Using CrowdSourcing for Data Analytics

Hector Garcia-Molina

(work with Steven Whang, Peter Lofgren, Aditya Parameswaran and others)

Stanford University

- Big Data Analytics
- CrowdSourcing

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"

3

Real World Examples

4

Many Crowdsourcing Marketplaces!

uTest
Elance
CloudCrowd
GRABCAD
liveops
oDesk
mygengo
freelancer
amazonmechanical turk
CrowdFlower
agentanything
crowdcontent
socialvibe
DeskLancer
servio
CTR Boost
DESIGN RIDER
CAPSEO
samasource
SpeakerText
Truveli
trada
guru
microtask
task us
DoNanza
Lingotek
expertplanet
MOB4HIRE
peopleperhour.com
clickworker
PEER TO PATENT
agasauris
the smart way to tag!
mediapiston

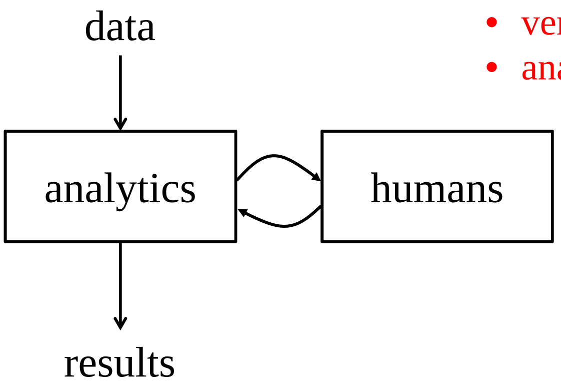
Many Research Projects!

W
UNIVERSITY of WASHINGTON
Carnegie Mellon University
NYU
I
Cal
HARVARD
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Universität Bremen
Microsoft Research
YAHOO! RESEARCH
MIT
Massachusetts Institute of Technology
SYRACUSE UNIVERSITY
THE UNIVERSITY OF TEXAS AT AUSTIN

6

Example tasks:

- get missing data
- verify results
- analyze data



data

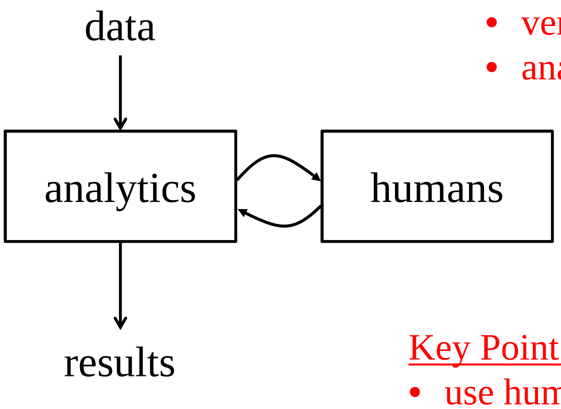
analytics

humans

results

Example tasks:

- get missing data
- verify results
- analyze data



data

analytics

humans

results

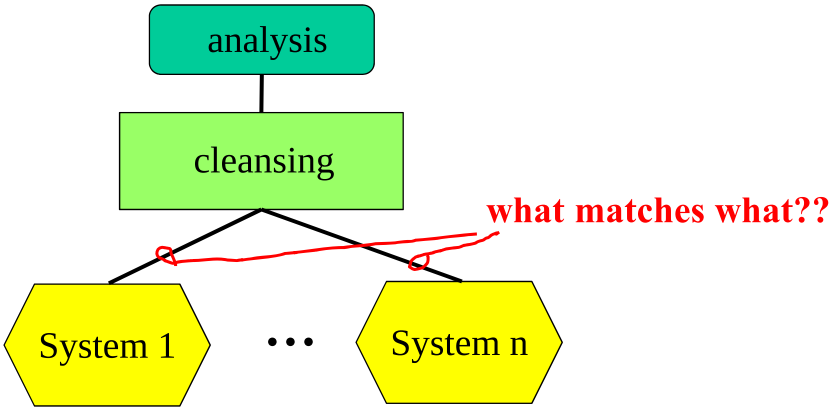
Key Point:

- use humans judiciously

Today will illustrate with

- Entity Resolution
- (may cover another topic briefly)

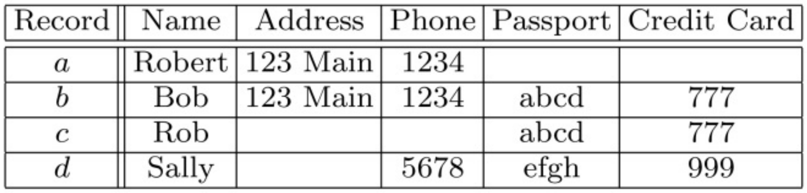
Traditional Entity Resolution



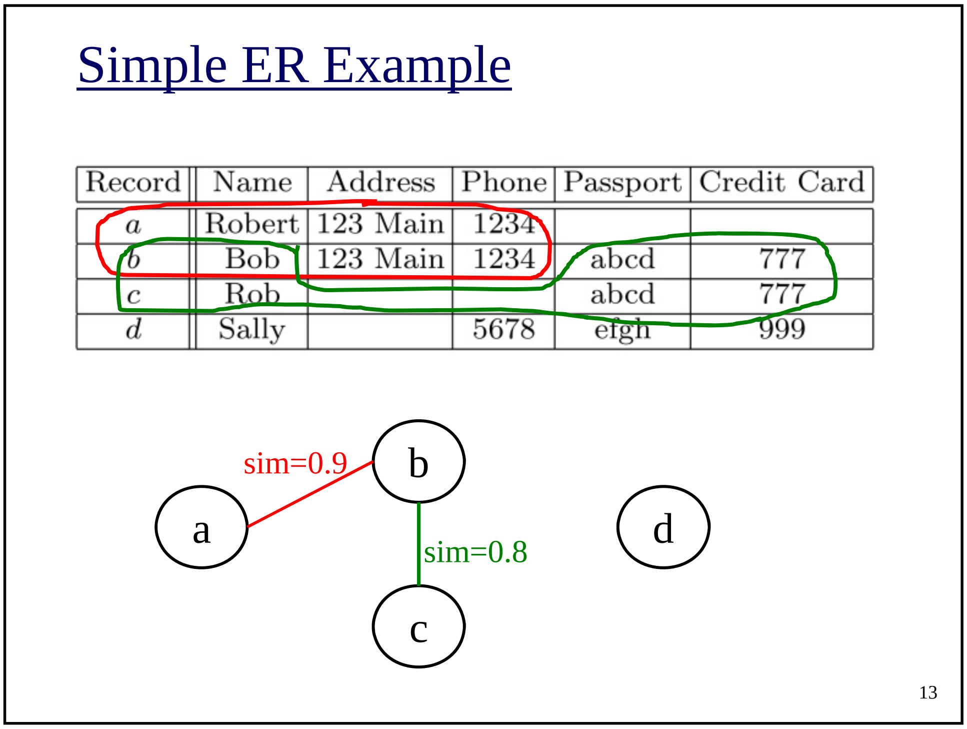
Why is ER Challenging?

- Huge data sets
- No unique identifiers
- Missing data
- Lots of uncertainty
- Many ways to skin the cat

Simple ER Example

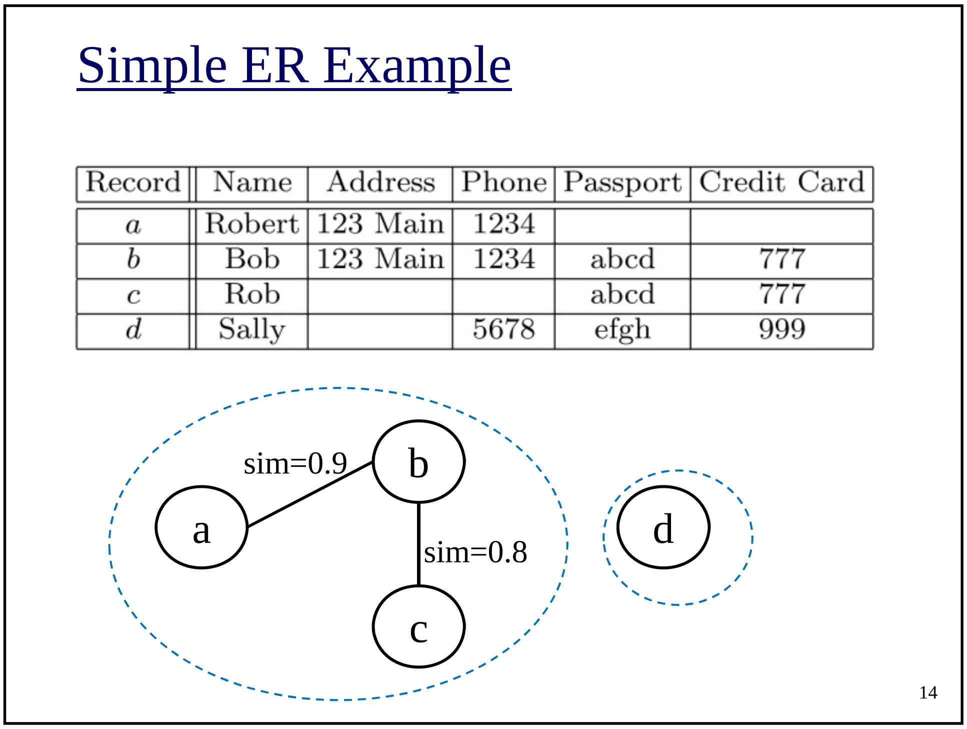


Record	Name	Address	Phone	Passport	Credit Card
a	Robert	123 Main	1234		
b	Bob	123 Main	1234	abcd	777
c	Rob			abcd	777
d	Sally		5678	efgh	999



Simple ER Example

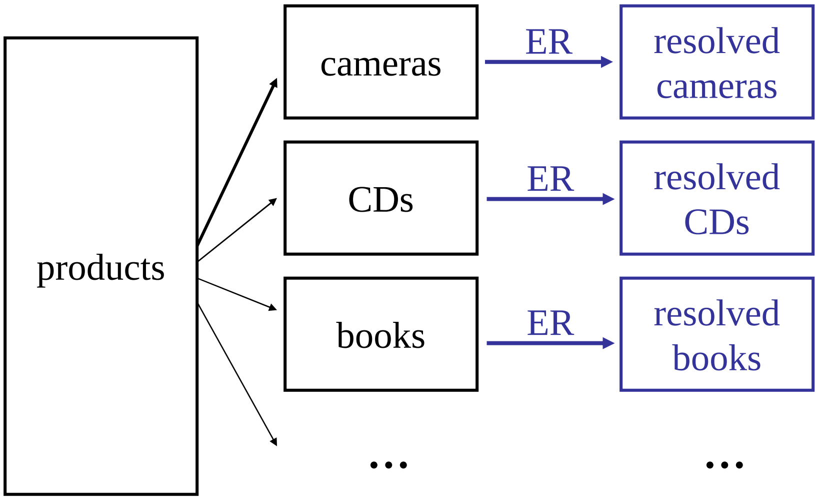
Record	Name	Address	Phone	Passport	Credit Card
a	Robert	123 Main	1234		
b	Bob	123 Main	1234	abcd	777
c	Rob			abcd	777
d	Sally		5678	efgh	999



Simple ER Example

Record	Name	Address	Phone	Passport	Credit Card
a	Robert	123 Main	1234		
b	Bob	123 Main	1234	abcd	777
c	Rob			abcd	777
d	Sally		5678	efgh	999

ER: Exact vs Approximate



products

cameras

CDs

books

…

ER

ER

ER

resolved cameras

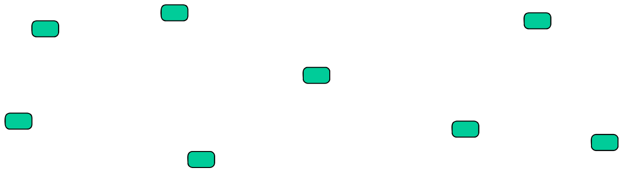
resolved CDs

resolved books

…

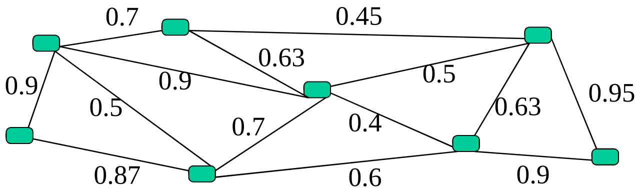
Simple ER Algorithm

- Compute pairwise similarities
- Apply threshold
- Perform transitive closure



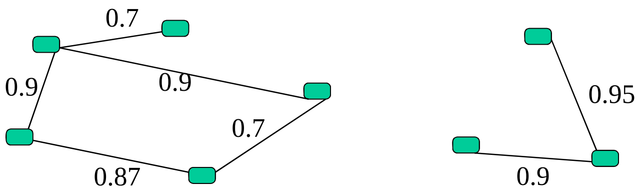
Simple ER Algorithm

- Compute pairwise similarities
- Apply threshold
- Perform transitive closure



Simple ER Algorithm

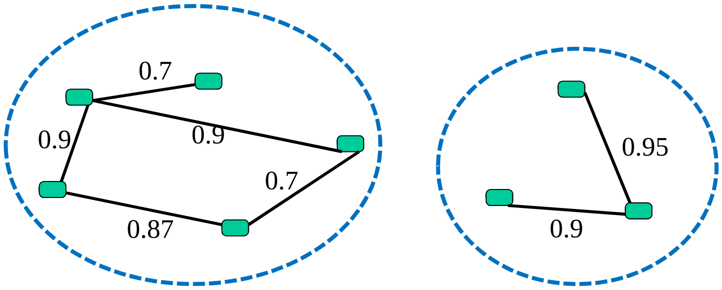
- Compute pairwise similarities
- Apply threshold
- Perform transitive closure



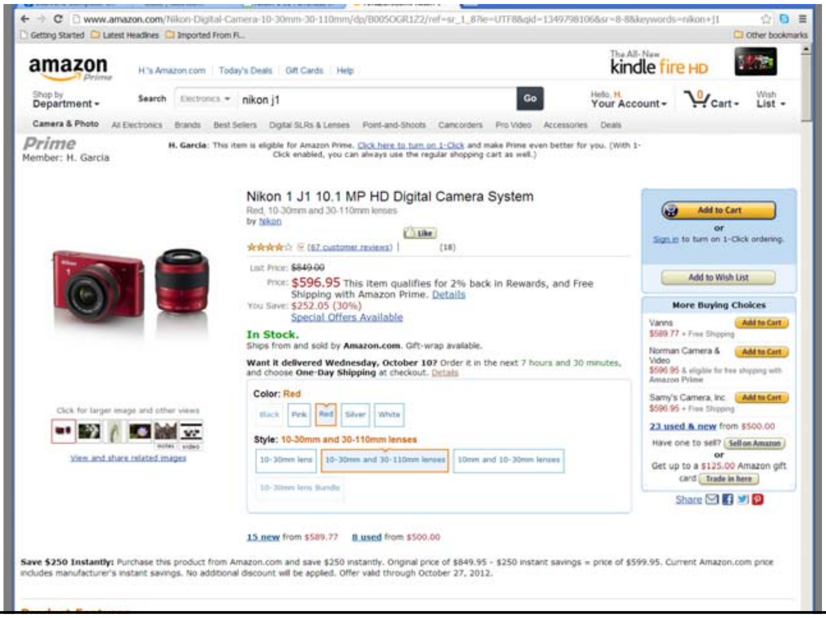
threshold = 0.7

Simple ER Algorithm

- Compute pairwise similarities
- Apply threshold
- Perform transitive closure



Crowd ER



H. Garcia: This item is eligible for Amazon Prime. Click here to turn on 1-Click and make Prime even better for you. (With 1-Click enabled, you can always use the regular shopping cart as well.)

Nikon 1 J1 10.1 MP HD Digital Camera System
Red, 10-30mm and 30-110mm lenses
by Nikon

★★★★☆ (67 customer reviews) | (18)

List Price: ~~$849.00~~
Price: $596.95 This item qualifies for 2% back in Rewards, and Free Shipping with Amazon Prime. Details
You Save: $252.05 (30%)
Special Offers Available

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Wednesday, October 10? Order it in the next 7 hours and 30 minutes, and choose **One-Day Shipping** at checkout. Details

Color: Red
Black | Pink | Red | Silver | White

Style: 10-30mm and 30-110mm lenses
10-30mm lens | 10-30mm and 30-110mm lenses | 10mm and 10-30mm lenses | 10-30mm lens Bundle

15 new from $589.77   8 used from $500.00

Add to Cart
or
Sign in to turn on 1-Click ordering.

More Buying Choices

Vanns
$589.77 + Free Shipping

Norman Camera & Video
$596.95 & eligible for free shipping with Amazon Prime

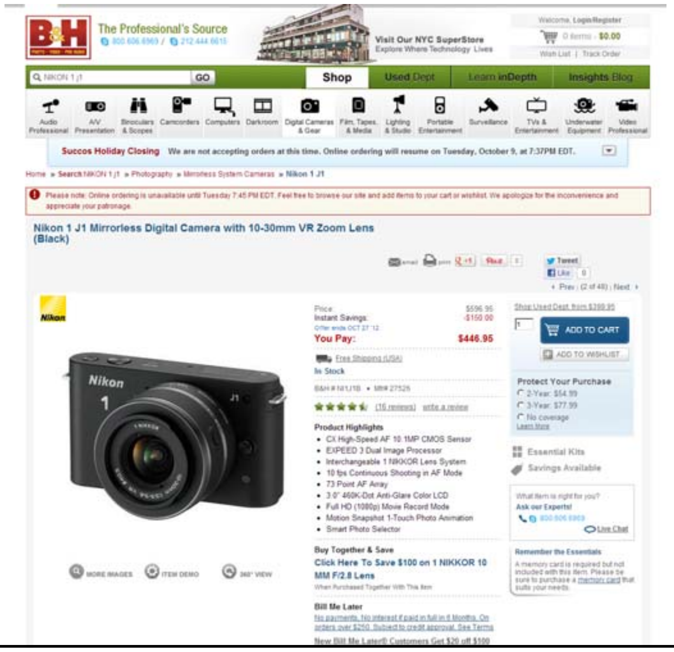
Samy's Camera, Inc
$596.95 + Free Shipping

23 used & new from $500.00

Have one to sell?
or
Get up to a $125.00 Amazon gift card

Save $250 Instantly: Purchase this product from Amazon.com and save $250 instantly. Original price of $849.95 - $250 instant savings = price of $599.95. Current Amazon.com price includes manufacturer's instant savings. No additional discount will be applied. Offer valid through October 27, 2012.

Same as this?



Succos Holiday Closing We are not accepting orders at this time. Online ordering will resume on Tuesday, October 9, at 7:37PM EDT.

Please note: Online ordering is unavailable until Tuesday 7:45 PM EDT. Feel free to browse our site and add items to your cart or wishlist. We apologize for the inconvenience and appreciate your patronage.

Nikon 1 J1 Mirrorless Digital Camera with 10-30mm VR Zoom Lens (Black)

Price:	$596.95
Instant Savings:	-$150.00
Offer ends OCT 27 '12	
You Pay:	$446.95

Free Shipping (USA)

In Stock

B&H # NI1J1B • MFR # 27526

★★★★½ (16 reviews) write a review

Product Highlights

- CX High-Speed AF 10.1MP CMOS Sensor
- EXPEED 3 Dual Image Processor
- Interchangeable 1 NIKKOR Lens System
- 10 fps Continuous Shooting in AF Mode
- 73 Point AF Array
- 3.0" 460K-Dot Anti-Glare Color LCD
- Full HD (1080p) Movie Record Mode
- Motion Snapshot 1-Touch Photo Animation
- Smart Photo Selector

Buy Together & Save

Click Here To Save $100 on 1 NIKKOR 10 MM F/2.8 Lens

When Purchased Together With This Item

Bill Me Later

No payments. No interest if paid in full in 6 Months. On orders over $250. Subject to credit approval. See Terms

New Bill Me Later® Customers Get $20 off $100

Shop Used Dept. from $399.95

1

Protect Your Purchase

○ 2-Year: $54.99
○ 3-Year: $77.99
○ No coverage

Learn More

What item is right for you?

Ask our Experts!

800 606 6969

Remember the Essentials

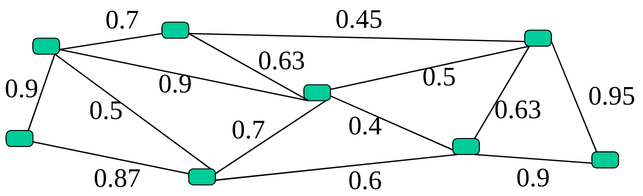
A memory card is required but not included with this item. Please be sure to purchase a memory card that suits your needs.

Crowd ER

- First Cut: For every pair of records, ask workers if they match (i.e., get similarity)

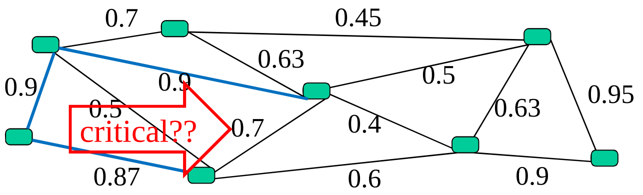
Crowd ER

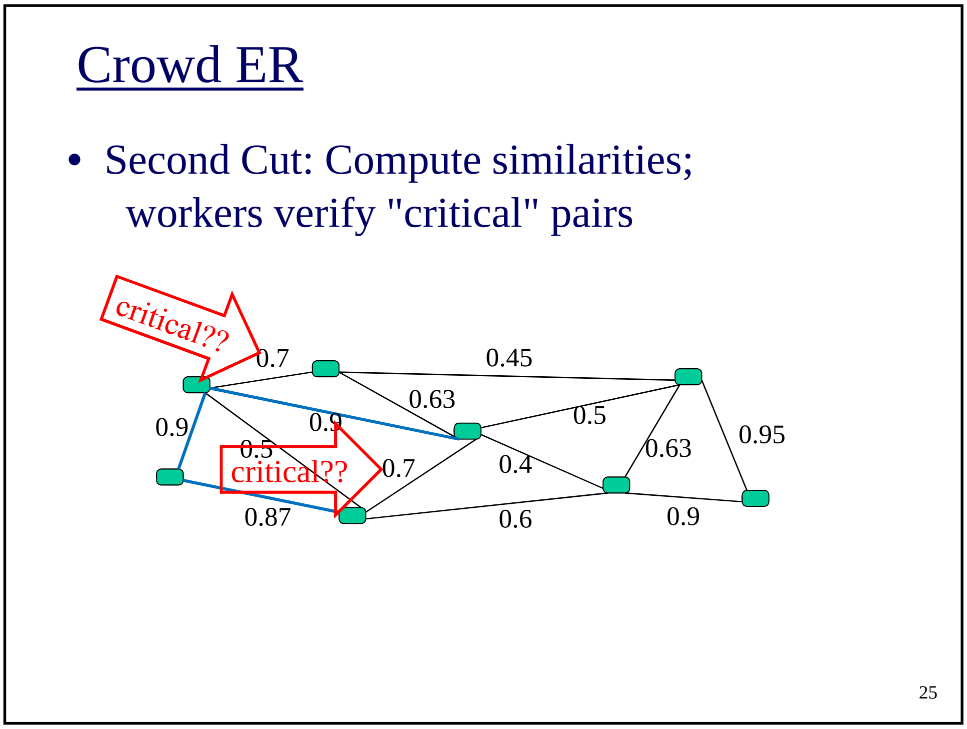
- First Cut: For every pair of records, ask workers if they match (i.e., get similarity)
- Too expensive!



Crowd ER

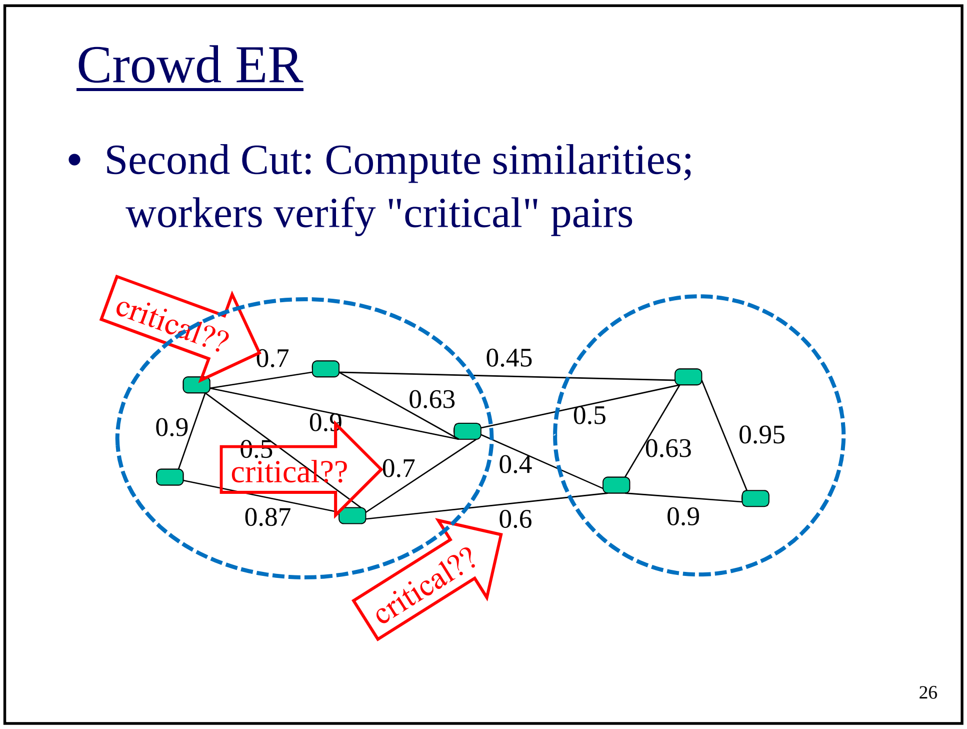
- Second Cut: Compute similarities; workers verify "critical" pairs





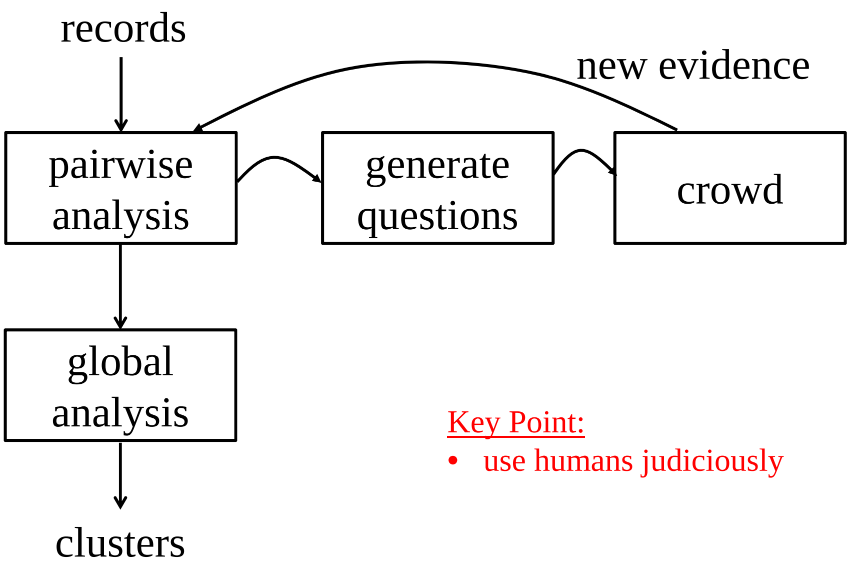
Crowd ER

- Second Cut: Compute similarities; workers verify "critical" pairs



Crowd ER

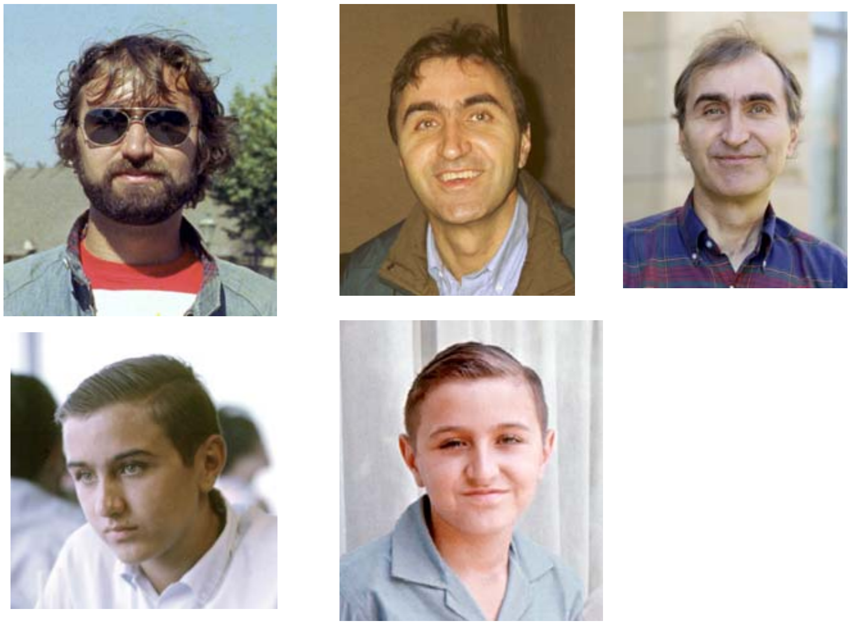
- Second Cut: Compute similarities; workers verify "critical" pairs

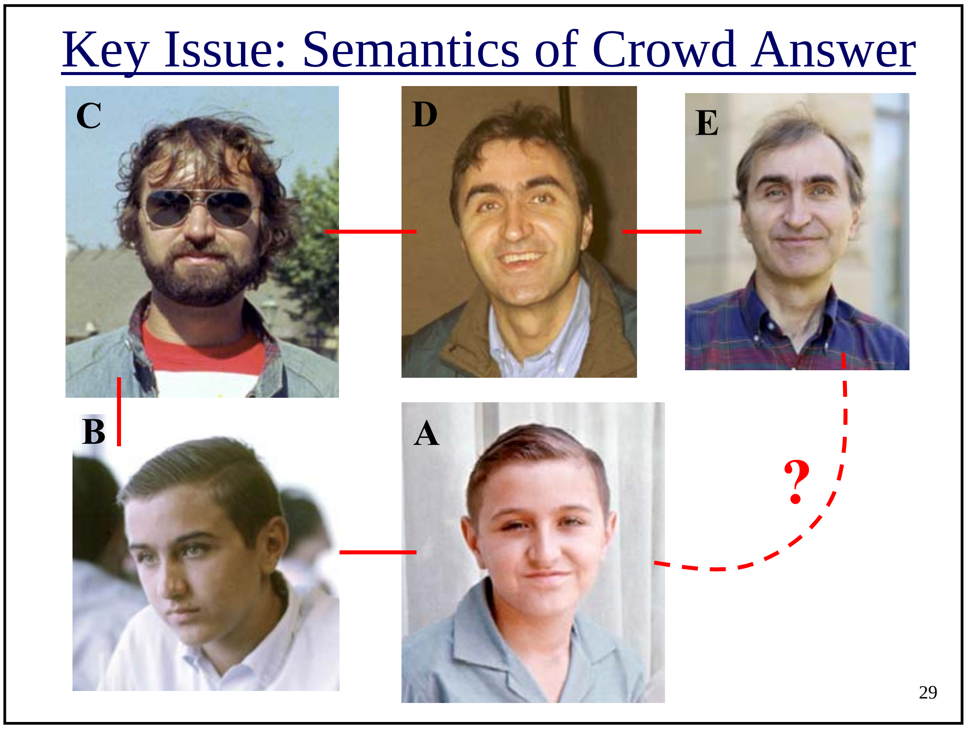


Key Point:

- use humans judiciously

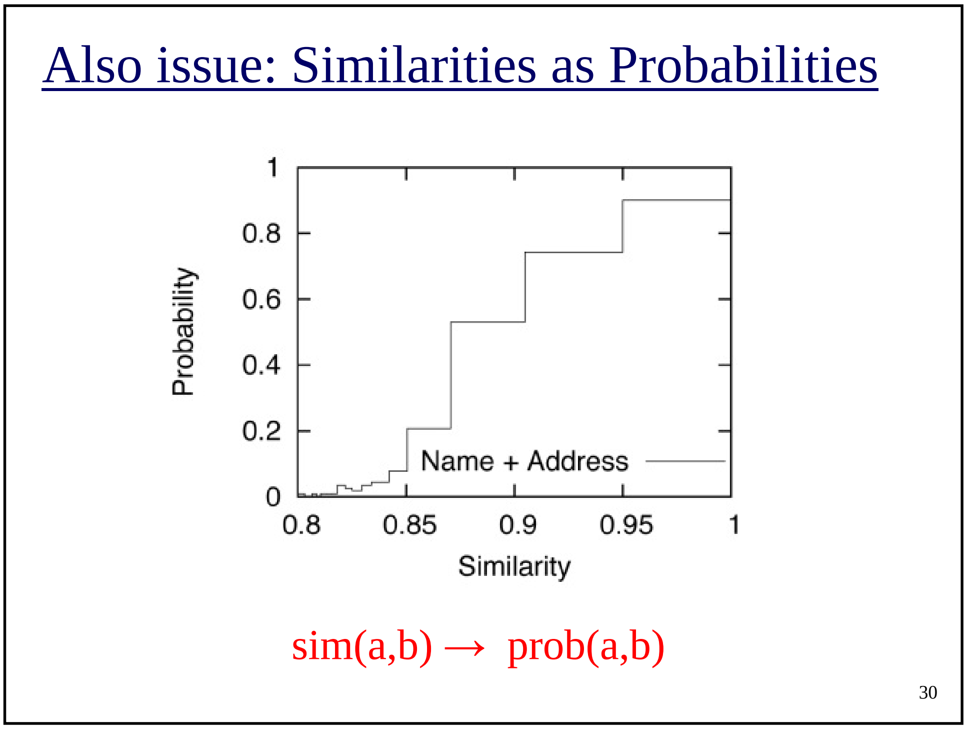
Key Issue: Semantics of Crowd Answer





Key Issue: Semantics of Crowd Answer

29

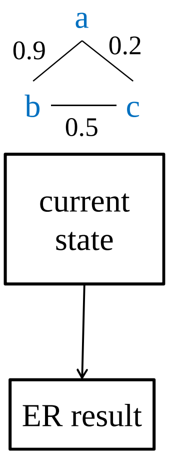


Also issue: Similarities as Probabilities

sim(a,b) → prob(a,b)

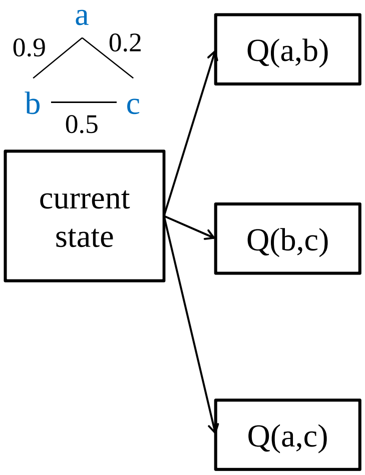
30

Strategy



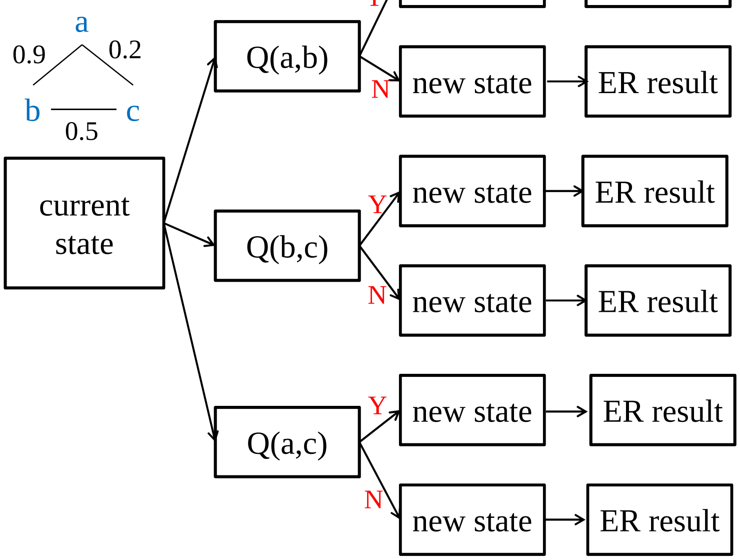
use any given ER algorithm

Strategy



consider ALL possible questions
(three in this example)

Strategy



a

0.9 0.2

b c

0.5

current state

Q(a,b)

Q(b,c)

Q(a,c)

Y new state → ER result

N new state → ER result

Y new state → ER result

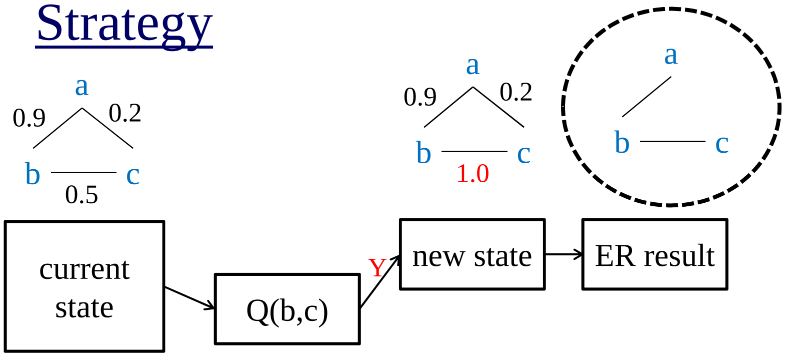
N new state → ER result

Y new state → ER result

N new state → ER result

consider possible outcomes

Strategy



a

0.9 0.2

b c

0.5

current state

Q(b,c)

Y new state → ER result

a

0.9 0.2

b c

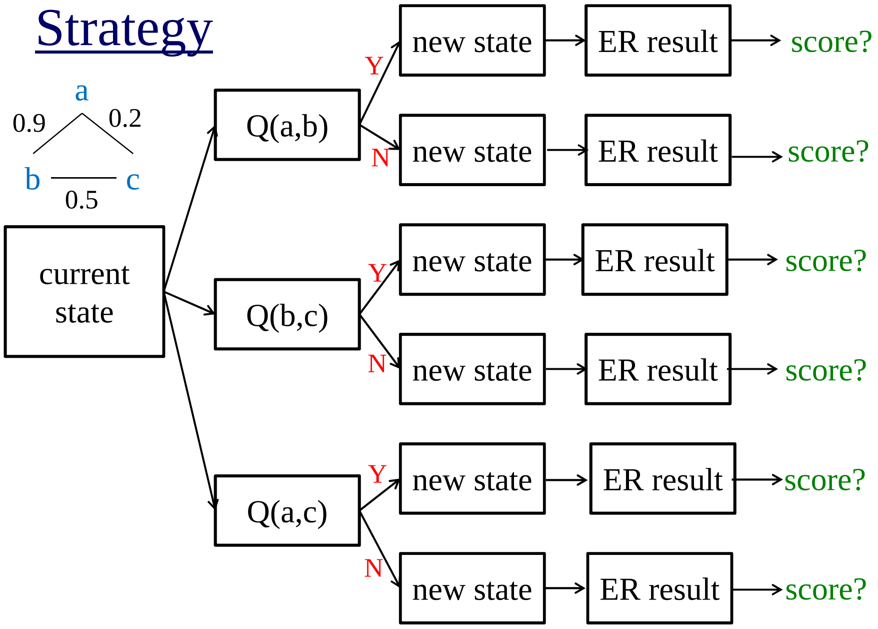
1.0

a

b c

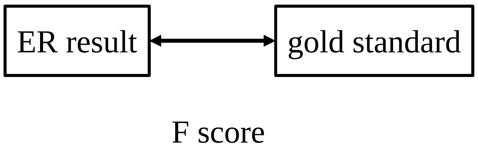
example

Strategy



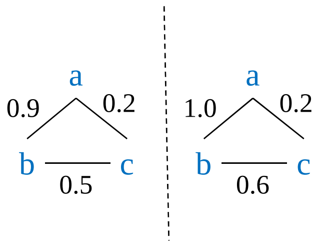
Two Remaining Issues

- How do we score an ER result?



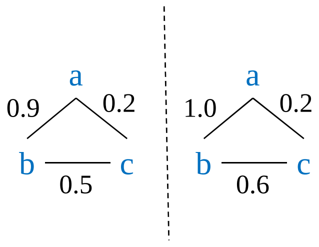
- Efficiency?

Gold Standard?



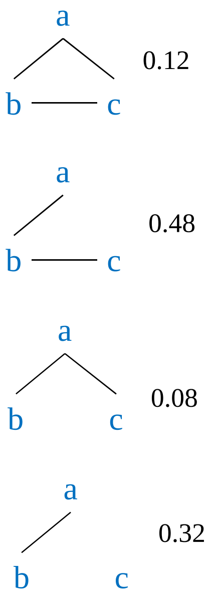
sim to prob

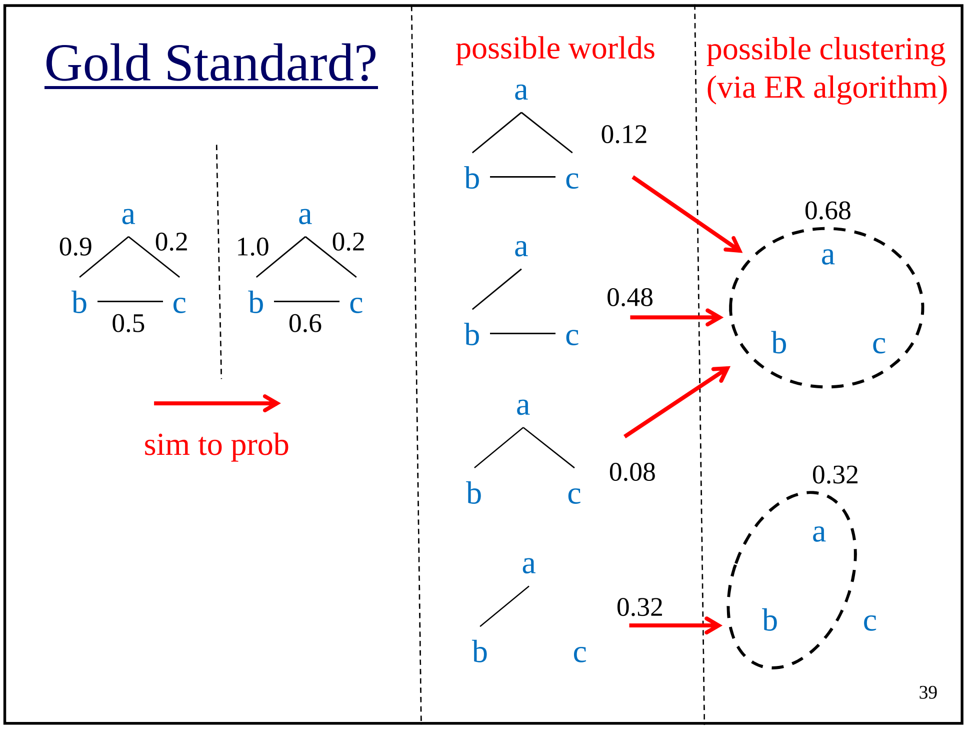
Gold Standard?



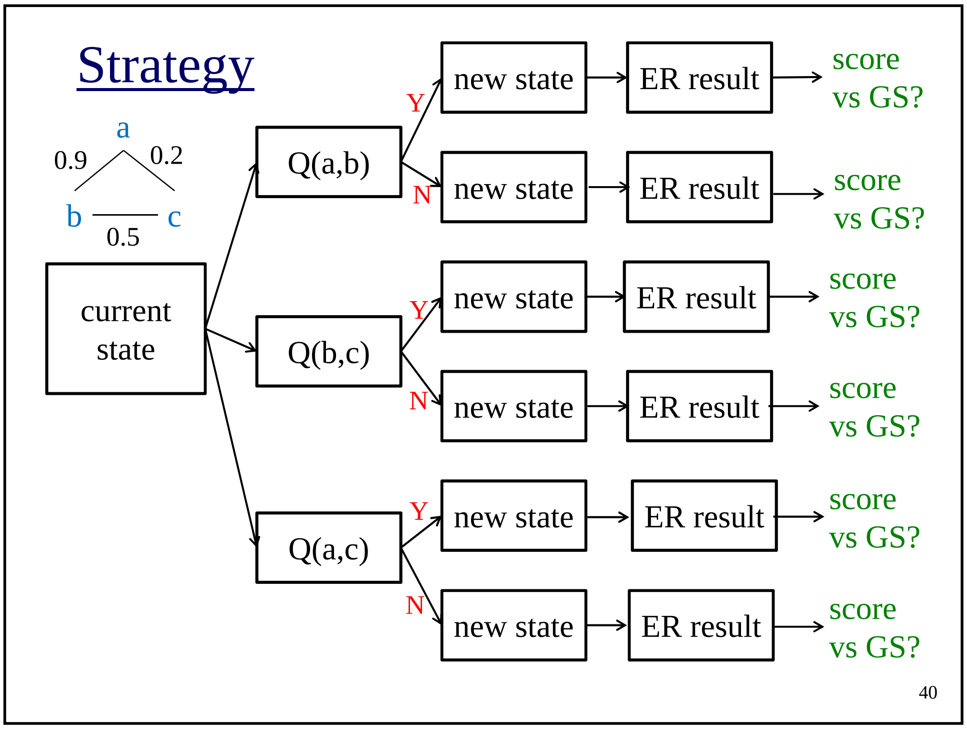
sim to prob

possible worlds





Gold Standard?

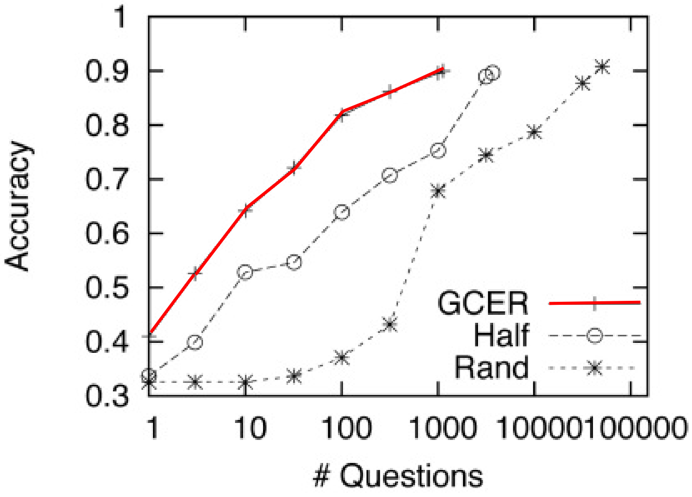


Strategy

Evaluating Efficiently

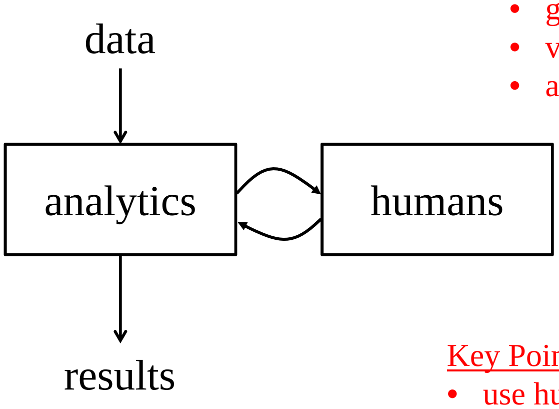
- See: Steven E. Whang, Peter Lofgren, and H. Garcia-Molina. Question Selection for Crowd Entity Resolution. To appear in Proc. 39th Int'l Conf. on Very Large Data Bases (PVLDB), Trento, Italy, 2013.

Sample Result



': Cora Results using High Threshold

Summary



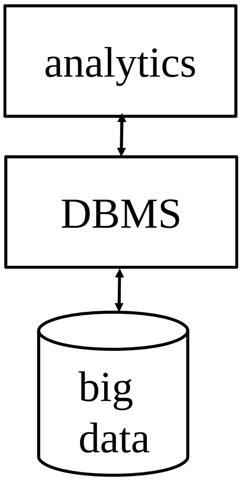
Example tasks:

- get missing data
- verify results
- analyze data

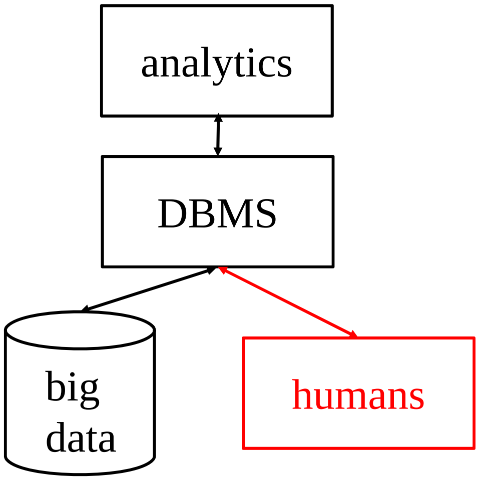
Key Point:

- use humans judiciously

Now for something completely different!



Now for something completely different!



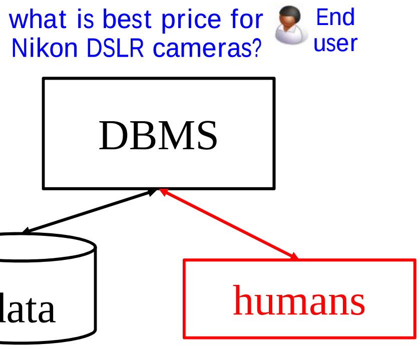
analytics

DBMS

big data

humans

DeCo: Declarative CrowdSourcing



what is best price for Nikon DSLR cameras?

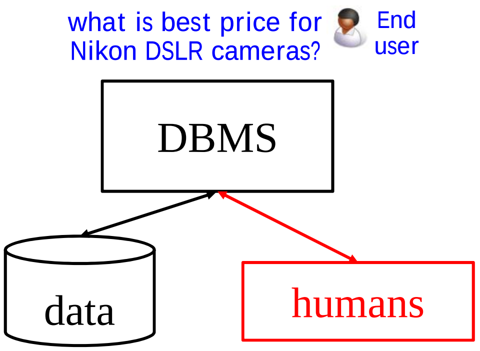
End user

DBMS

data

humans

DeCo: Declarative CrowdSourcing



what is best price for Nikon DSLR cameras?

End user

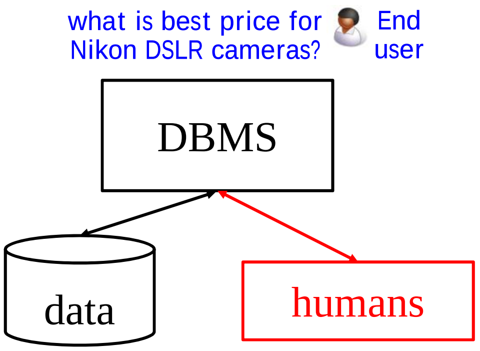
DBMS

data

humans

model	type	brand
D7100	DSLR	Nikon
7D	DSLR	Canon
P5000	comp	Nikon
•••	•••	•••

DeCo: Declarative CrowdSourcing



what is best price for Nikon DSLR cameras?

End user

DBMS

data

humans

model	type	brand
D7100	DSLR	Nikon
7D	DSLR	Canon
P5000	comp	Nikon
•••	•••	•••

what is best price for Nikon D7100 camera?

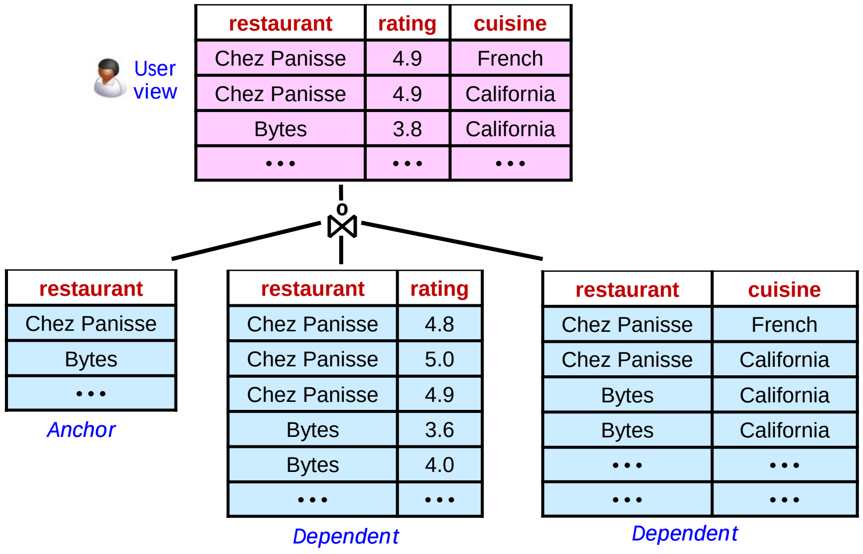
Crowd

Example with a bit more detail:

User view

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

Example with a bit more detail:



User view

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

restaurant
Chez Panisse
Bytes
• • •

Anchor

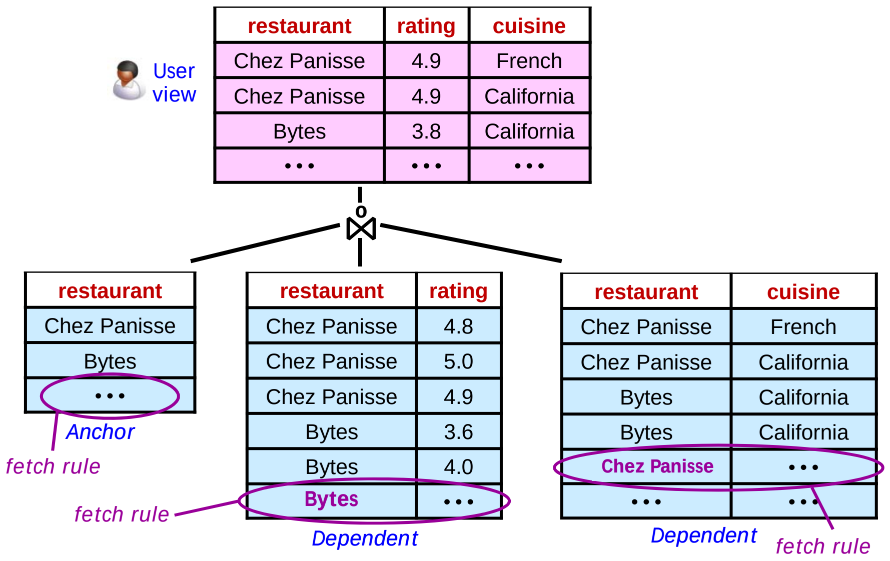
restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
• • •	• • •

Dependent

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

Dependent

Example with a bit more detail:



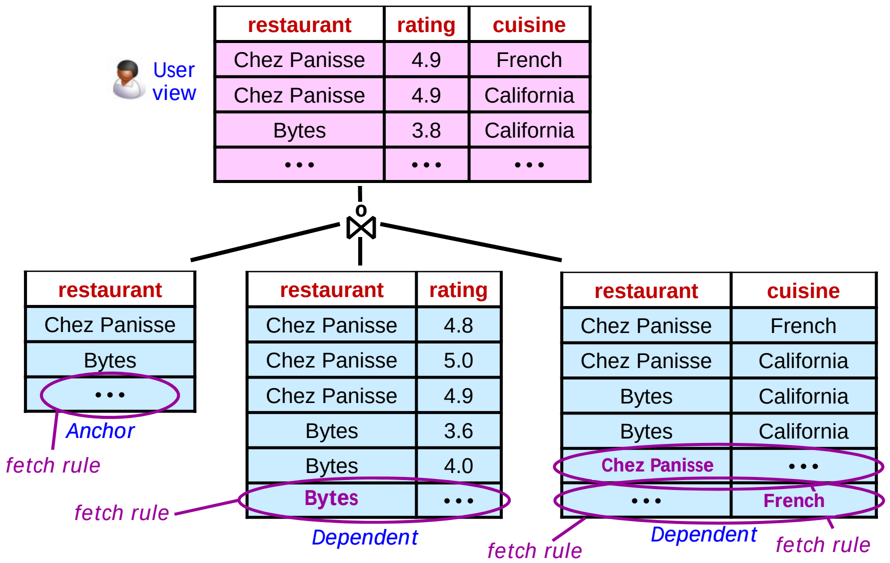
restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
•••	•••	•••

restaurant
Chez Panisse
Bytes
•••

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

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

Example with a bit more detail:

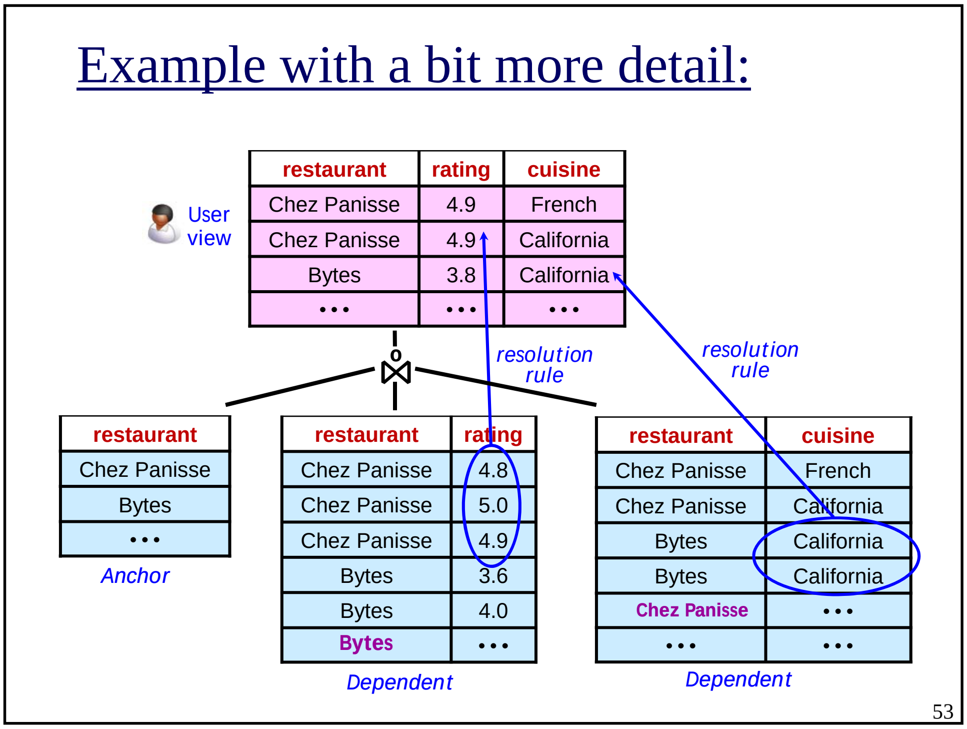


restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
•••	•••	•••

restaurant
Chez Panisse
Bytes
•••

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

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



Example with a bit more detail:

User view

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

restaurant
Chez Panisse
Bytes
• • •

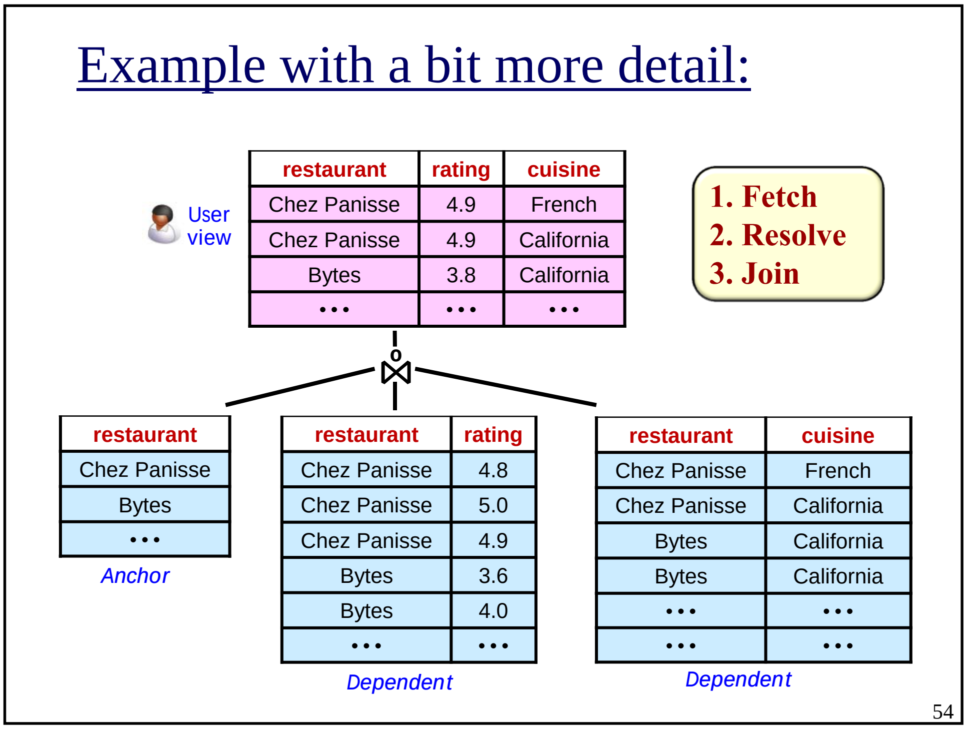
Anchor

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

Dependent

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

Dependent



Example with a bit more detail:

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
• • •	• • •

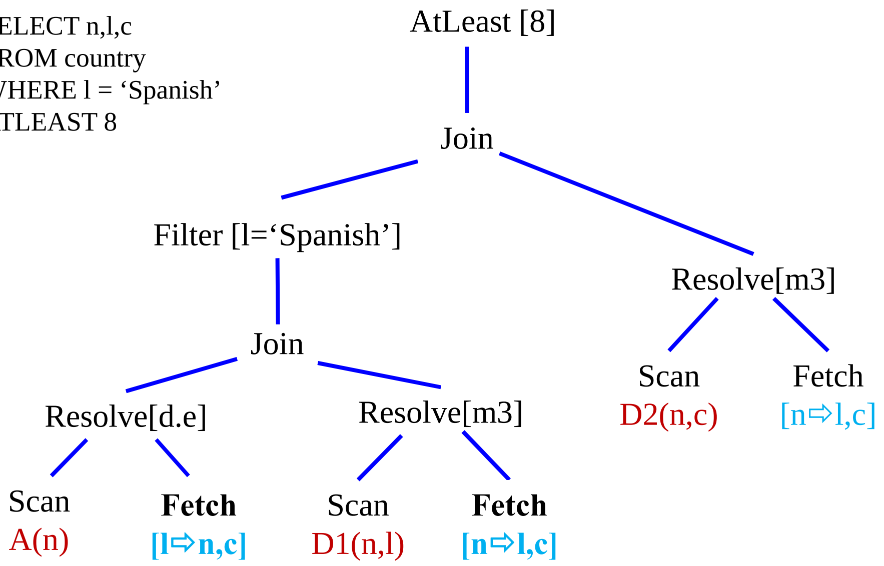
Dependent

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

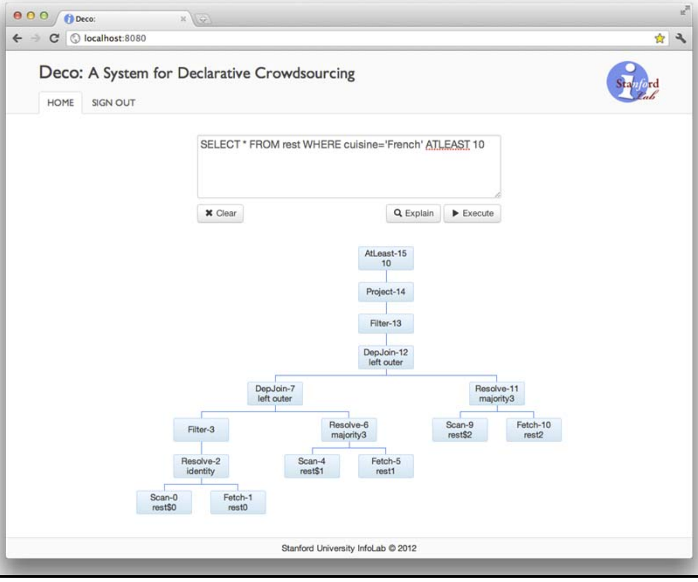
Dependent

Many Query Processing Challenges

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



Deco Prototype V1.0



Deco: A System for Declarative Crowdsourcing

SELECT * FROM rest WHERE cuisine='French' ATLEAST 10

Stanford University InfoLab © 2012

Conclusion

- Crowdsourcing is important for managing data!
- Still many challenges ahead!