The Demographics of Web Search

Ingmar Weber, Carlos Castillo

Yahoo! Research Barcelona

Warm-up DEMO

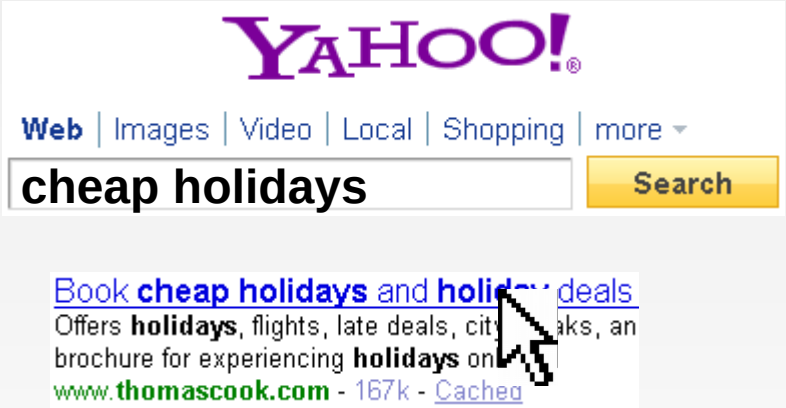
The DEMOgraphics of a query

offline slides

http://adlab.microsoft.com/Demographics-Prediction/DPUI.aspx

How the Data was Obtained

Gender: Male
Birth year: 1978
ZIP code: 95054



Q

D

US Census Data
factfinder.census.gov

Expected income: $ 31k

Expected education: 45% BA

Race distribution: 38% w, 47% A

Label (Q,D) with $31k, 45%BA, ...
Income_5, education_5, white_1, ...

quintiles

Yahoo! Users vs. US population

Feature	Y! p-q. aver.	US aver.	
P-c income $k	22.7	21.6	slightly richer
Bel. poverty %	11.1	12.4	
BA degree %	25.5	24.4	sl. m. educated
White %	76.9	75.1	
Afr. Amer. %	4.0	12.3	digital divide?!
Asian %	4.0	3.6	
Non-English %	17.3	17.9	
Year of birth	1970 med.	1974 med.	slightly older
Gender (f – m)	49.7 – 50.3	49.1 – 50.9	

Some Discriminating Queries

- Rich: “www.popsugar.com”
- Poor: “www.unitnet.com”
- Edu+: “spencer stuart executive search”
- White: “pulloff.com”
- Afr. Amer: “s2s magazine”
- Asian: “sina”
- Non-English: “mis novelas favoritas”
- Young: “free teen chatrooms”
- Old: “www.johnhopkinshealthalerts.com”

Experiments

- Want to rank a *target* for a certain *input*
  - P(“wiki.org/Richard_Wagner”|“wagner”)

target = URL U

input = query Q

- Add demographic condition
  - P(“wiki.org/ Richard_Wagner”|“wagner”,“male”)

demographic F

- (Q,D), (1st term, 2nd term), (D,Q)

Experiments

Only (input, target) pairs where for some demographic feature value F (a quintile)

users(input,F) ¸ 100 & users($\overline{\text{input}}$,F) ¸ 400

Only consider using demographic information when it is **not personalized**

Web Search

- Click behavior can depend on demographics
  - R. Wagner (female) vs. Wagner Spray Tech (male)
  - ESL Federal Credit Union vs. English as a Sec. L.

	# pairs	P@1 w/o F	P@1 with F
all (100+400)	207 Mio	.703	.713
H(D|Q)¸ 1.0	123 Mio	.557	.574
H(D|Q)¸ 2.0	60.6 Mio	.381	.408

Query Completion

- Given first term, suggest the second term
  - “frontpage X”, where X = ...
  - “2003” for most people
  - “free” for young people
  - “africa” for African Americans link
  - “magazine” for educated people link

	# pairs	P@1 w/o D	P@1 with D
all (100+400)	**459 Mio**	**.250**	**.276**

Differences to Personalization

- No per-person information aggregated
  - Fewer privacy concerns
  - Similar to publishing census information
- Make explanatory factors explicit
  - Age, gender, income, education, ...
  - Attractive for advertisers
- Should cope better with “cold start”
  - ZIP information gives a reasonable prior
  - Personalization still better for more data

Articles in NewScientist & Slashdot

Slashdot NEWS FOR NERDS. STUFF THAT MATTERS.

Stories Recent Popular Search

Bieeanda: So the search I did last night, for 'how to fix a cracked toilet', might result in 'hire a plumber, lady' instead of 'go to Home Depot for a replacement, dude'.

top-ranked result 7 per cent more often than in the standard Yahoo search." New Scientist mentions this research and two other innovative adjuncts to current search practice: following the mouse cursor as a proxy for eye tracking, and taking back bearings on online criminals by

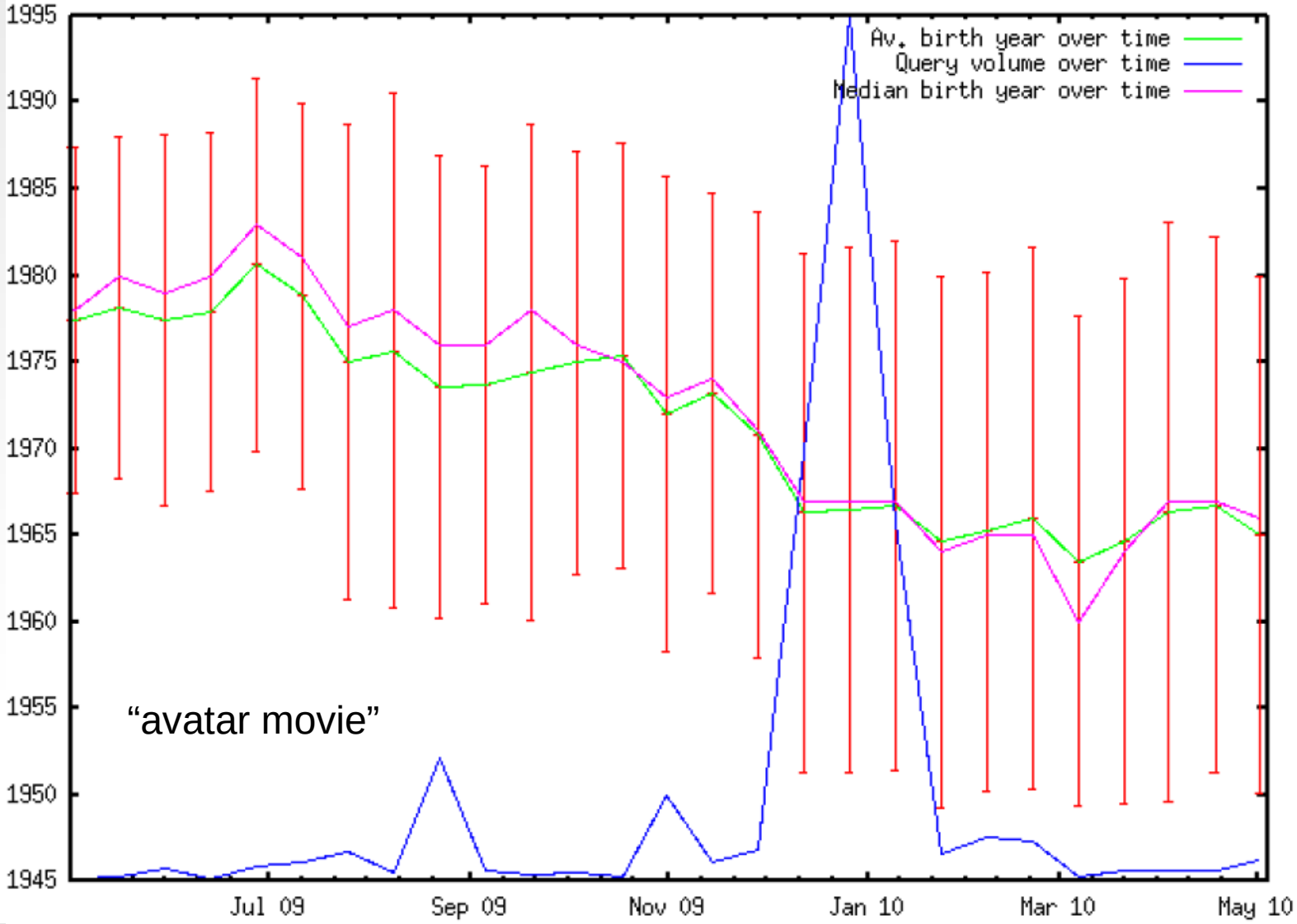
Should we avoid reinforcing stereotypes?
C.f. “Daily Me” (Negroponte)

Related Stories

Submission: The demographics of Web search by adaviel (1189751)

The Demographics of Web Search (50 Comments)

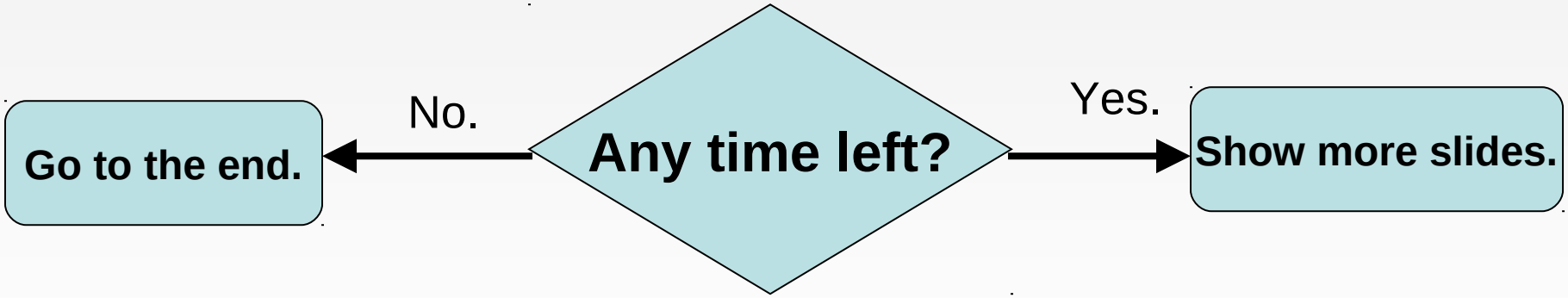
“Demographic Information Flows” @ CIKM 2010



“Demographic Information Flows” @ CIKM 2010

- “sonia sotomayor”
  - Pre-burst: large fraction of hispanic users
  - Burst: general population
  - Post-burst: large fraction of hispanic users
- Similarly: “ben bernanke” with BA degree

Parallel Universes



The End!

Thank you! **(~70% female query)**

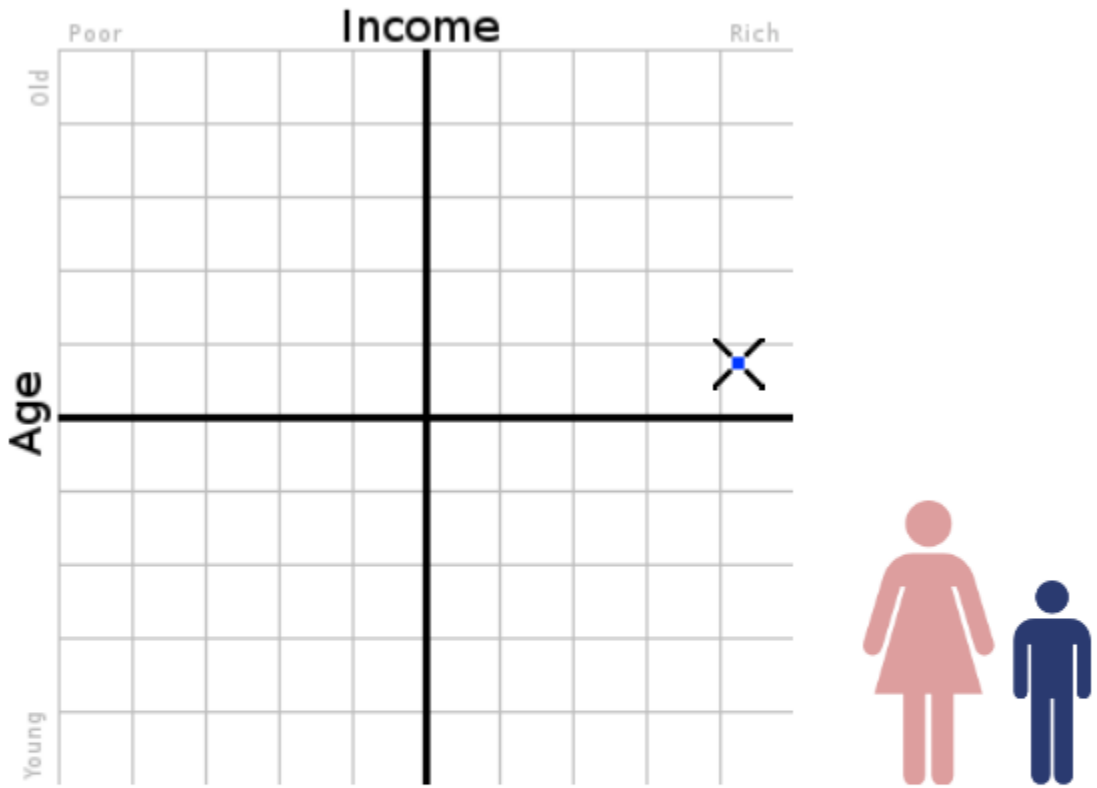
ingmar @ + chato @ yahoo-inc.com

Upcoming: “Demographic Information Flows”, CIKM 2010, Weber & Jaimes

Extra Slides

Extra Slides

“luxury resort”

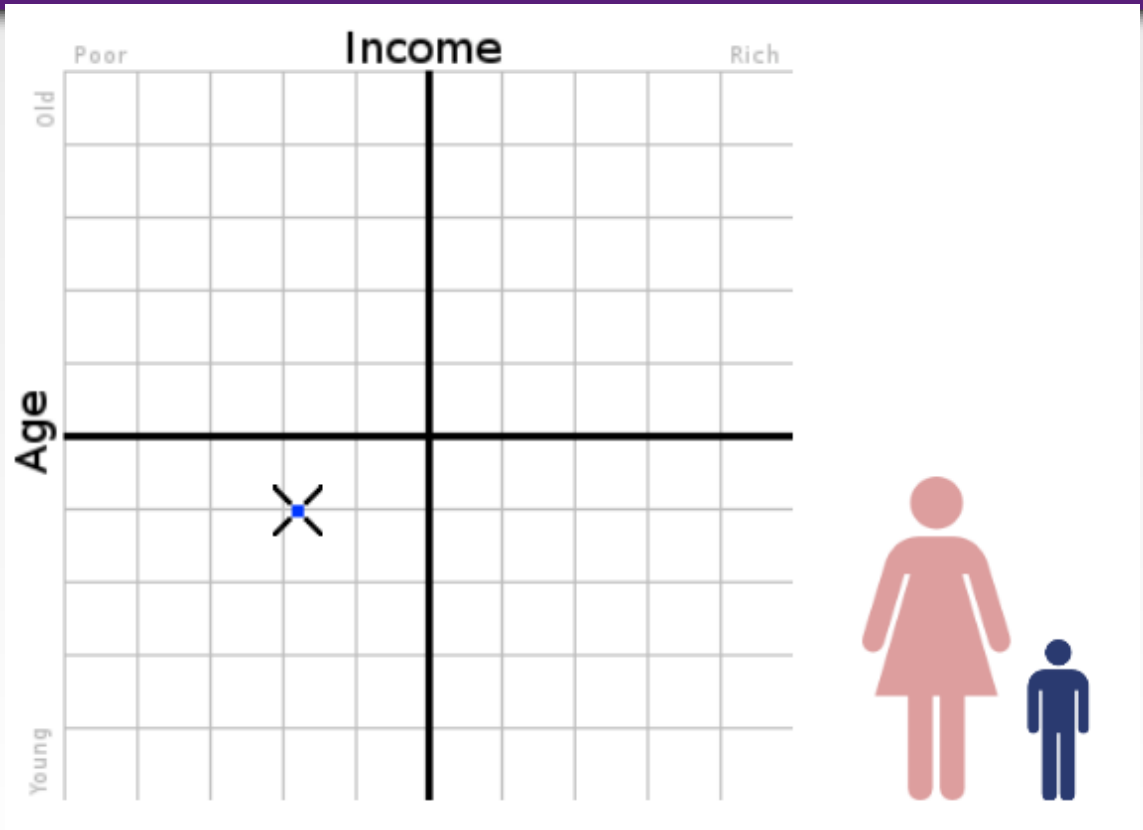


Query: luxury resort OK Method: Regular expression prefix

About this query:
Per-family income: USD $85403 (avg: USD $60000)
Average age: 43 (avg: 40)
Probability man: 42% woman: 58%

Back.

“food stamps”

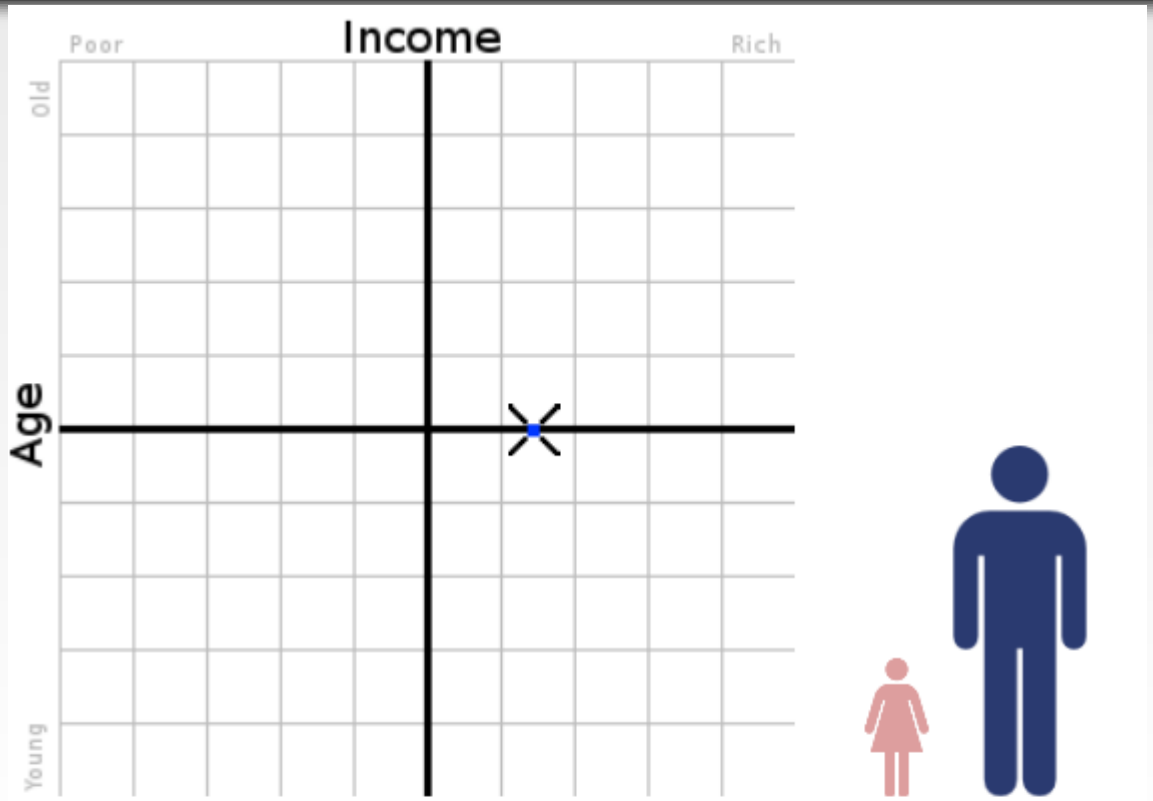


Query: food stamps OK Method: Regular expression prefix

About this query:
Per-family income: USD $49085 (avg: USD $60000)
Average age: 36 (avg: 40)
Probability man: 33% woman: 67%

Back.

“porsche”

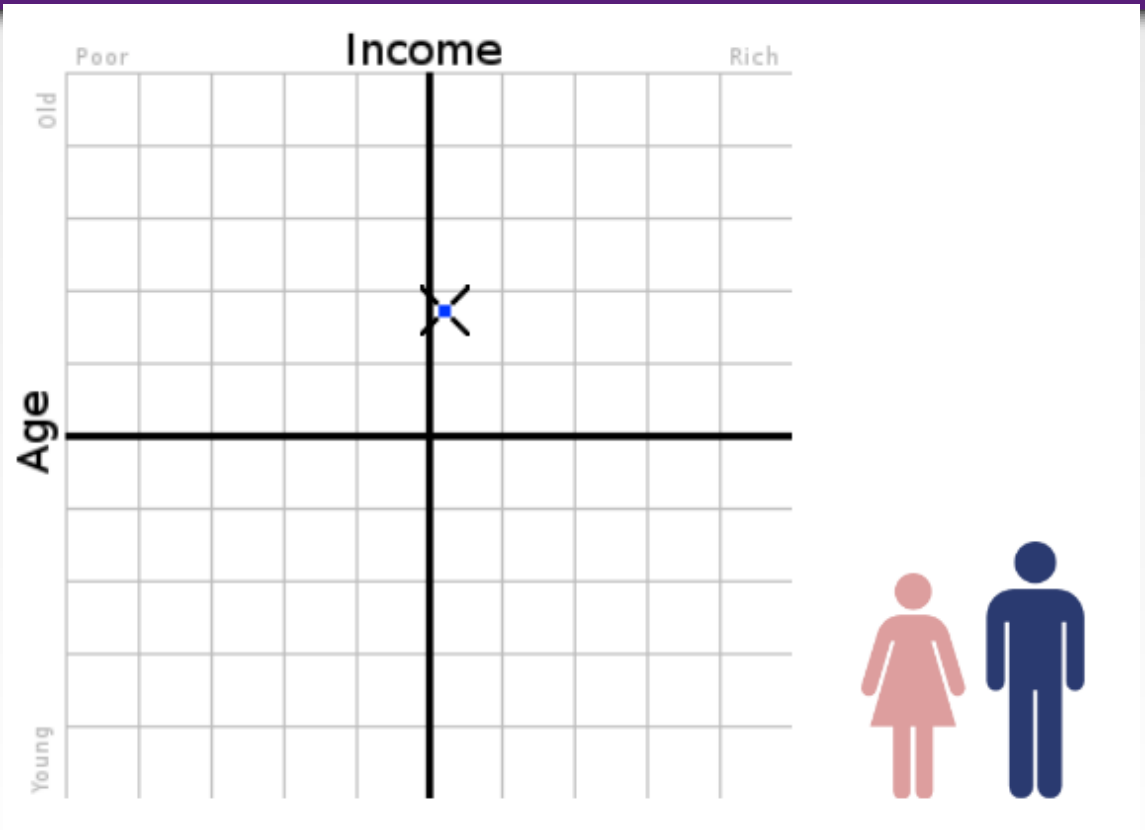


Query: porsche OK Method: Regular expression prefix

About this query:
Per-family income: USD $68622 (avg: USD $60000)
Average age: 40 (avg: 40)
Probability man: 72% woman: 28%

Back.

“retirement”



Query: retirement OK Method: Regular expression prefix

About this query:
Per-family income: USD $61196 (avg: USD $60000)
Average age: 47 (avg: 40)
Probability man: 53% woman: 47%

Back.

Finding “Deep Interest” Queries

- Low click entropy H(U|Q)
  - Usually navigational queries
  - No “deep interest”
- High click entropy H(Q|U)
  - “difficult” queries
  - “deep interest”

Examples: “scrapbooking” for young users

“civil war” for old users

The end.

URL Labeling

- Given a URL, what is the most likely query?
  - Automatic tagging

www.weedsthatplease.com/growing.htm

“how to grow weed” (young) vs. “marijuana growing” (old)

	# pairs	P@1 w/o D	P@1 with D
all (100+400)	**246 Mio**	**.461**	**.483**

The end.

Removing Localized Queries

- Keep the first two digits of each ZIP code
- For each query look at its “zip entropy”
- 6.23 bits across all queries
- Require 4.00 bits for a “nation-wide” query
- Example list of discriminative queries only shows nation-wide queries

The end.