



Information Management via CrowdSourcing

Hector Garcia-Molina

Stanford University

Crowdsourcing

REWARD
(\$5,000.00)

Reward for the capture, dead or alive,
of one Wm. Wright, better known as
"BILLY THE KID"

Age, 18. Height, 5 feet, 3 inches.
Weight, 125 lbs. Light hair, blue
eyes and even features. He is
the leader of the worst band of
desperadoes the Territory has
ever had to deal with. The above
reward will be paid for his capture
or positive proof of his death.

JIM DALTON, Sheriff.



DEAD OR ALIVE!
"BILLY THE KID"

Not to be confused with:

- Wisdom of the Crowd
- Cloud Computing :-)

Does figure show > 45 dots?

Question A

Does figure show > 45 dots?



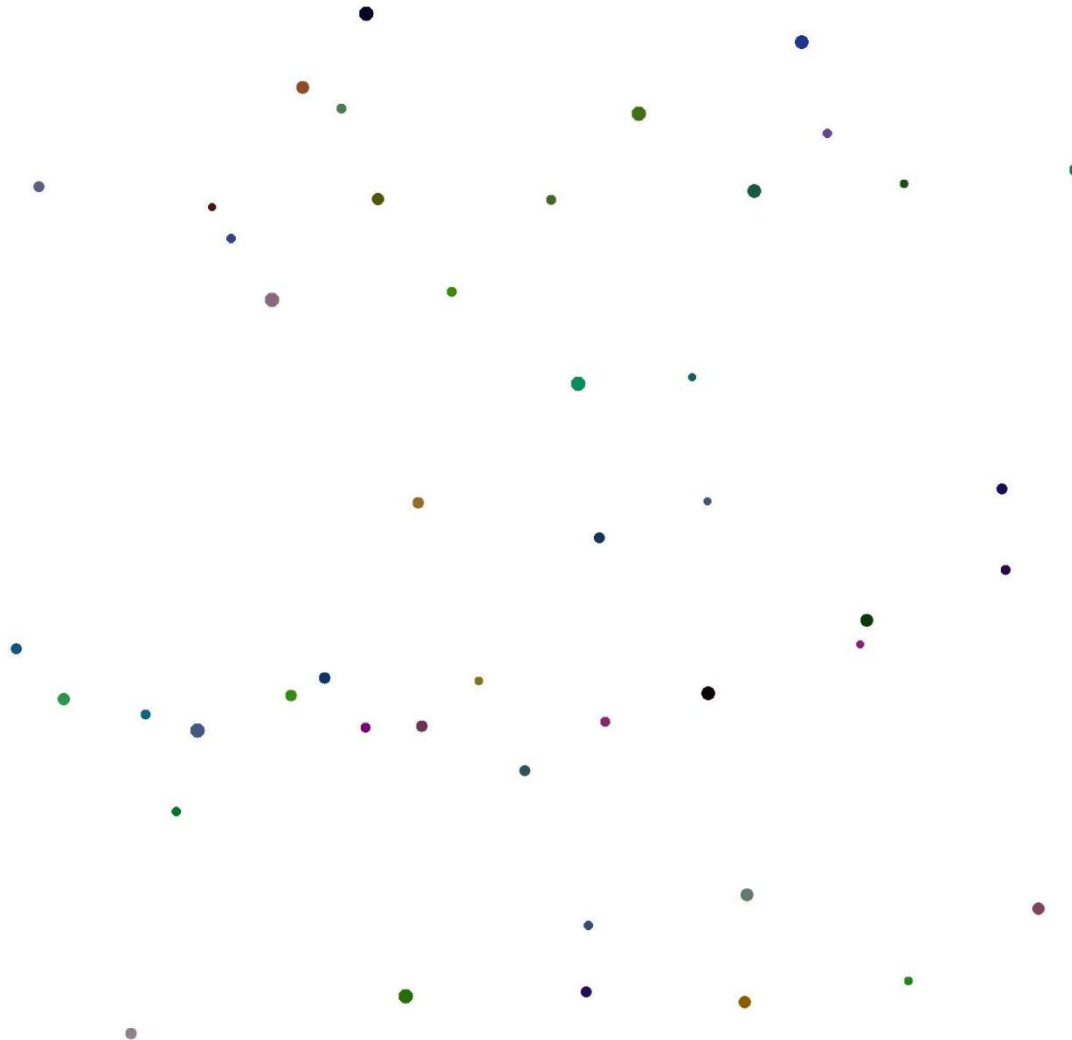
Does figure show > 45 dots?

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

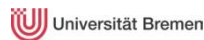
CrowdFlower



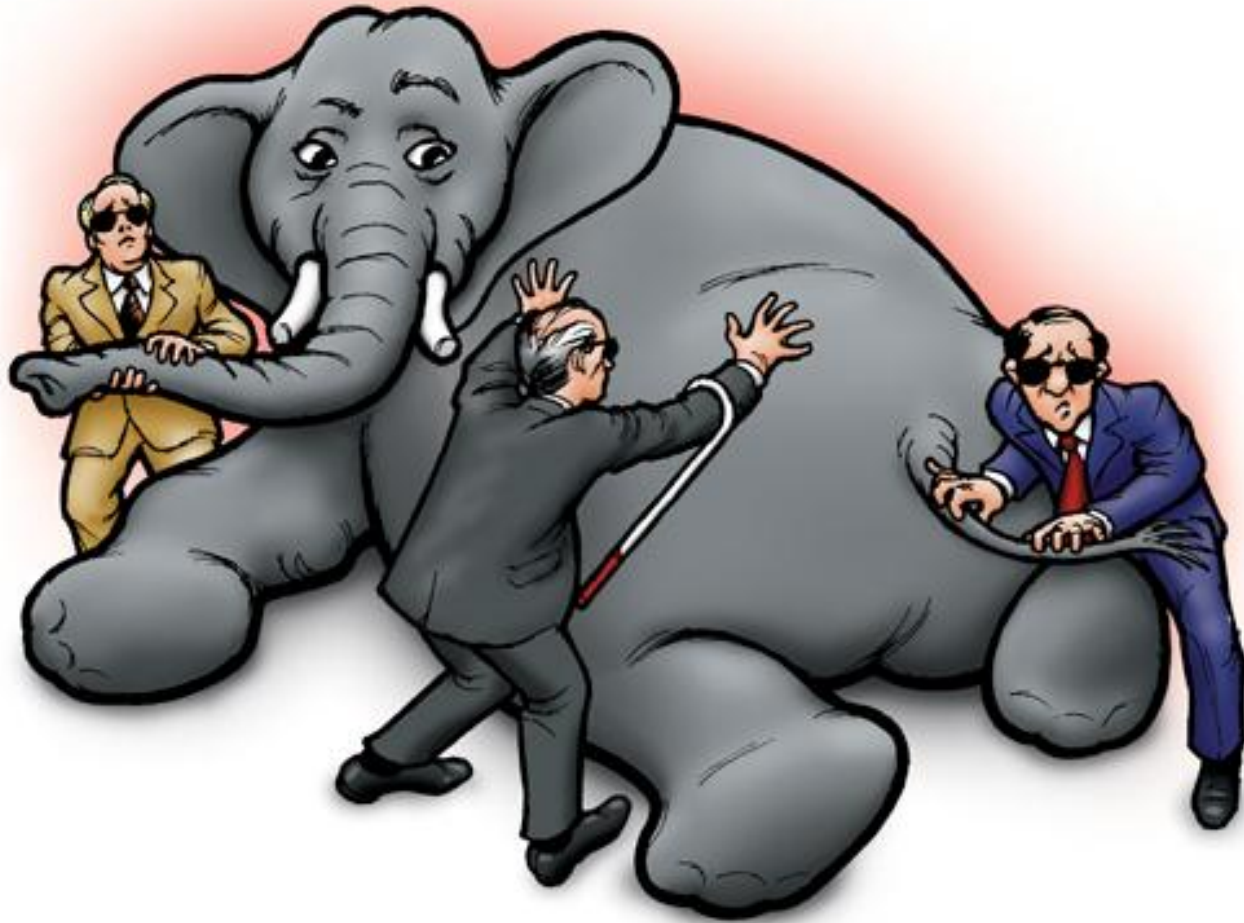
Image Matching
Translation



Many Research Projects!



The Many Faces of Crowdsourcing

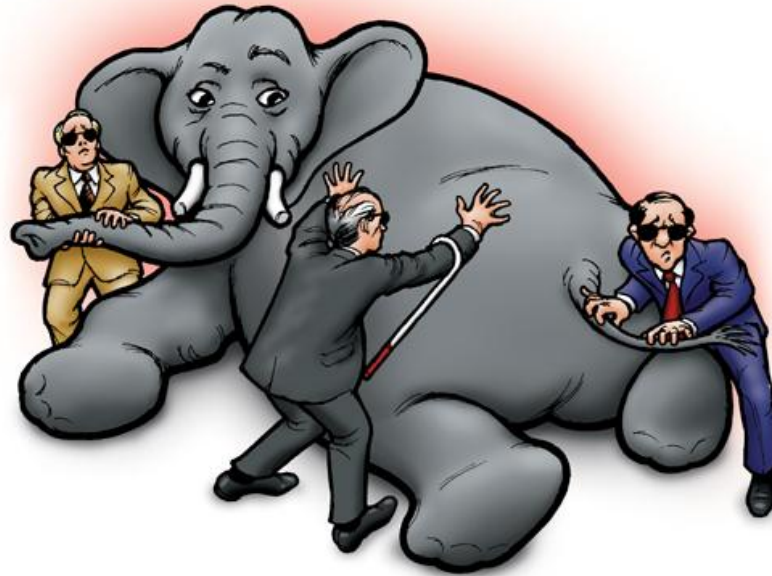


The Many Faces of Crowdsourcing

Human-Computer Interaction

Software Systems

Human Issues



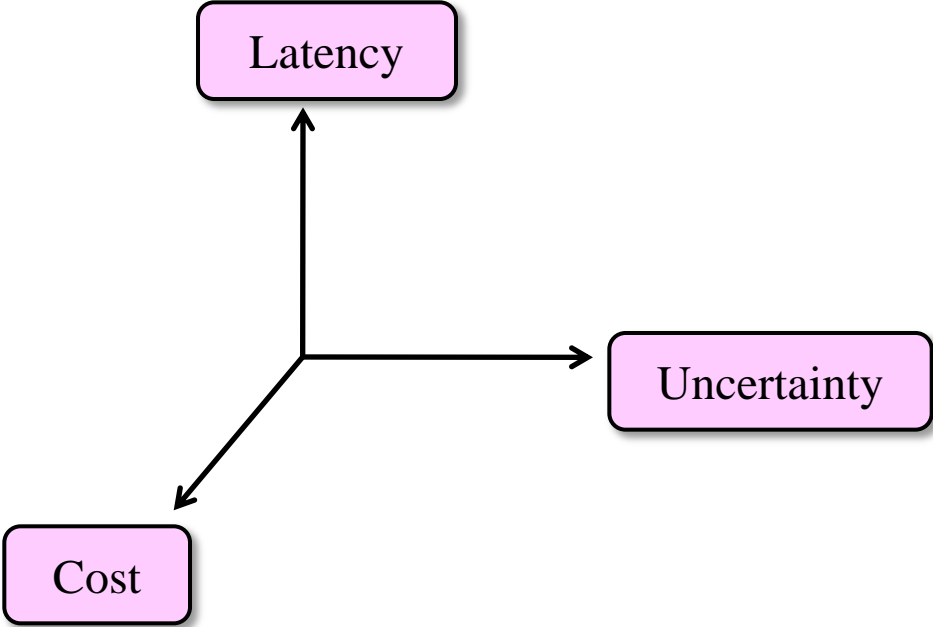
Machine Learning

Information Management

Crowd Information Management

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

Fundamental Tradeoffs



Efficiency: Fundamental Tradeoffs

How long can I wait?

Latency

Uncertainty

What is the desired quality?

Cost

How much \$\$ can I spend?

Efficiency: Fundamental Tradeoffs

How long can I wait?

Latency

- 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?

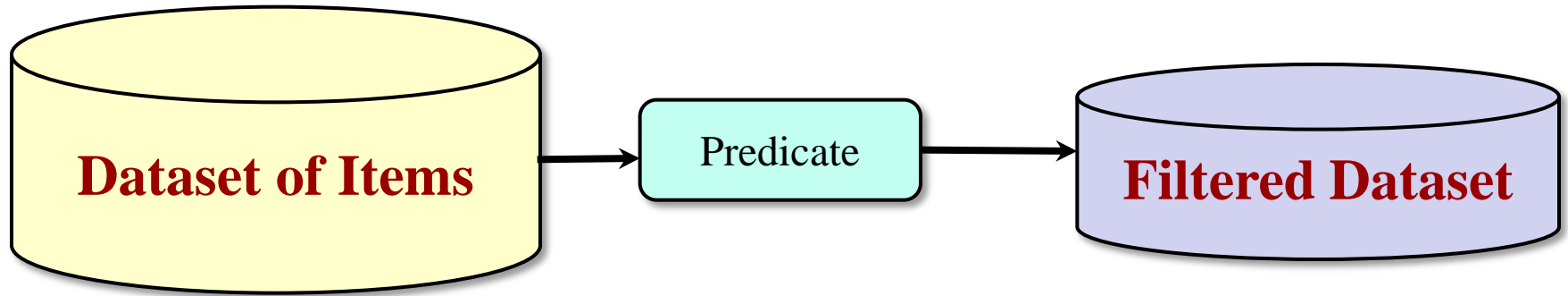
Uncertainty

What is the desired quality?

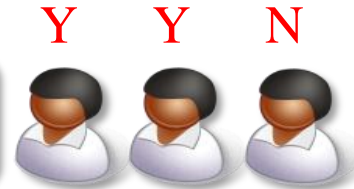
Cost

How much \$\$ can I spend?

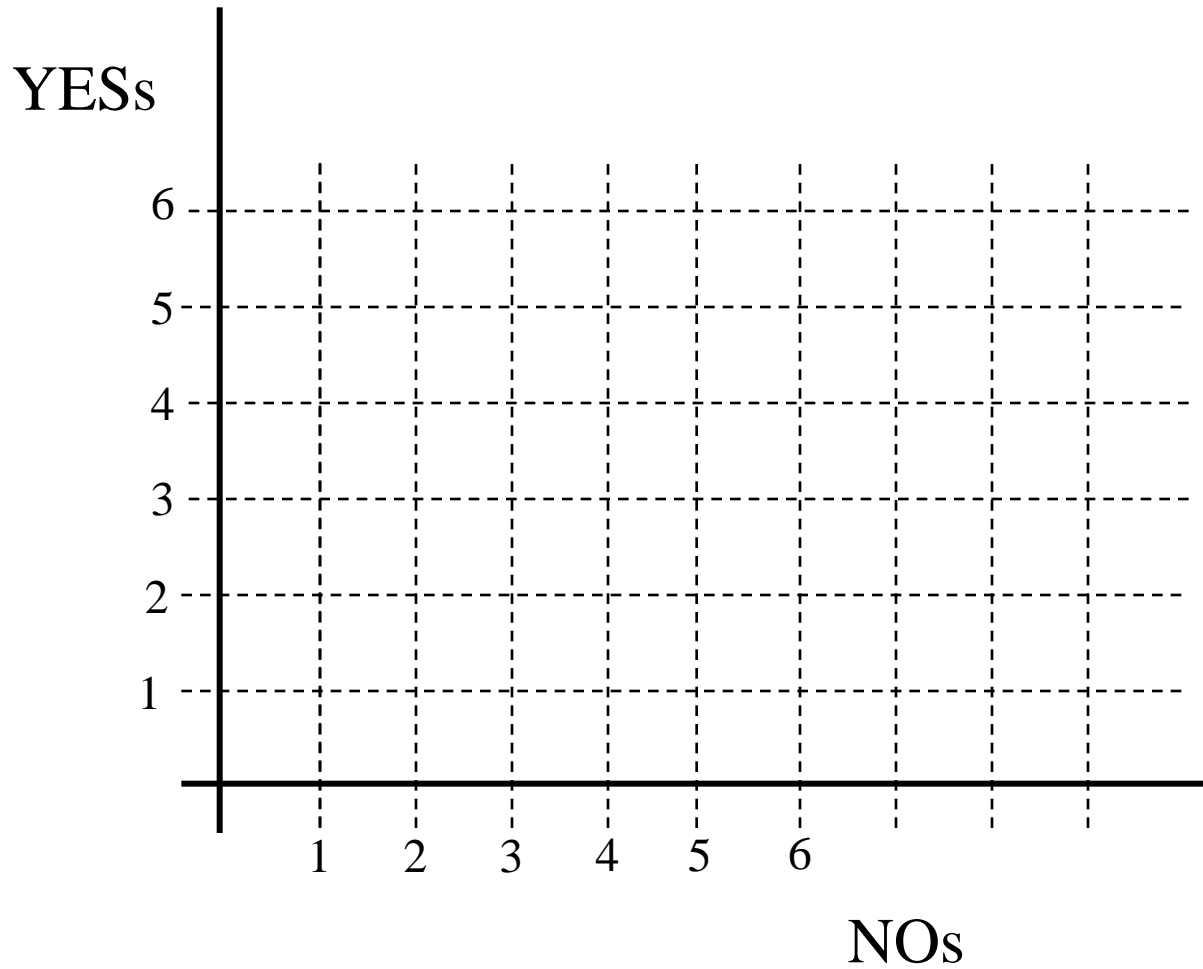
Example: CrowdScreen



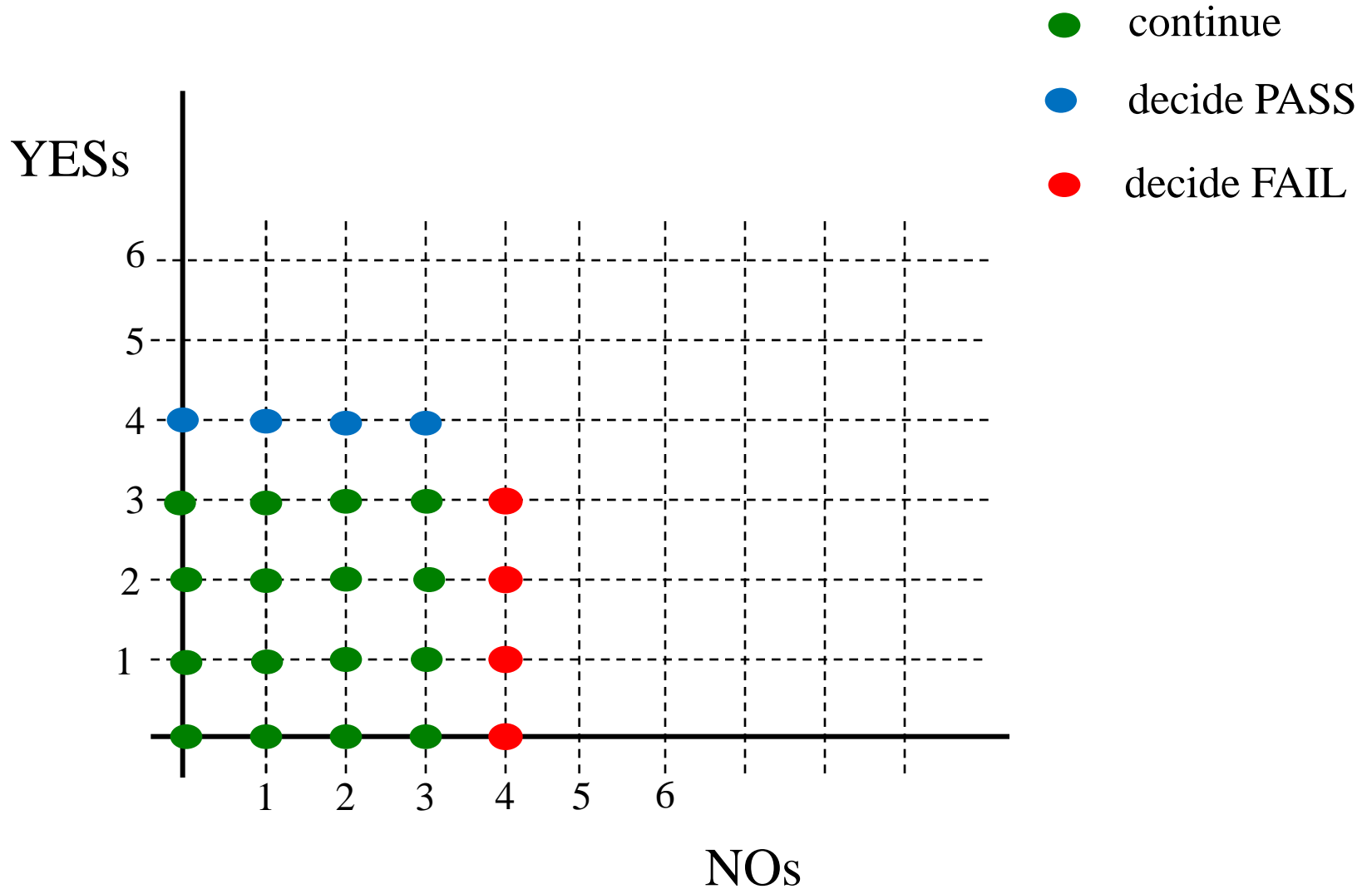
Item X satisfies predicate?



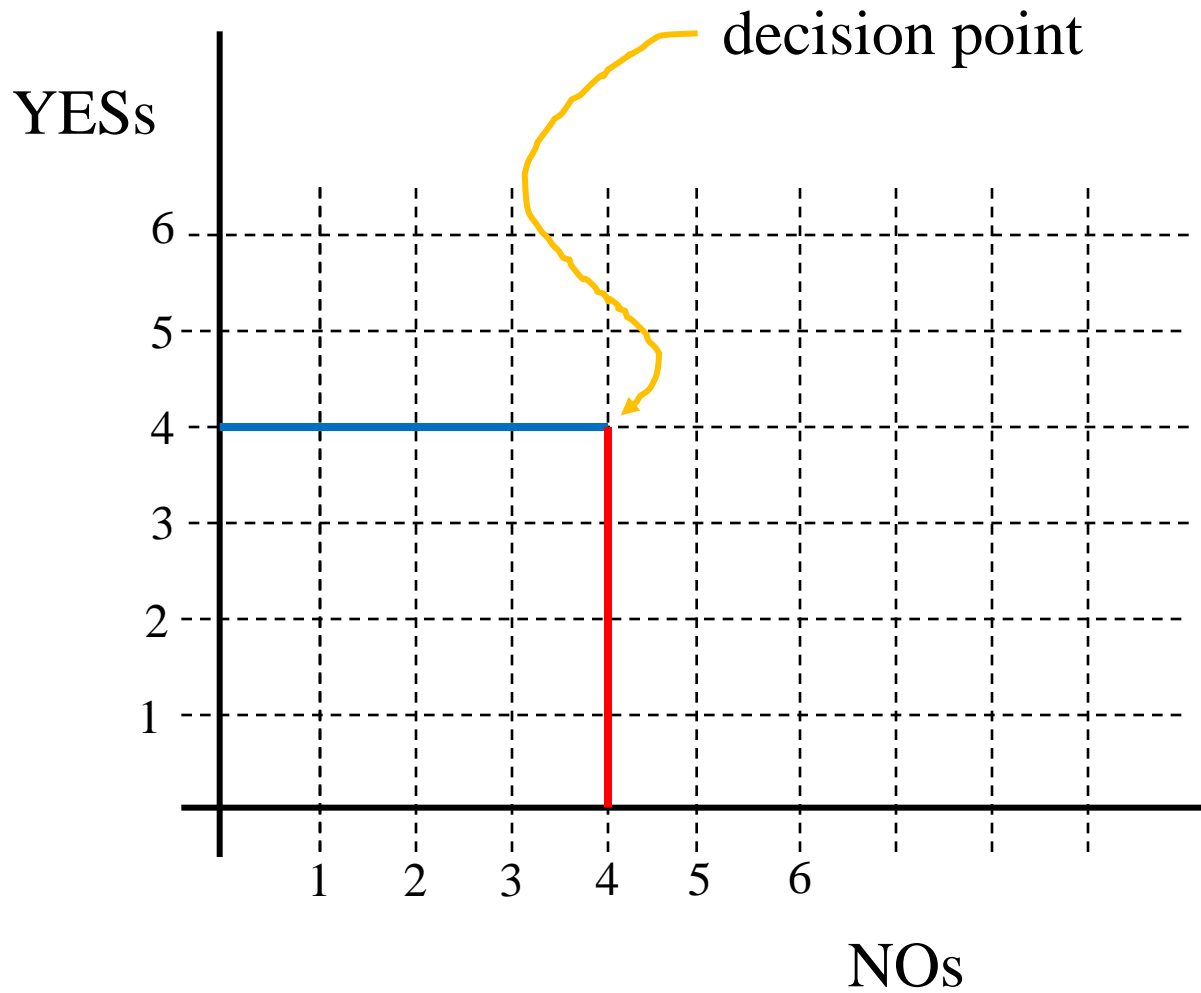
Strategy



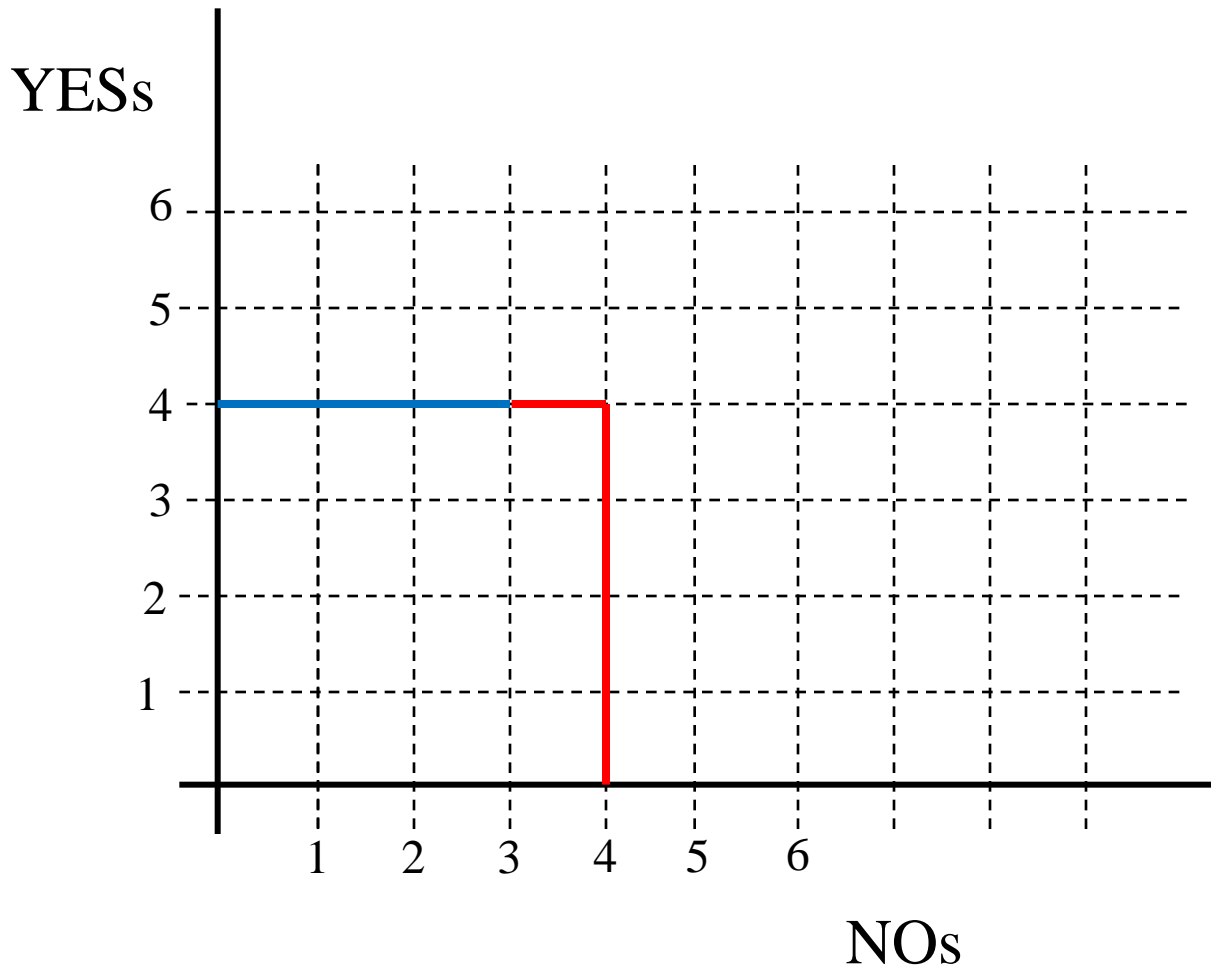
Strategy



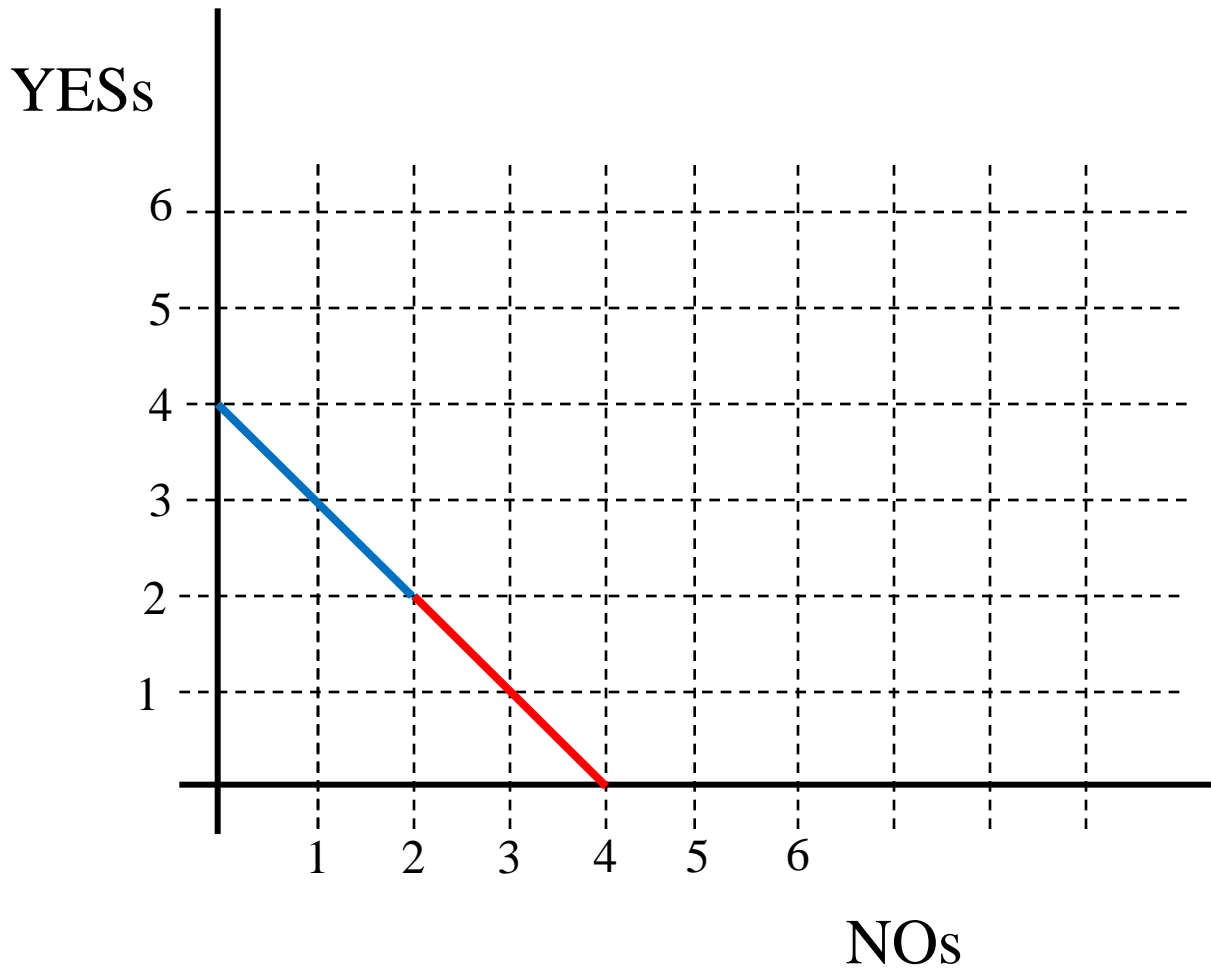
Strategy



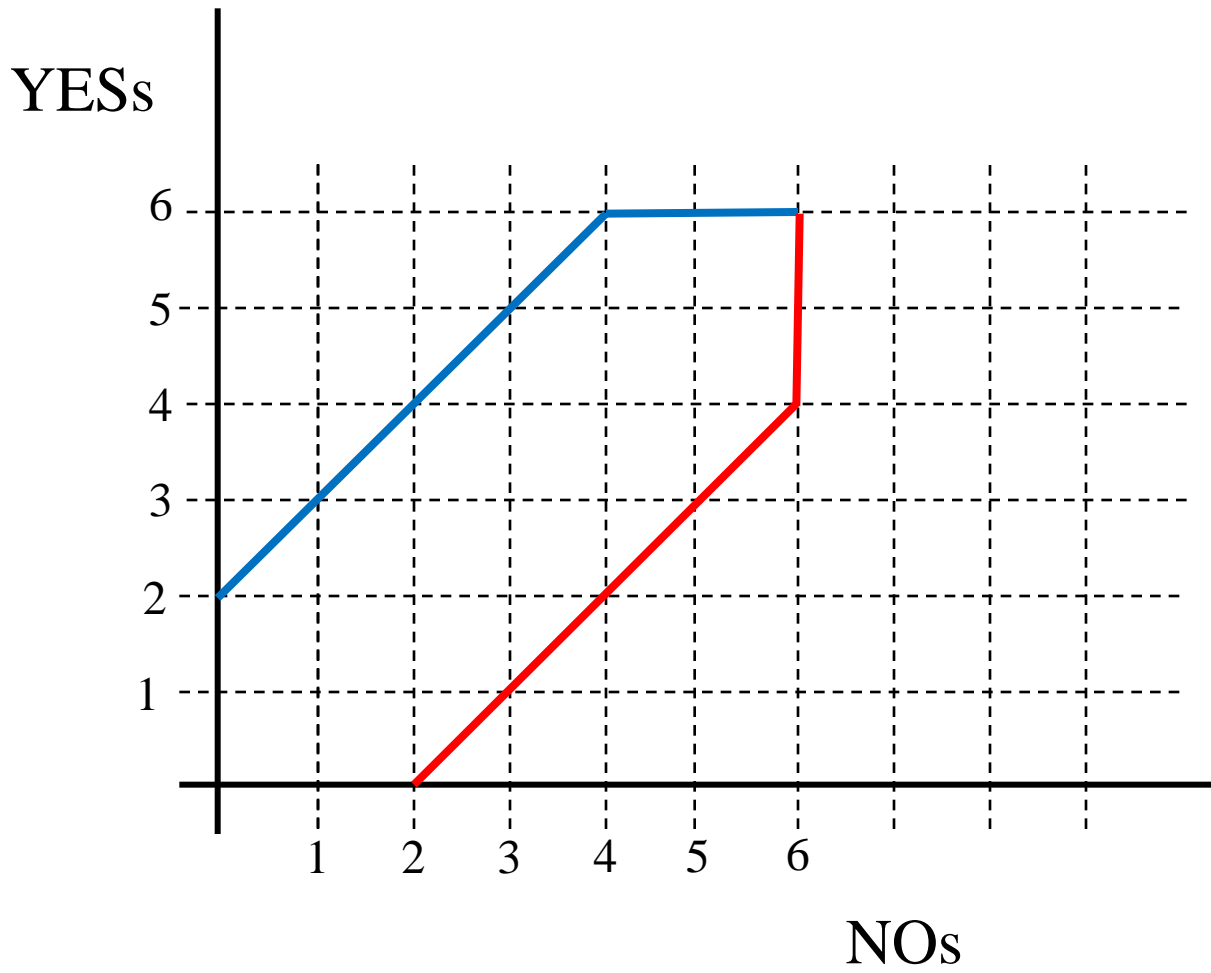
More Examples



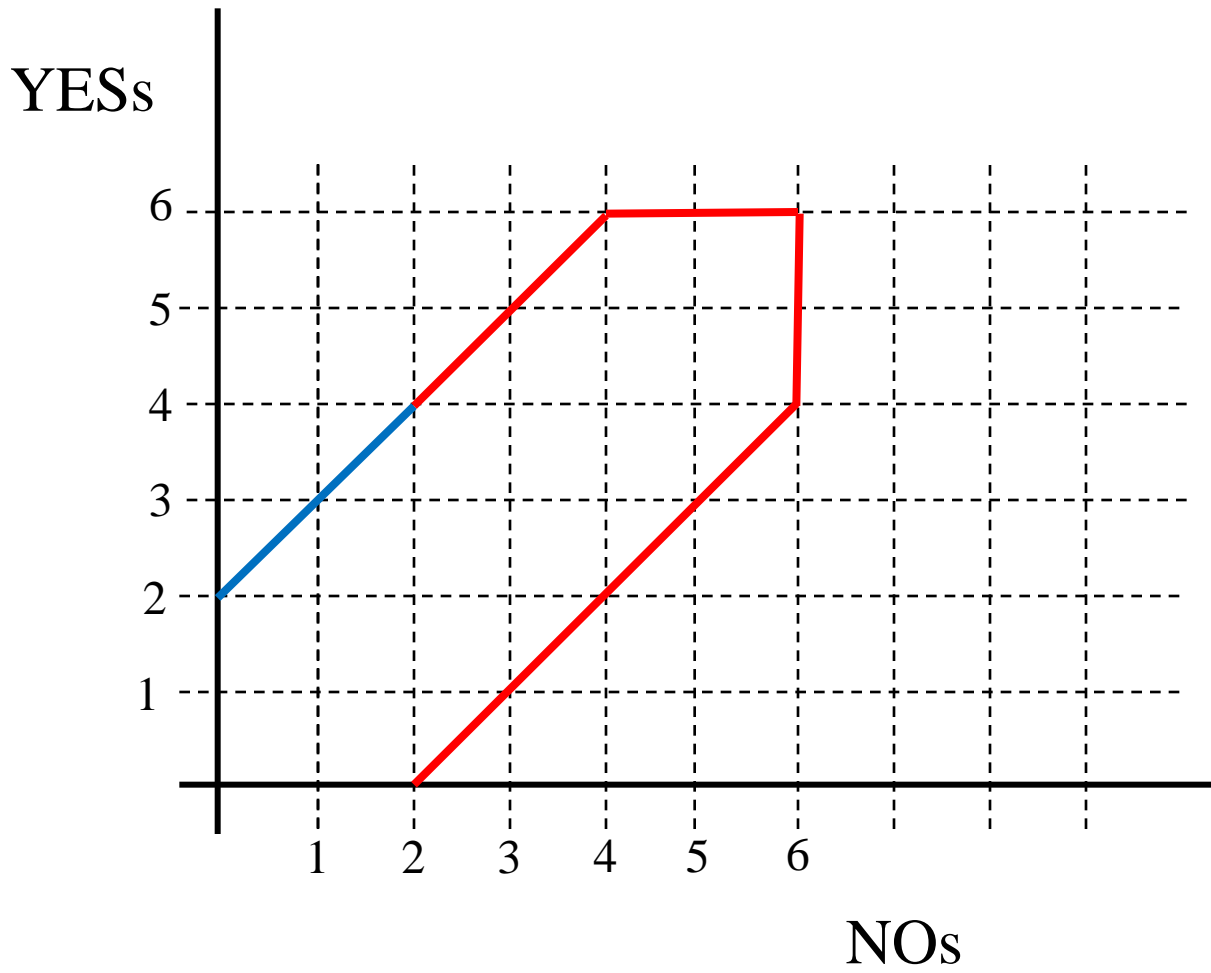
More Examples



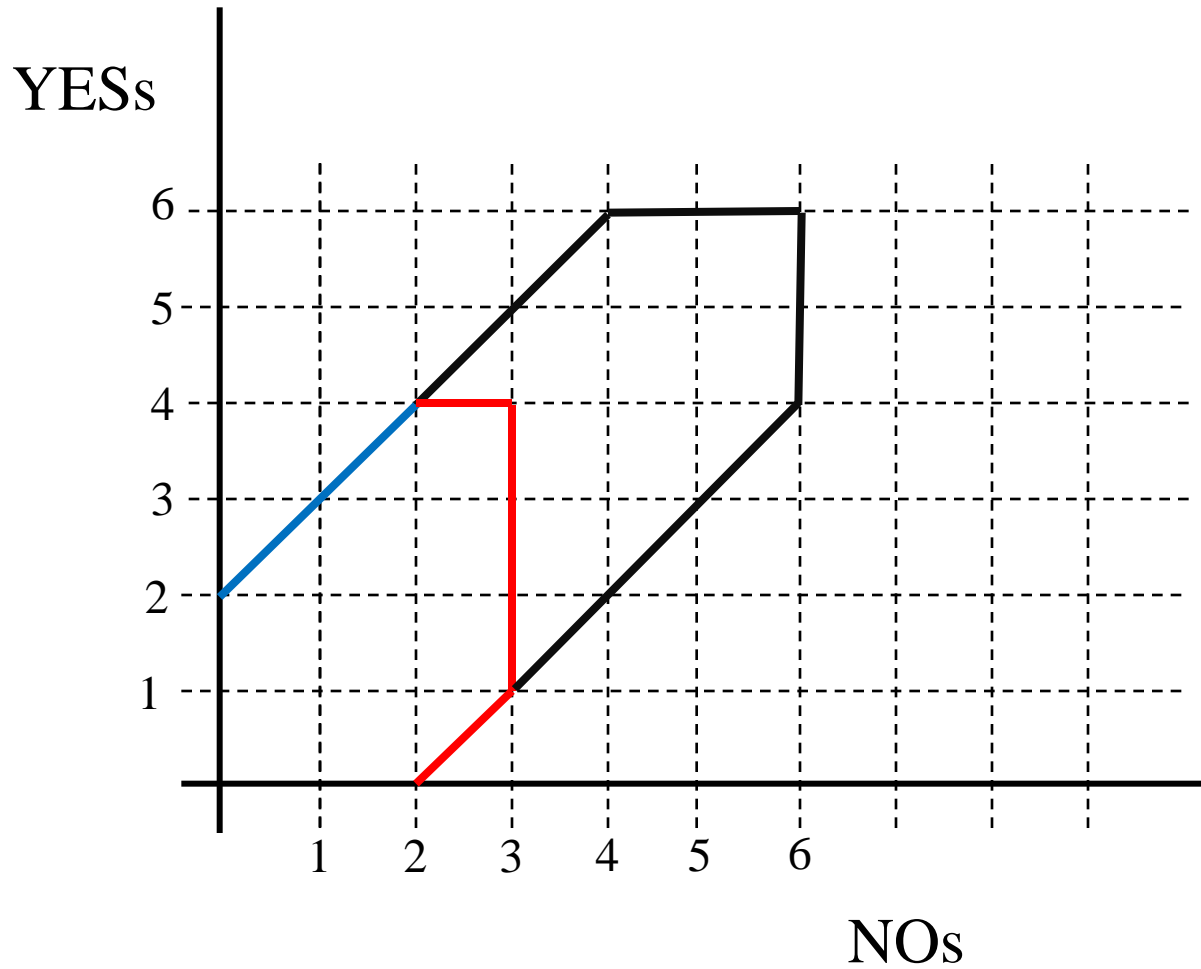
More Examples



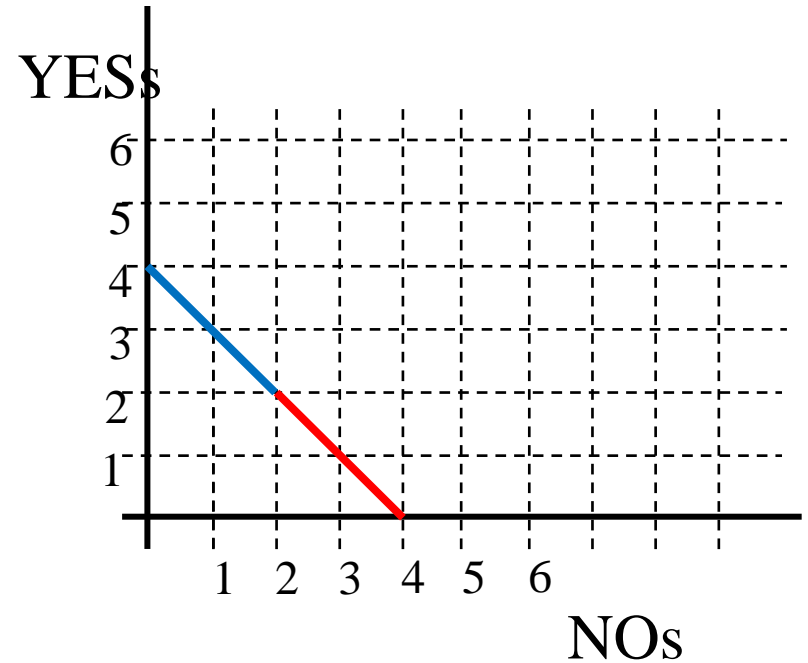
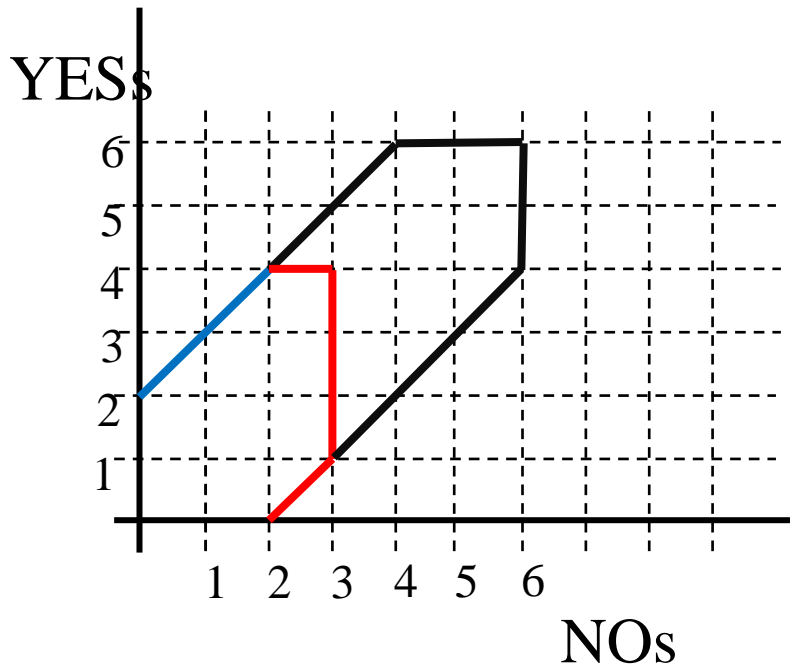
More Examples



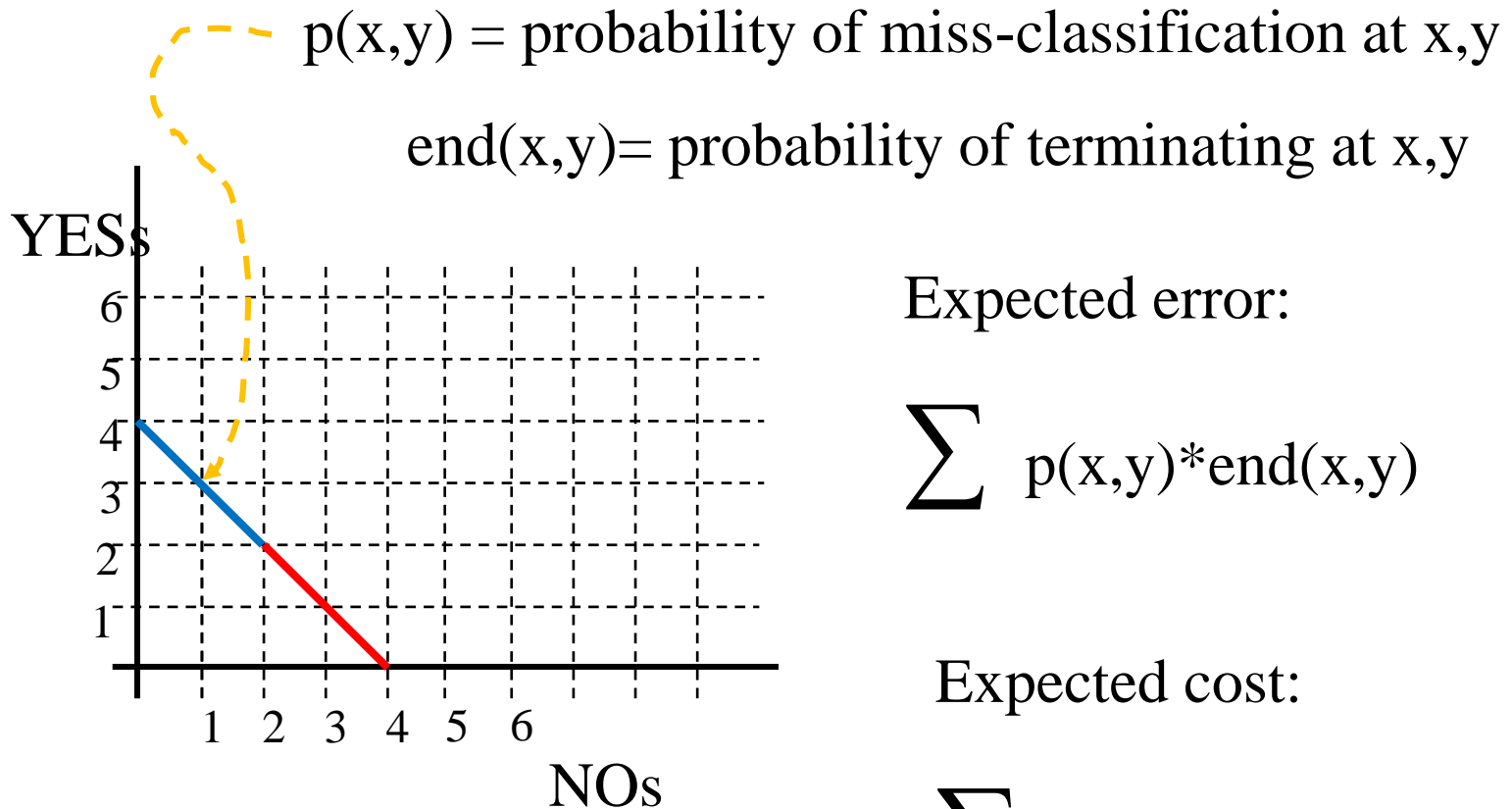
Some Optimizations



What is “best” strategy?



What is “best” strategy?



Expected error:

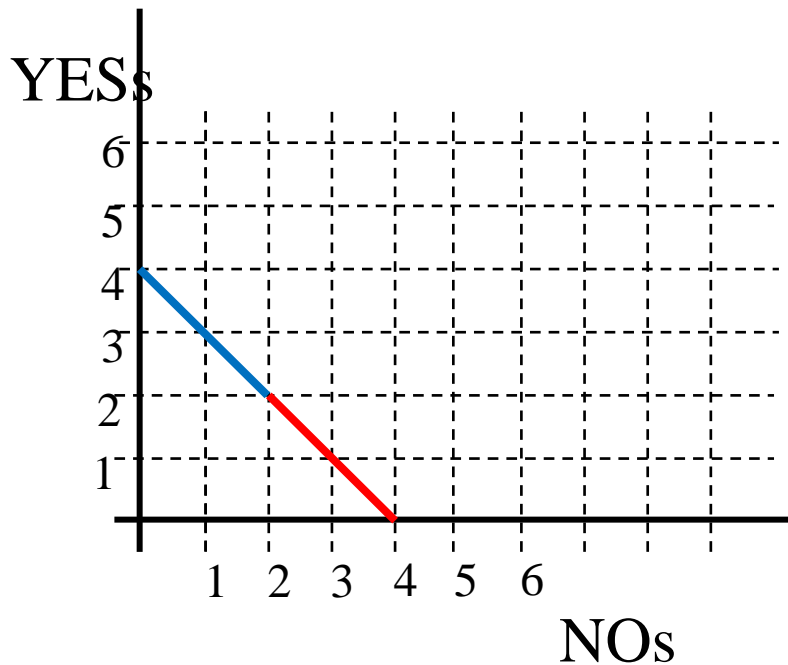
$$\sum p(x,y)*end(x,y)$$

Expected cost:

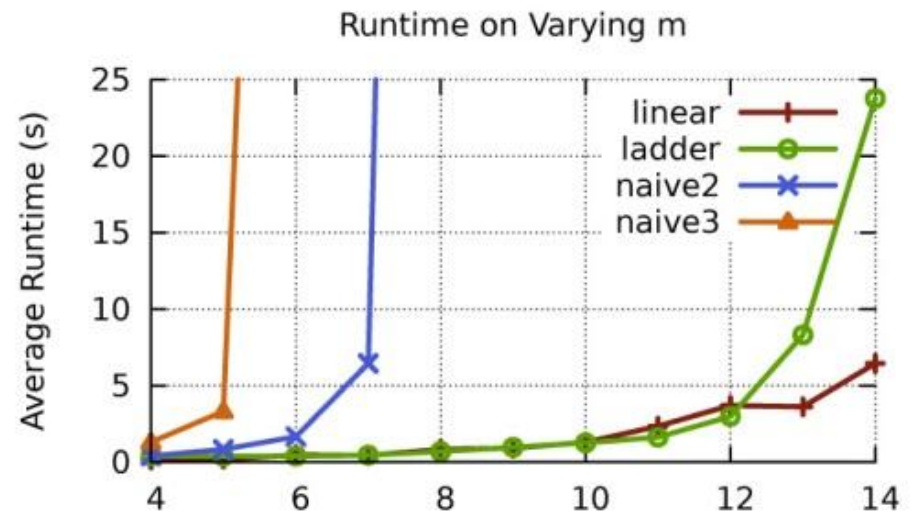
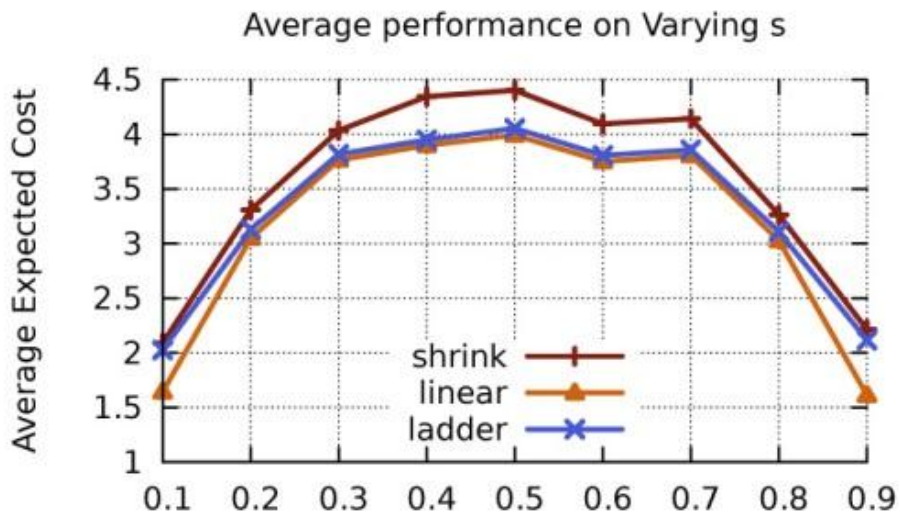
$$\sum (x+y)*end(x,y)$$

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)

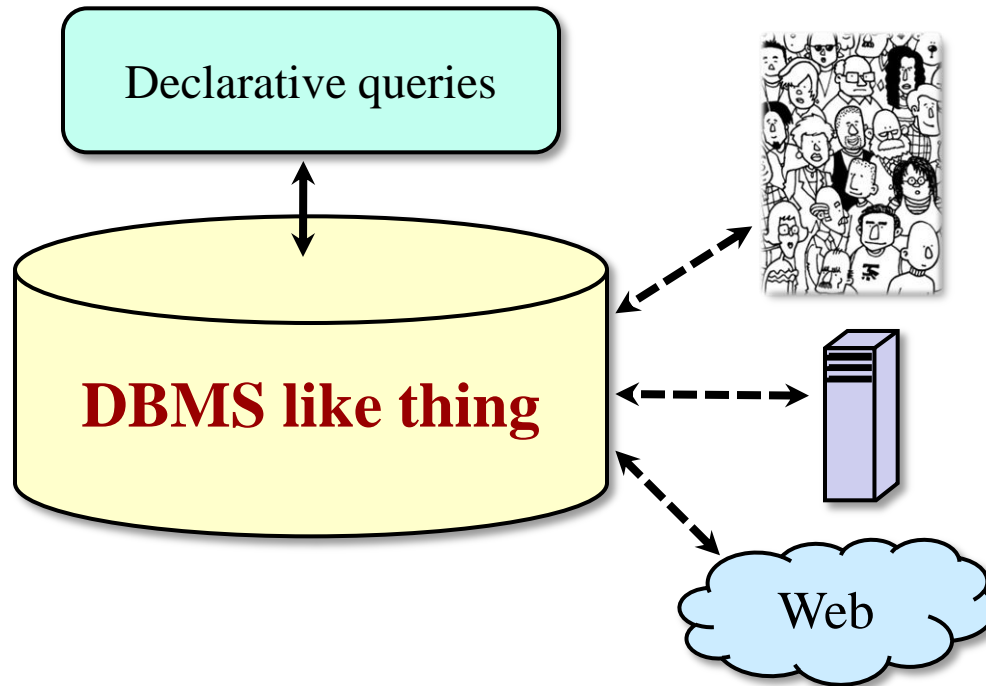
Beyond Filtering

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

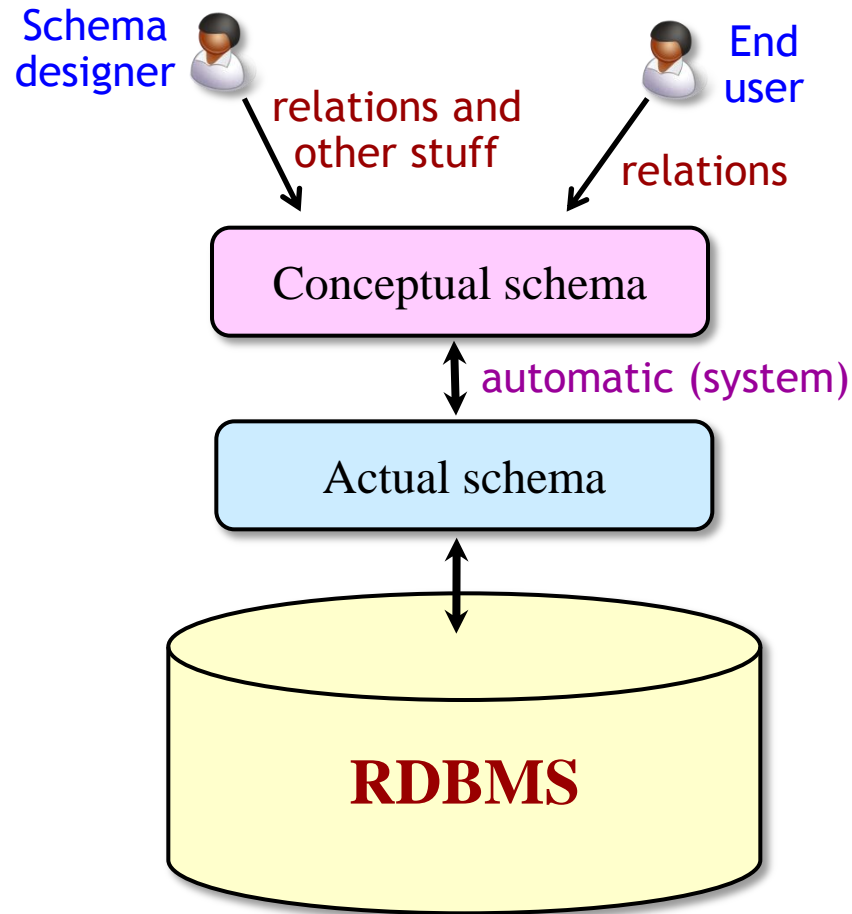
Beyond Simple Models

- 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



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



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



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



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	...

Dependent

fetch rule

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	...
...	...

Dependent

fetch rule

Small Example



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



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	...

Dependent

fetch rule

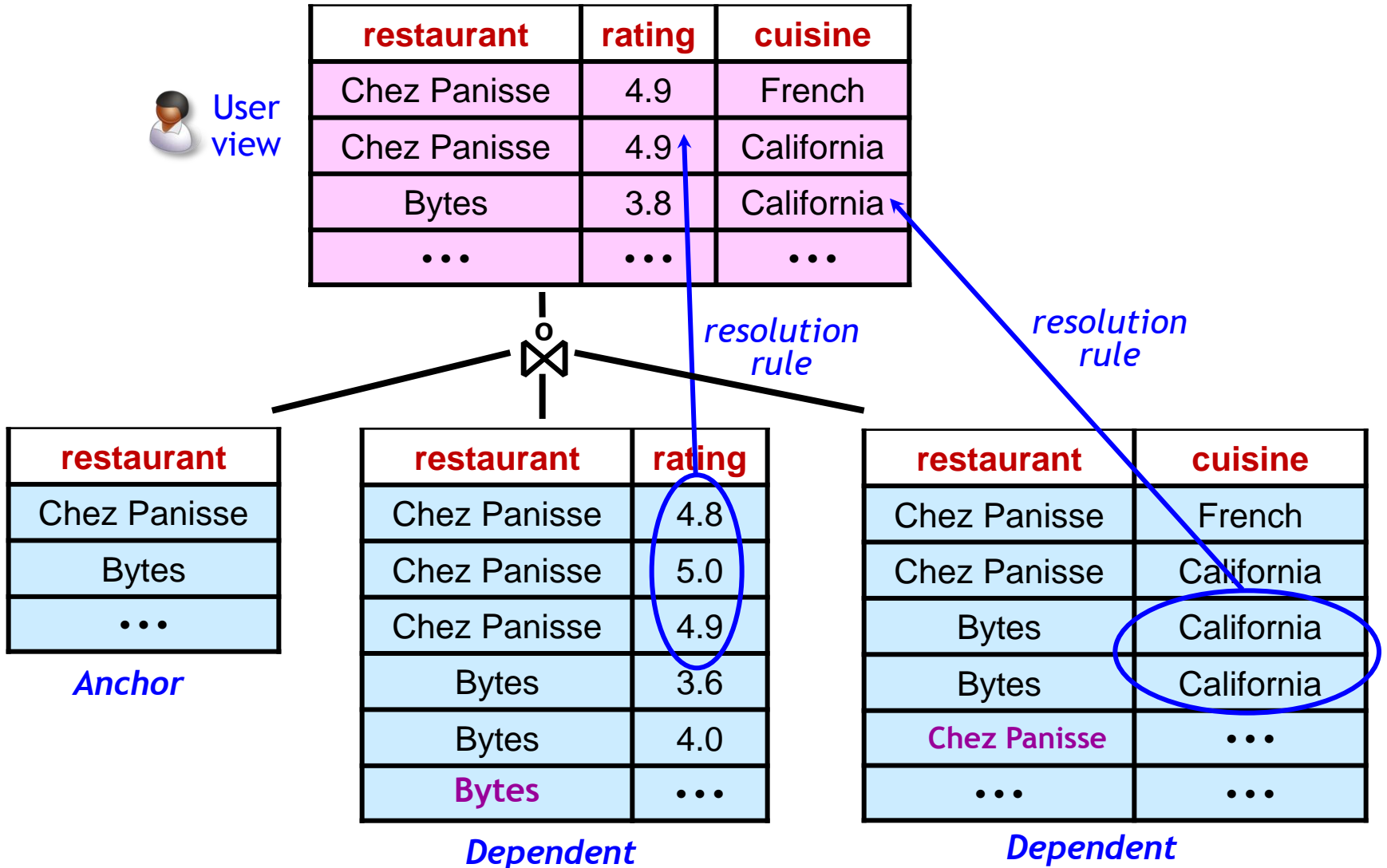
restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	...
...	French

Dependent

fetch rule

fetch rule

Small Example



Small Example



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
...	...

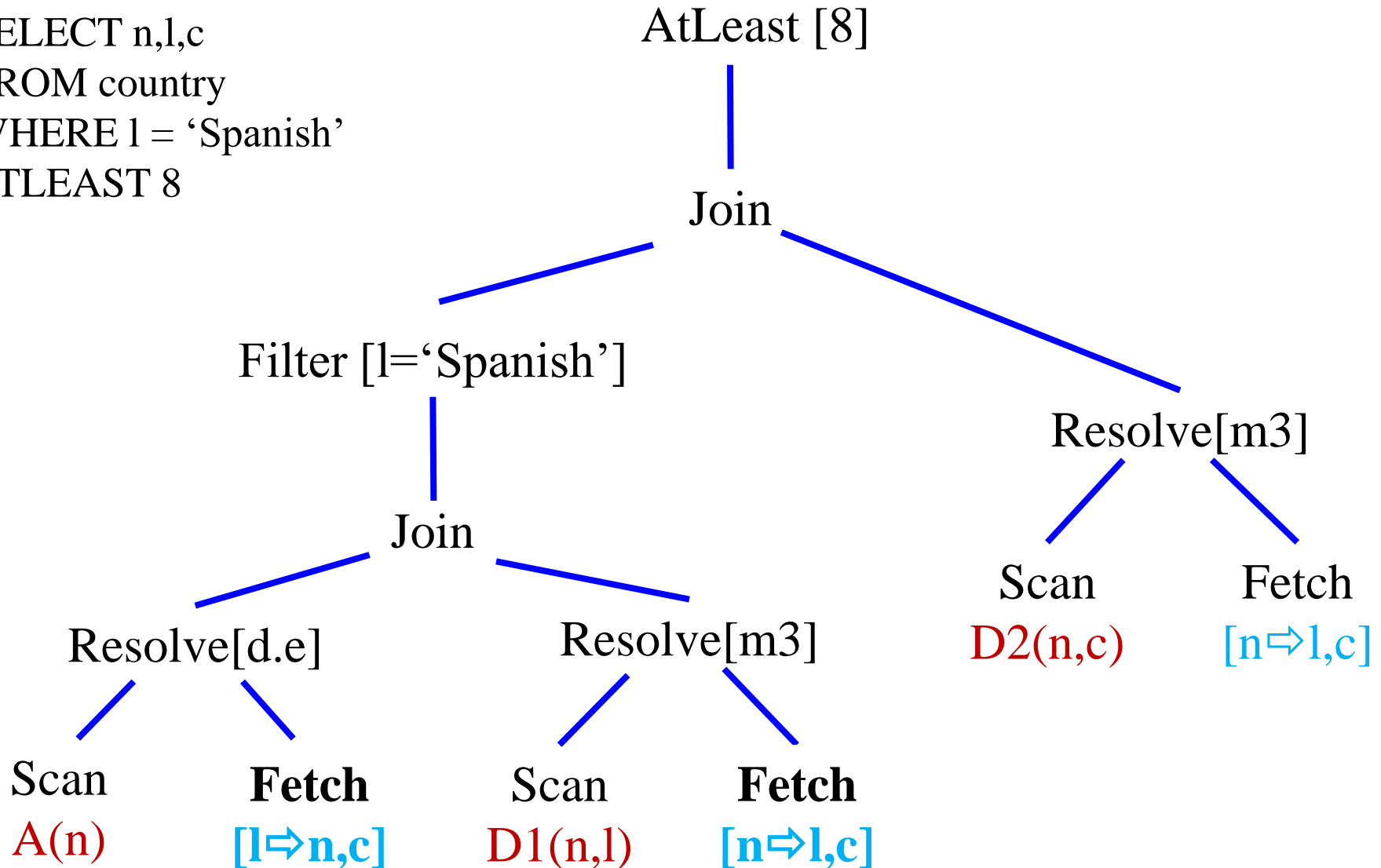
Dependent

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

The screenshot shows a web browser window with the URL `localhost:8080`. The page title is "Deco: A System for Declarative Crowdsourcing". In the top right corner, there is a "Stanford InfoLab" logo. Below the title, there are two buttons: "HOME" and "SIGN OUT".

The main content area features a text input field containing the SQL query: `SELECT * FROM rest WHERE cuisine='French' ATLEAST 10`. Below the input field are three buttons: "Clear", "Explain", and "Execute".

The "Explain" button has been clicked, resulting in a query execution plan diagram. The plan is a tree structure of nodes, each representing a step in the query execution. The nodes are:

- AtLeast-15 (10)
- Project-14
- Filter-13
- DepJoin-12 (left outer)
- DepJoin-7 (left outer)
- Resolve-11 (majority3)
- Filter-3
- Resolve-6 (majority3)
- Scan-9 (rest\$2)
- Fetch-10 (rest2)
- Resolve-2 (identity)
- Scan-4 (rest\$1)
- Fetch-5 (rest1)
- Scan-0 (rest\$0)
- Fetch-1 (rest0)

The diagram shows a top-down flow from the "AtLeast-15" node to the leaf nodes "Scan-0" and "Fetch-1".

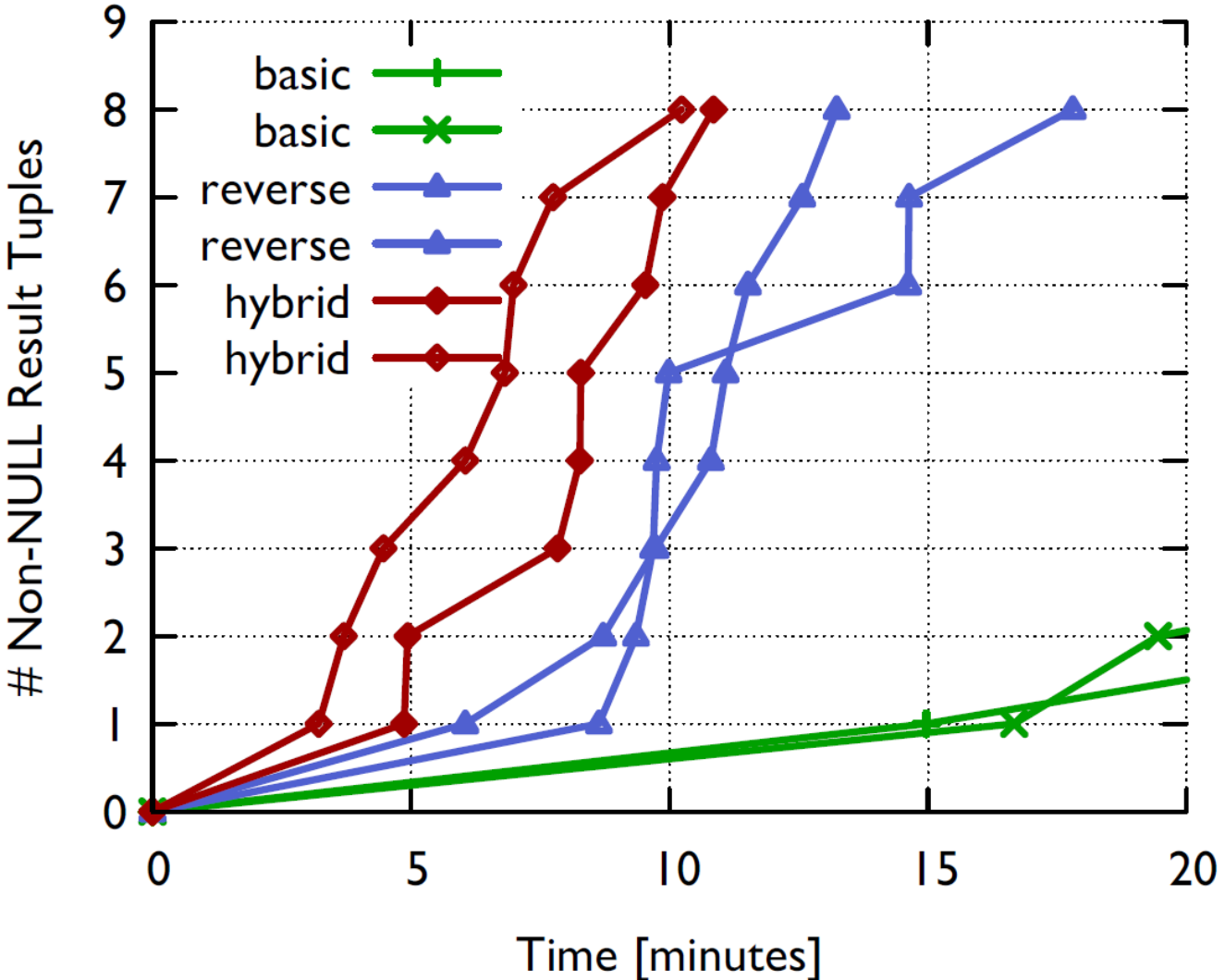
At the bottom of the page, there is a footer: "Stanford University InfoLab © 2012".

Experimental Setup

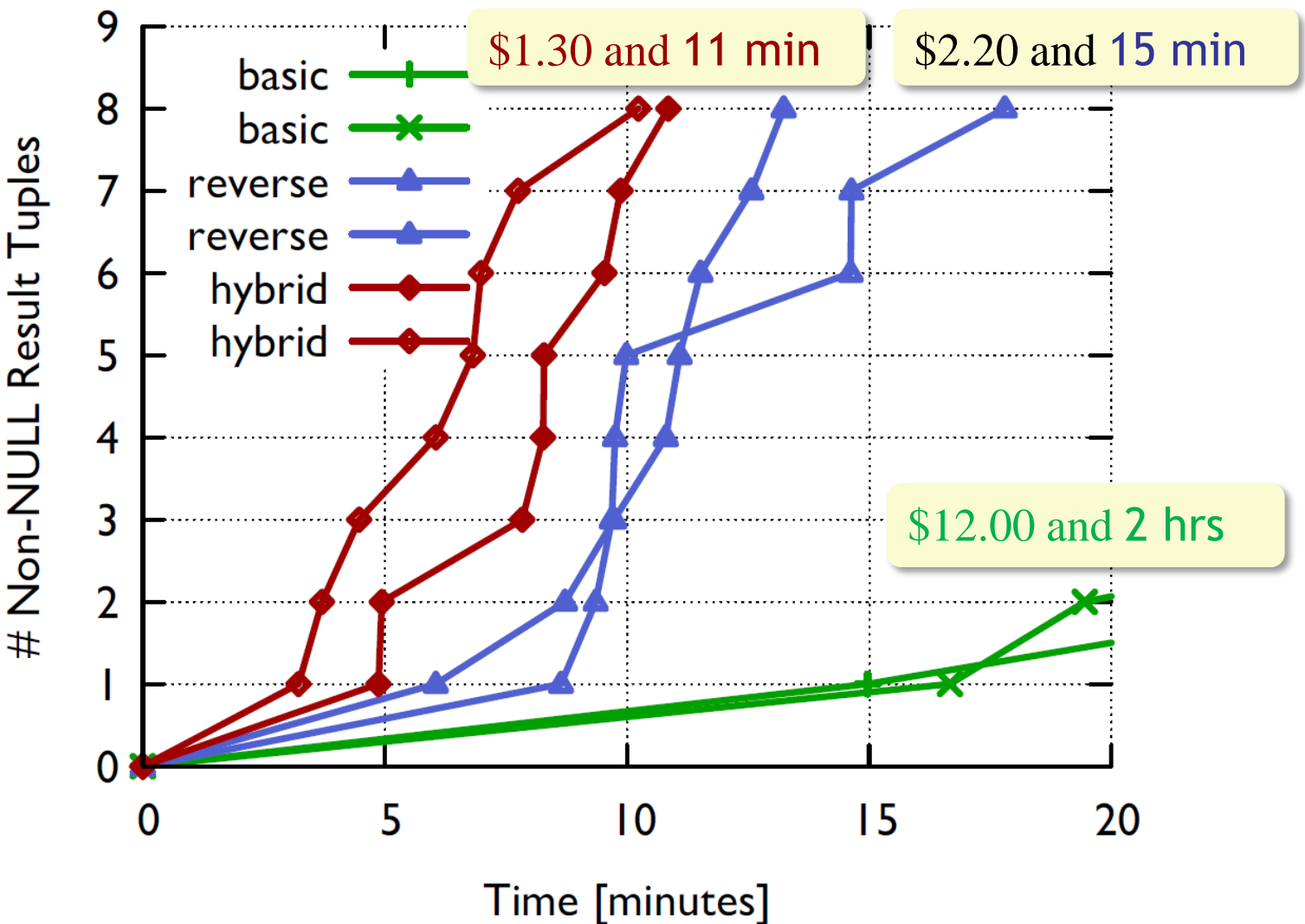
```
SELECT n,l,c FROM country WHERE l = 'Spanish' ATLEAST 8
```

- Experimental Goals:
 - Different fetch configurations
 - Basic: $\emptyset \Rightarrow n + n \Rightarrow l + n \Rightarrow c$
 - 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!

