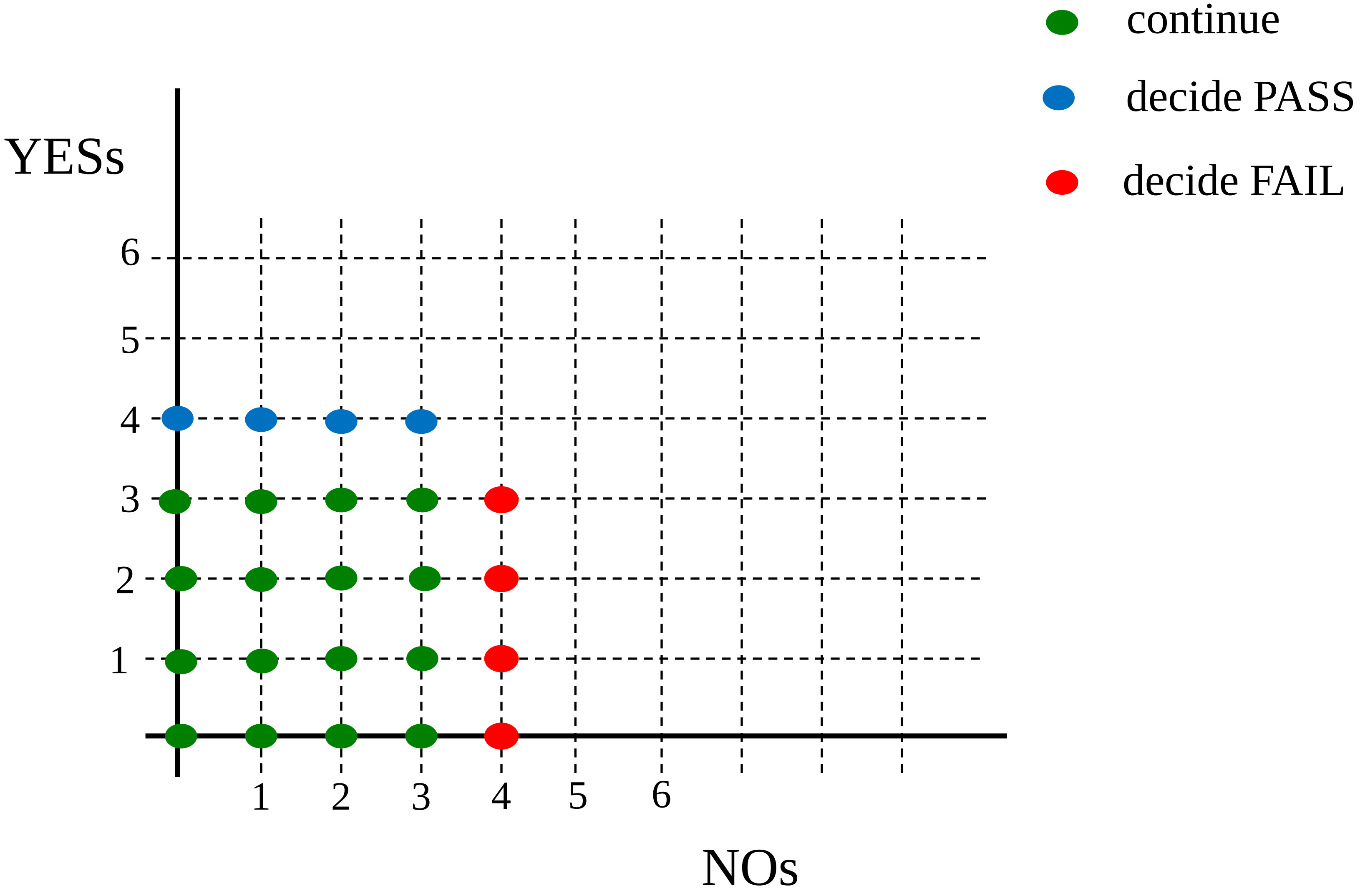
Item X satisfies predicate?



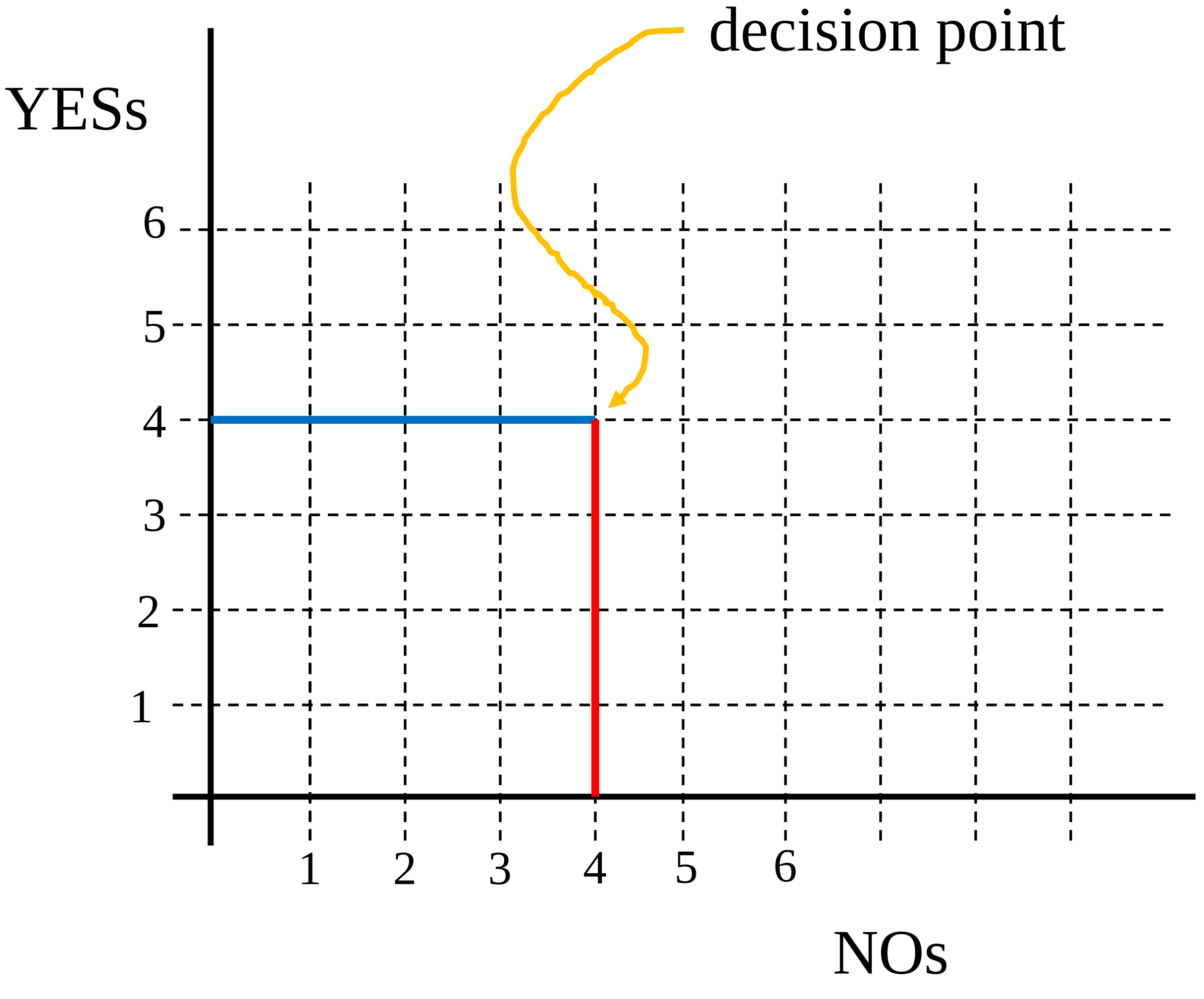
Strategy



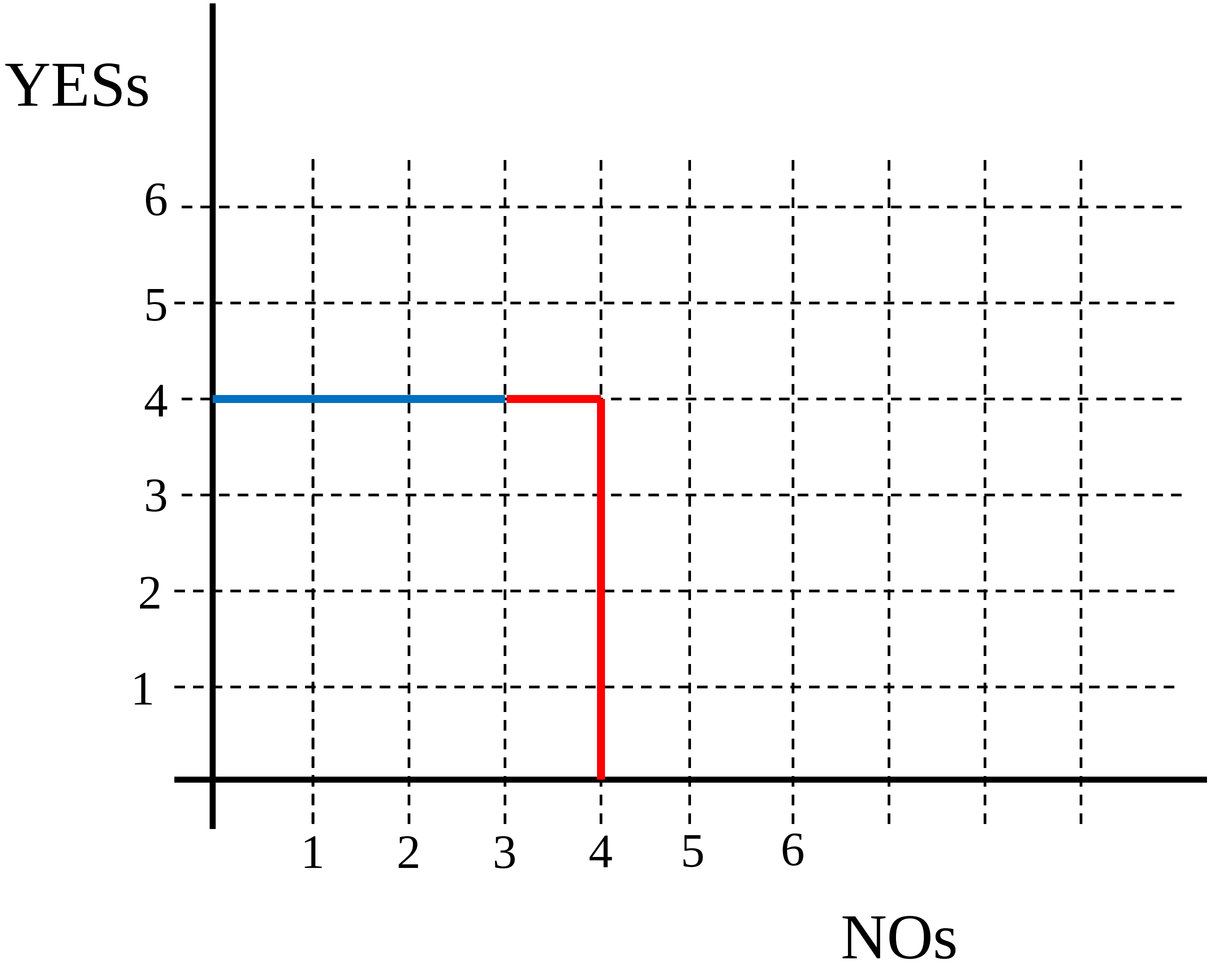
Strategy



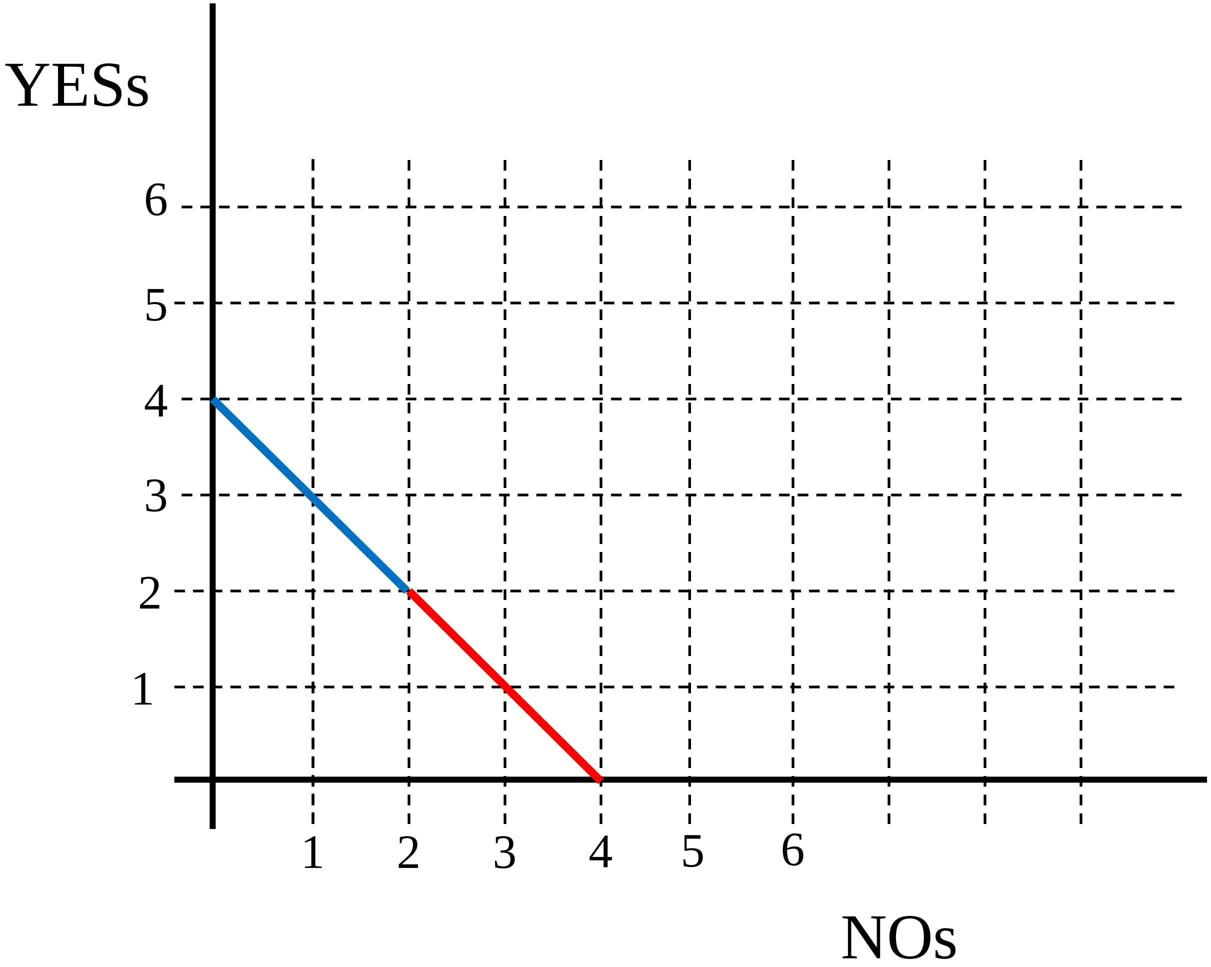
Strategy



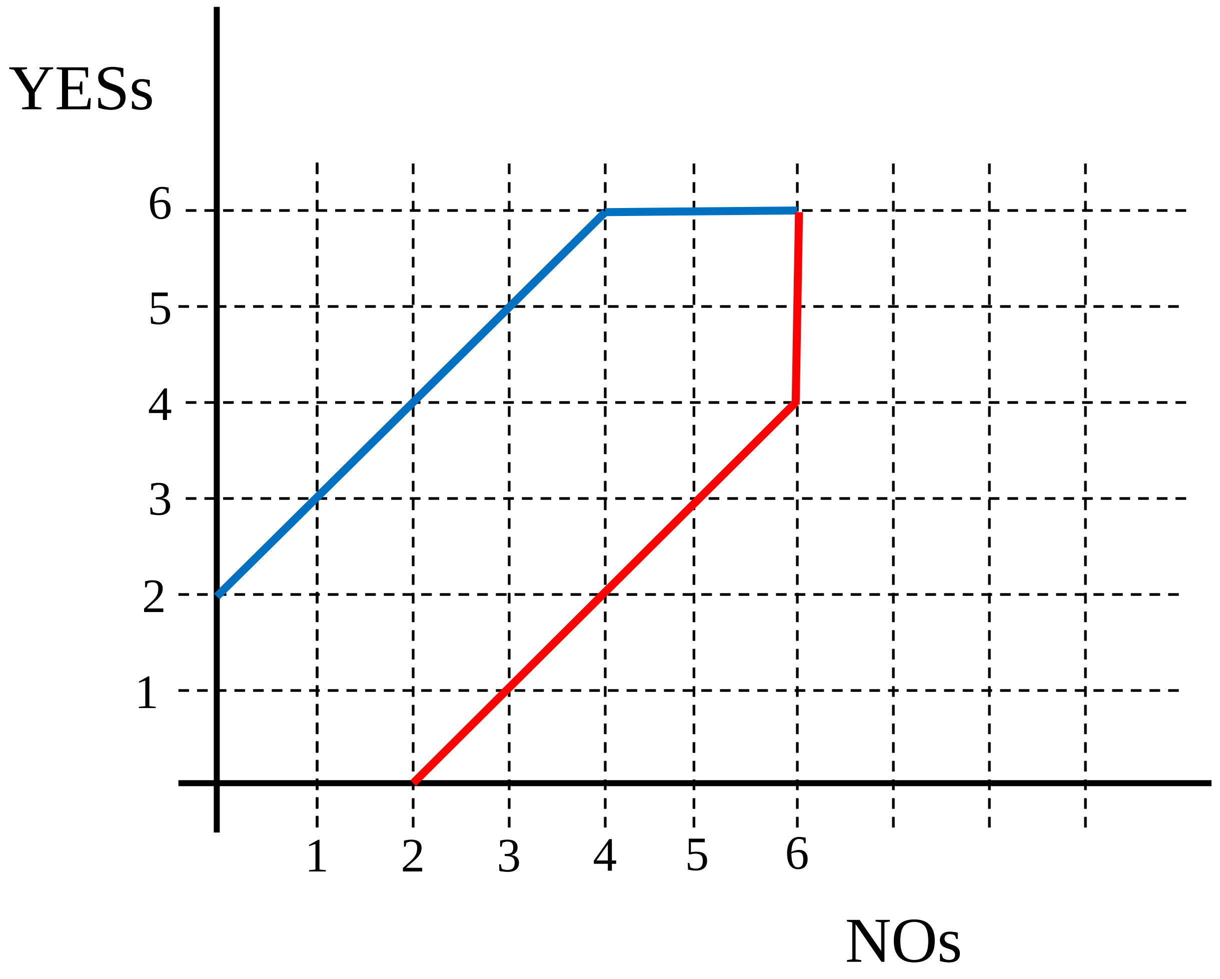
More Examples



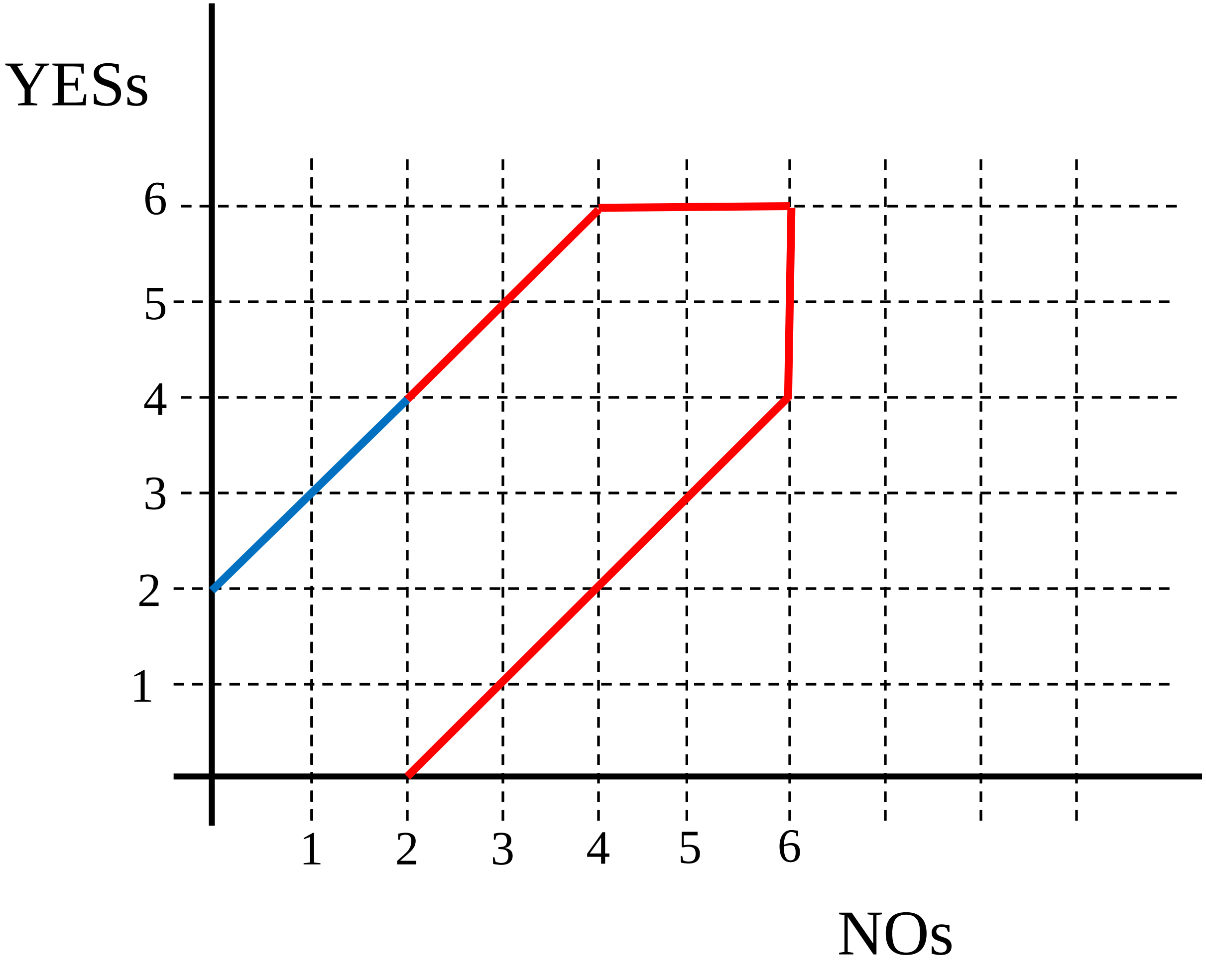
More Examples



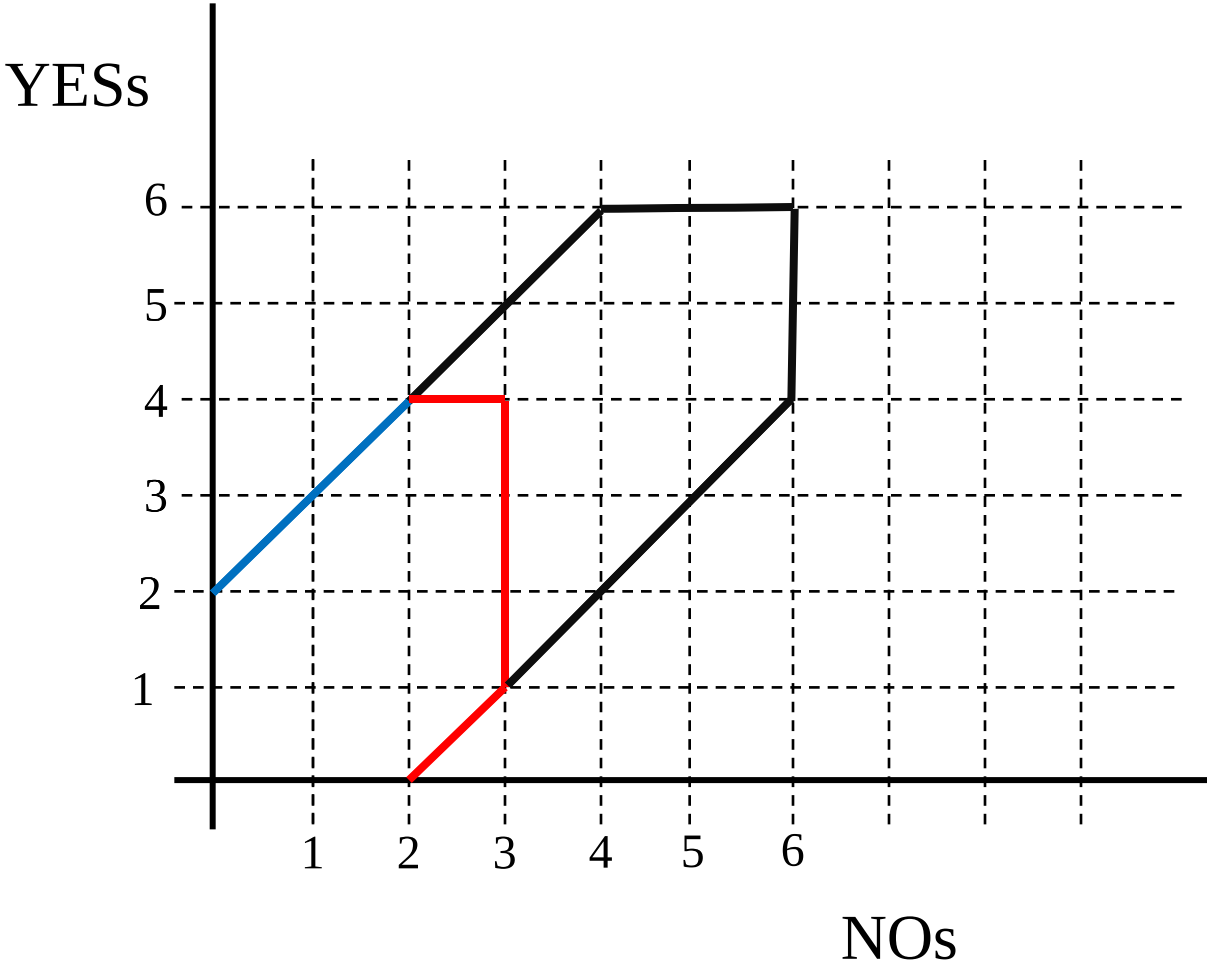
More Examples



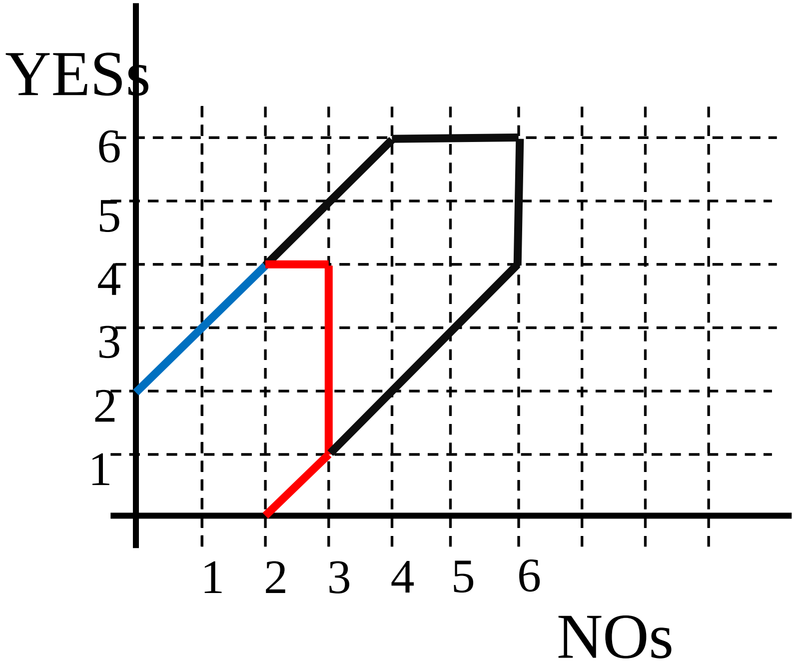
More Examples

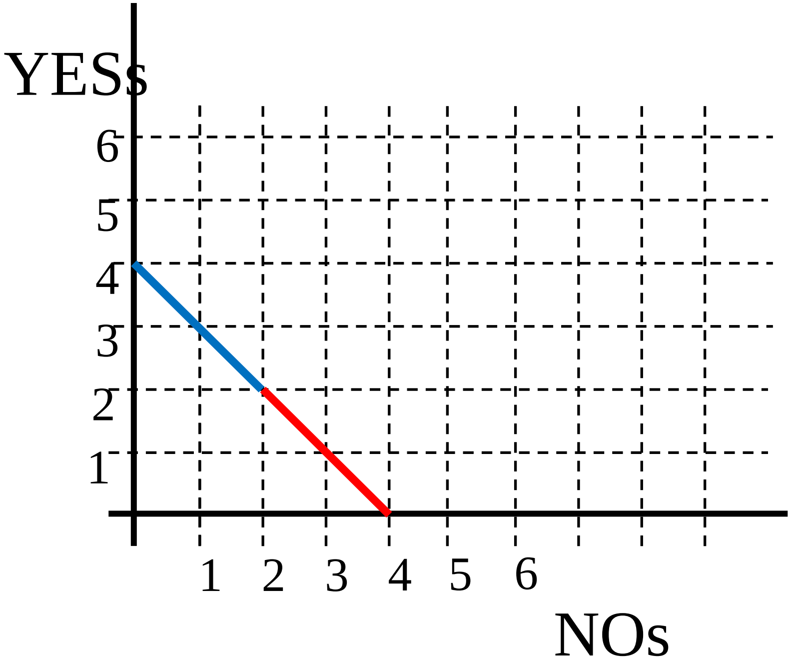


Some Optimizations



What is "best" strategy?

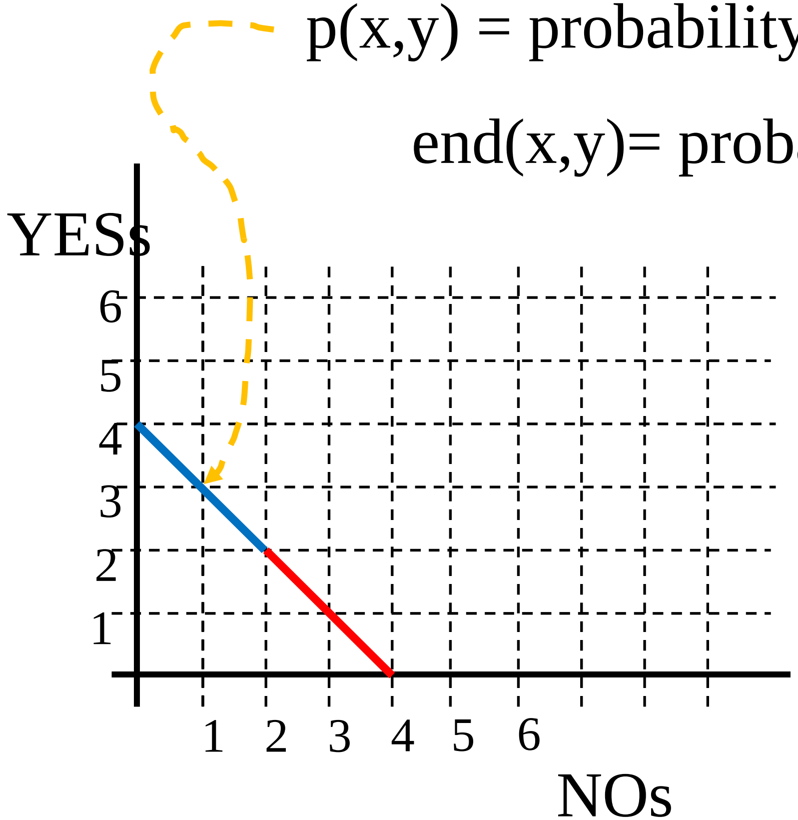




What is "best" strategy?

p(x,y) = probability of miss-classification at x,y

end(x,y)= probability of terminating at x,y



Expected error:

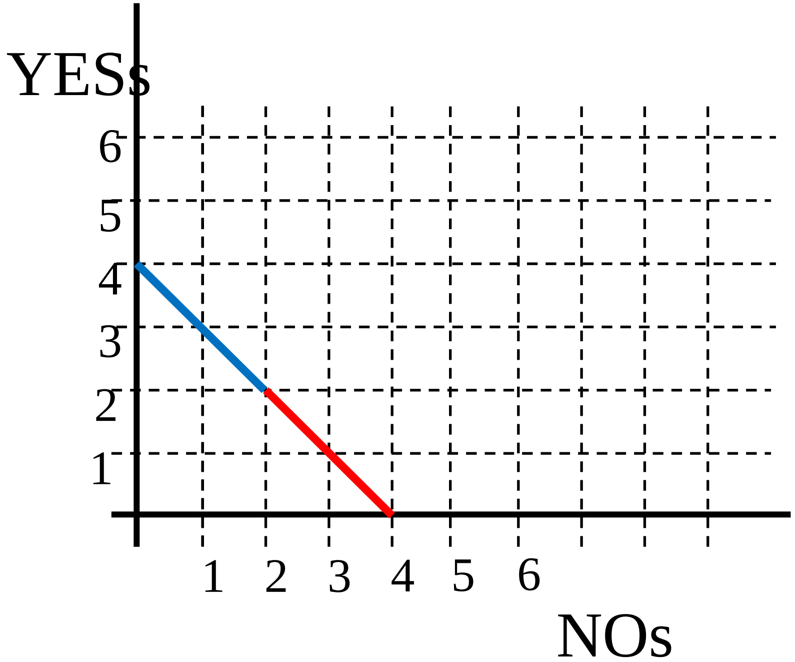
$$\sum p(x,y)*end(x,y)$$

Expected cost:

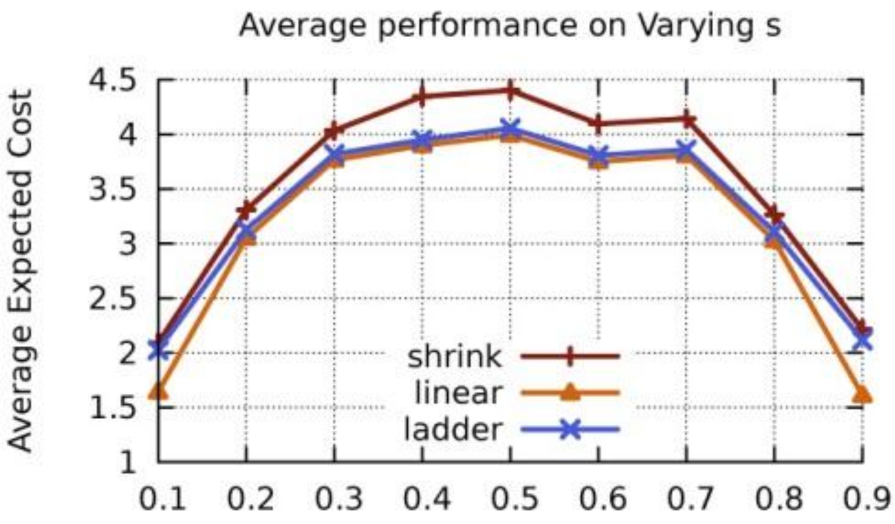
$$\sum (x+y)*end(x,y)$$

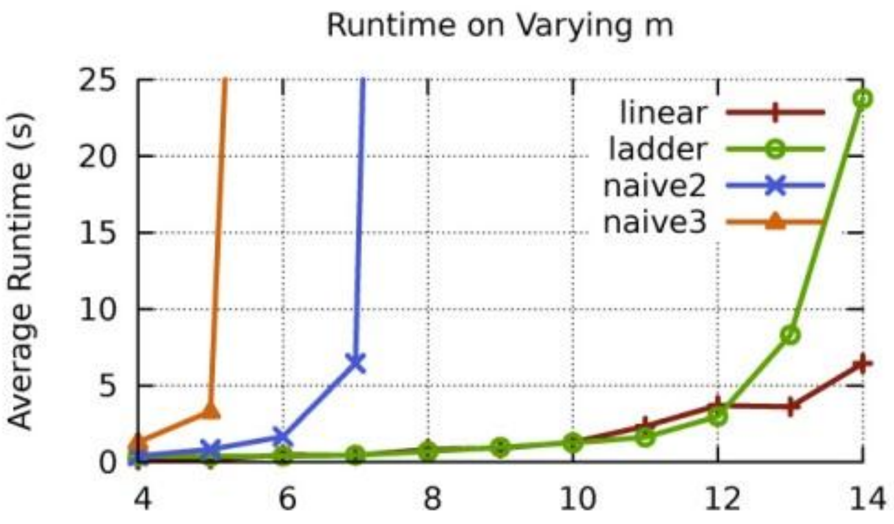
One (of many) optimization problems:

Find strategy that minimizes expected cost (# questions), such that expected error is less than threshold (and number of questions never exceeds m).



Example of Results





Beyond Single Filter

- Probabilistic Strategies
- Multiple Filters
- Categorizer (output more than 2 types)

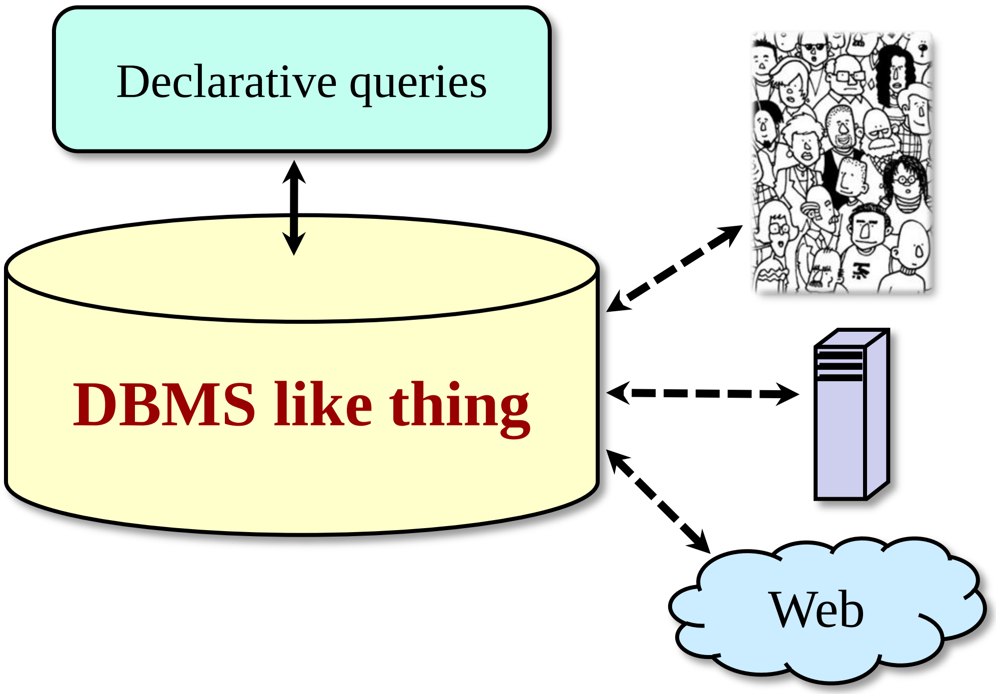
Beyond Filtering

- Finding Max
- Sorting
- Clustering
- Entity Resolution
- Adding terms to a taxonomy
- Building a Folksonomy
- ...

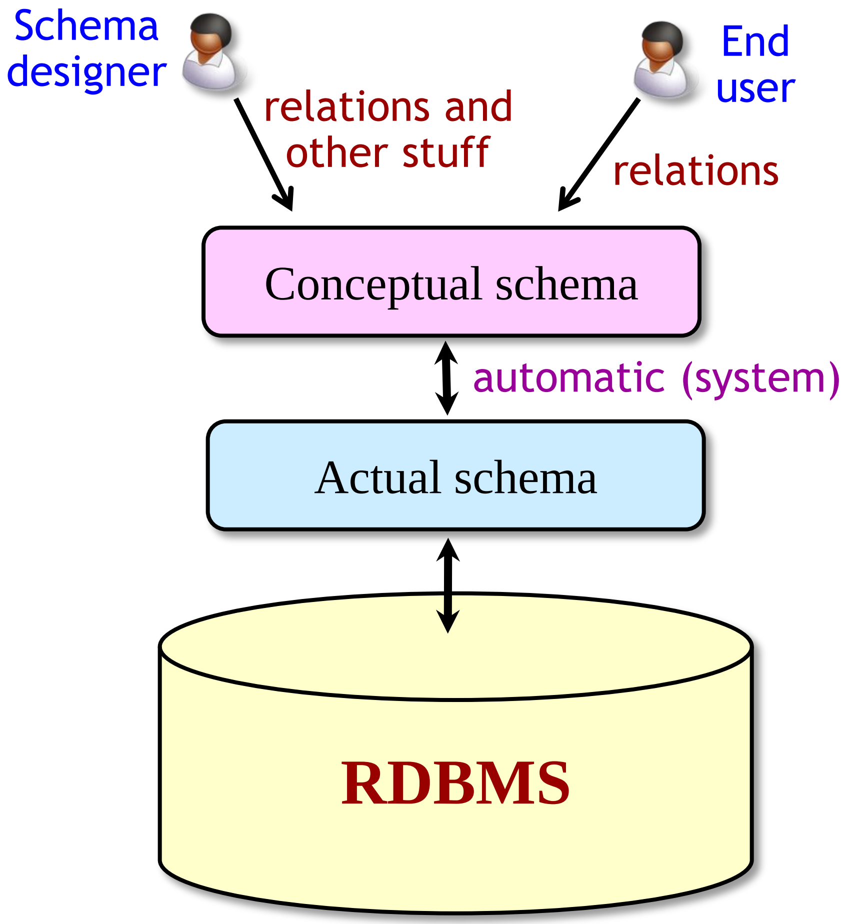
Beyond Simple Models

- Worker Error Models
- Task Design
- Tracking Worker Abilities
- Payments
- Response Time Issues
- ...

Crowd As Information Source



The Deco Data Model



Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
• • •	• • •	• • •

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
• • •	• • •	• • •

$\overset{o}{\bowtie}$

restaurant
Chez Panisse
Bytes
• • •

Anchor

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
• • •	• • •

Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
• • •	• • •
• • •	• • •

Dependent

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
• • •	• • •	• • •

$\overset{o}{\bowtie}$

restaurant
Chez Panisse
Bytes
• • •

Anchor

fetch rule

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
Bytes	• • •

fetch rule

Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	• • •
• • •	• • •

Dependent

fetch rule

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
• • •	• • •	• • •

$\bowtie$

restaurant
Chez Panisse
Bytes
• • •

Anchor

fetch rule

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
Bytes	• • •

fetch rule

Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	• • •
• • •	French

fetch rule

Dependent

fetch rule

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
• • •	• • •	• • •

resolution rule

resolution rule

restaurant
Chez Panisse
Bytes
• • •

Anchor

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
Bytes	• • •

Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	• • •
• • •	• • •

Dependent

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
• • •	• • •	• • •

1. Fetch
2. Resolve
3. Join

restaurant
Chez Panisse
Bytes
• • •

Anchor

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
• • •	• • •

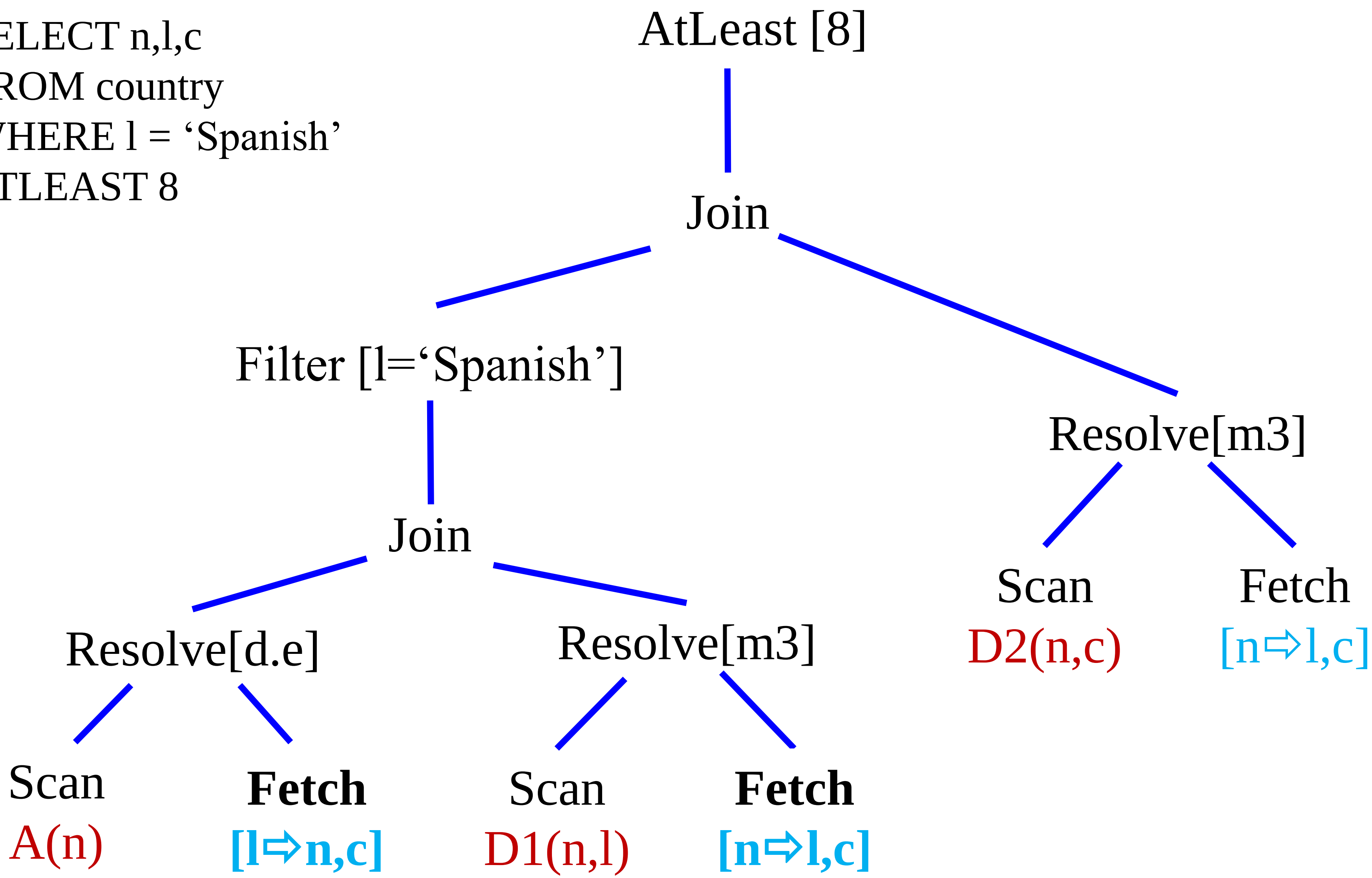
Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
• • •	• • •
• • •	• • •

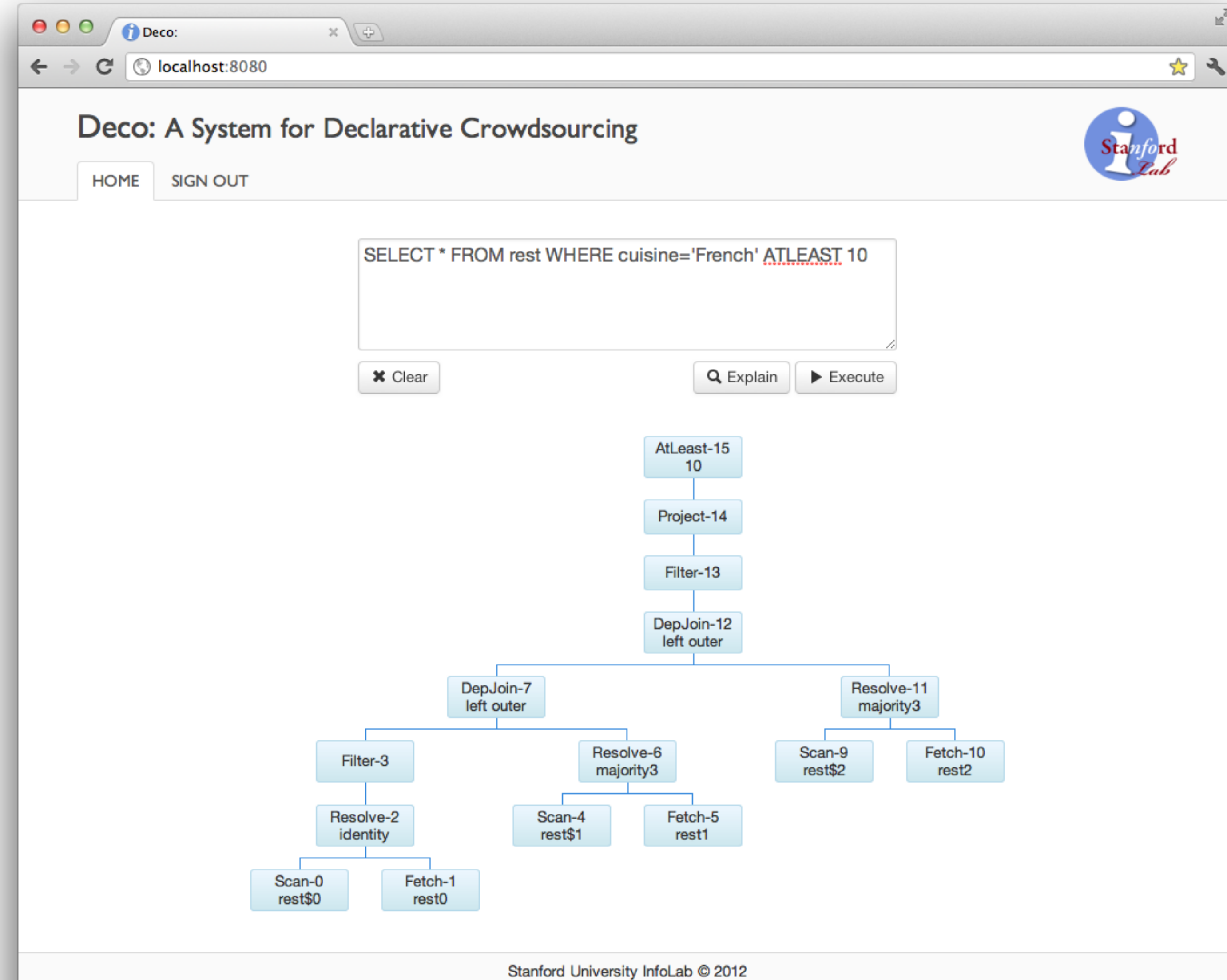
Dependent

Many Query Processing Challenges

SELECT n,l,c
FROM country
WHERE l = ‘Spanish’
ATLEAST 8



Deco Prototype V1.0

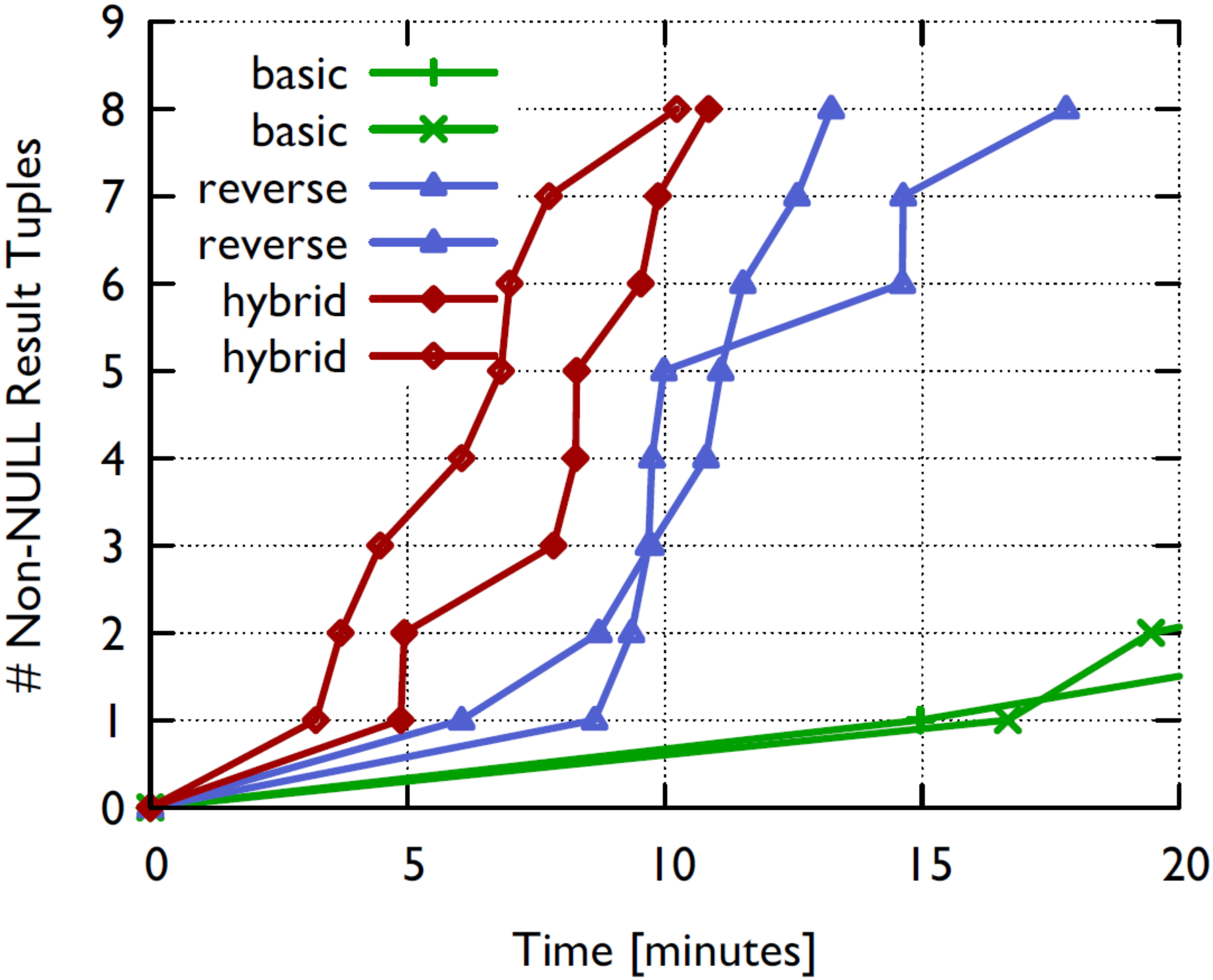


Experimental Setup

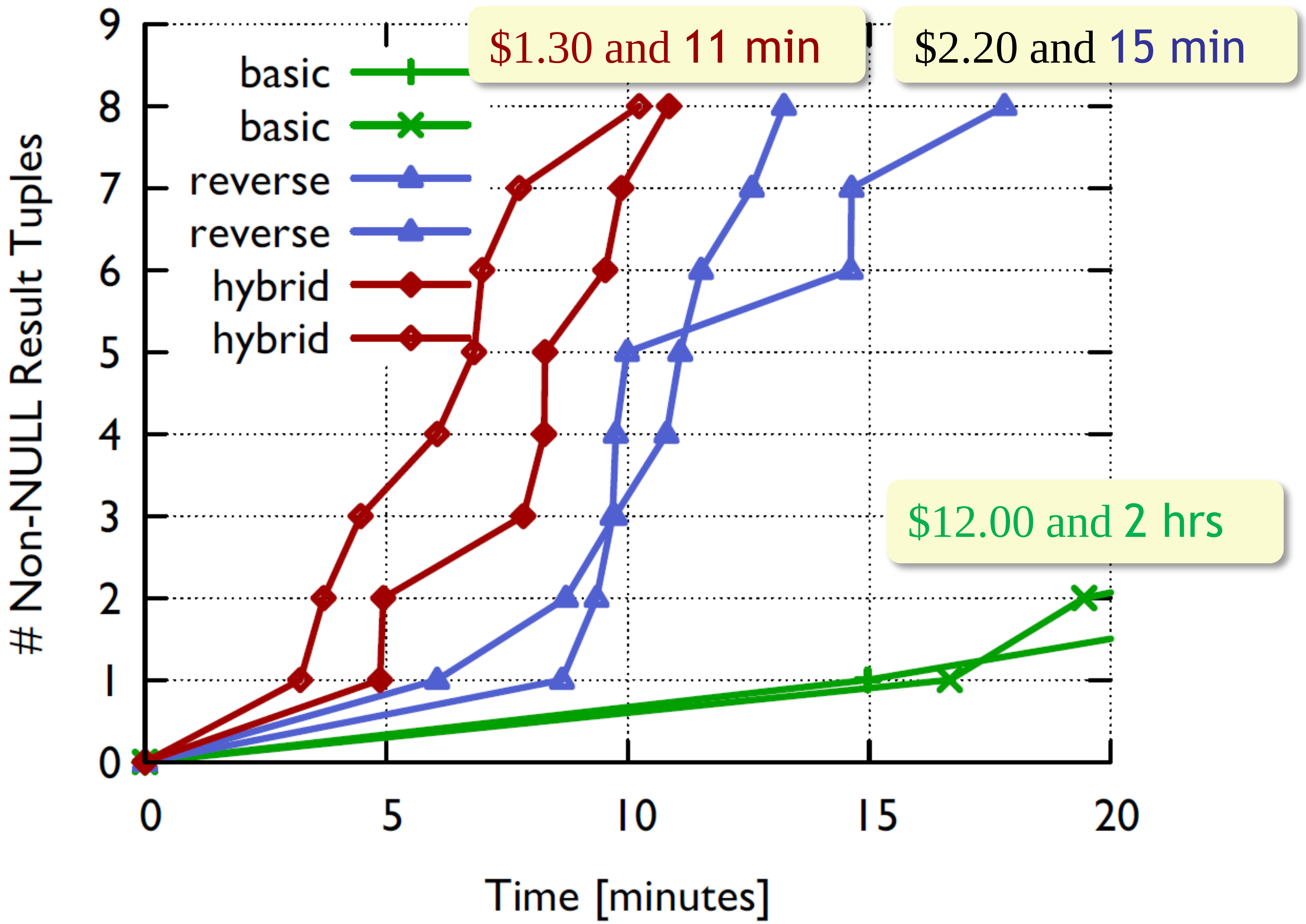
SELECT n,l,c FROM country WHERE l = ‘Spanish’ ATLEAST 8

- Experimental Goals:
  - Different fetch configurations
    - Basic: $\varnothing \Rightarrow n$ + $n \Rightarrow l$ + $n \Rightarrow c$
    - Reverse: $l \Rightarrow n$ + $n \Rightarrow l$ + $n \Rightarrow c$
    - Hybrid: $l \Rightarrow n,c$ + $n \Rightarrow l,c$
  - Different filter locations
    - After vs. between joins
- Experimental Setup:
  - 5 cents/task on MTurk
  - Empty tables initially
  - Default: reverse + between

Experiment 1: Different Fetch Rules



Experiment 1: Different Fetch Rules



Conclusion

- Crowdsourcing is exciting area!
- Many challenges!