Bayesian Magic for Complex Social Science Data:

Fusion, Nonparametrics, Dynamics, Dyads, Networks

ICOS **Big Data** Summer Camp

University of Michigan

June 5-9, 2017

Fred Feinberg

Ross School of Business and Department of Statistics
University of Michigan

MICHIGAN
ROSS SCHOOL OF BUSINESS

I know what you're thinking

Bayesian?

PLEASE MAKE IT STOP

I know.

I *know*.

Instead: Think "Pachyderm"

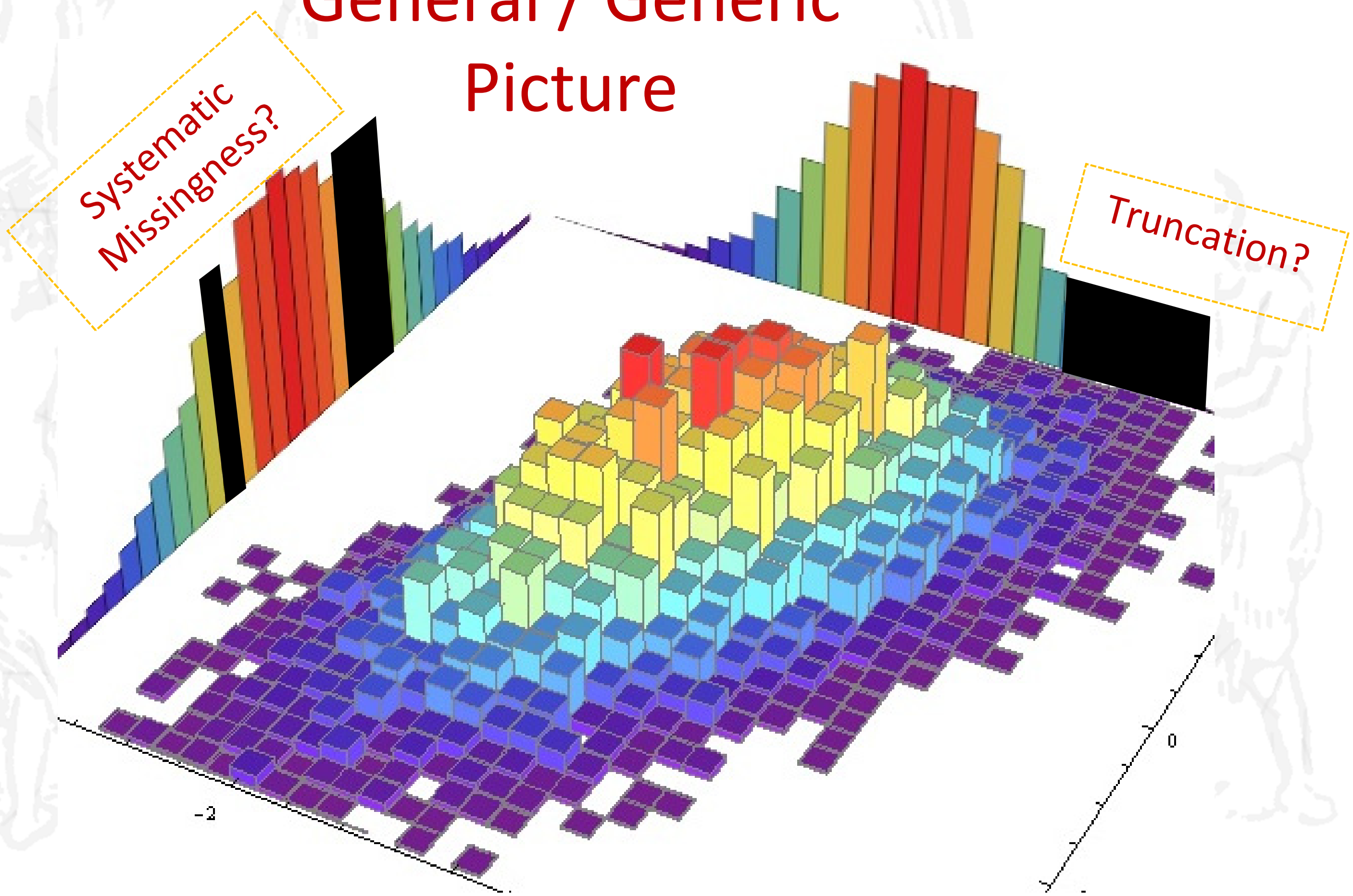
More intelligibly: It's a DATA POTLUCK

Everyone can "bring" their best data and FUSE them using a **behaviorally-plausible model**

T'was six wise men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind…

"TL;DR" Version:
#1: Side = Wall
#2: Tusk = Spear
#3: Trunk = Snake
#4: Knee = Tree
#5: Ear = Fan
#6: Tail = Rope

General / Generic Picture

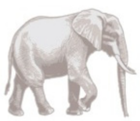


FAQ: Questions Surely on Someone's Mind

Q: Everyone's talking about Big Data, particularly **employers**.
What *is* Big Data anyway?

A:

DATA



Not all Big Data Created Equal

Olden Days

DV: Some Outcome (housing, jobs, marriages, …)

IVs: GeoDemographics (age, income, education…)

[Some can be “stated preferences”: e.g., surveys]

Then… use some (sophisticated!) regression approach to “figure out what’s going on”

Problem: <u>MORE DATA ALONE</u> don’t help!

Good Big Data = **PROCESS** Data

Electronic trails: online dating; real estate searches; Amazon clickstream; school and job applications; GPS tracking; housing patterns; etc.

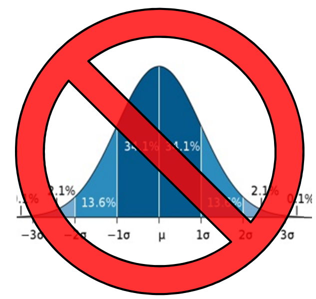
1) Novel **revealed preference** data on how people navigate social & physical environments

2) [Bayesianly!] Fuse data *with different deficiencies* to **jointly overcome them**

A Quasi-Cohesive Cornucopia of Important Opportunities for Data-Driven Social Science

“IMHO”

Fusion: Melding really different data sets



Nonparametrics: Minimize assumptions

Sparseness: Most data just ain’t there

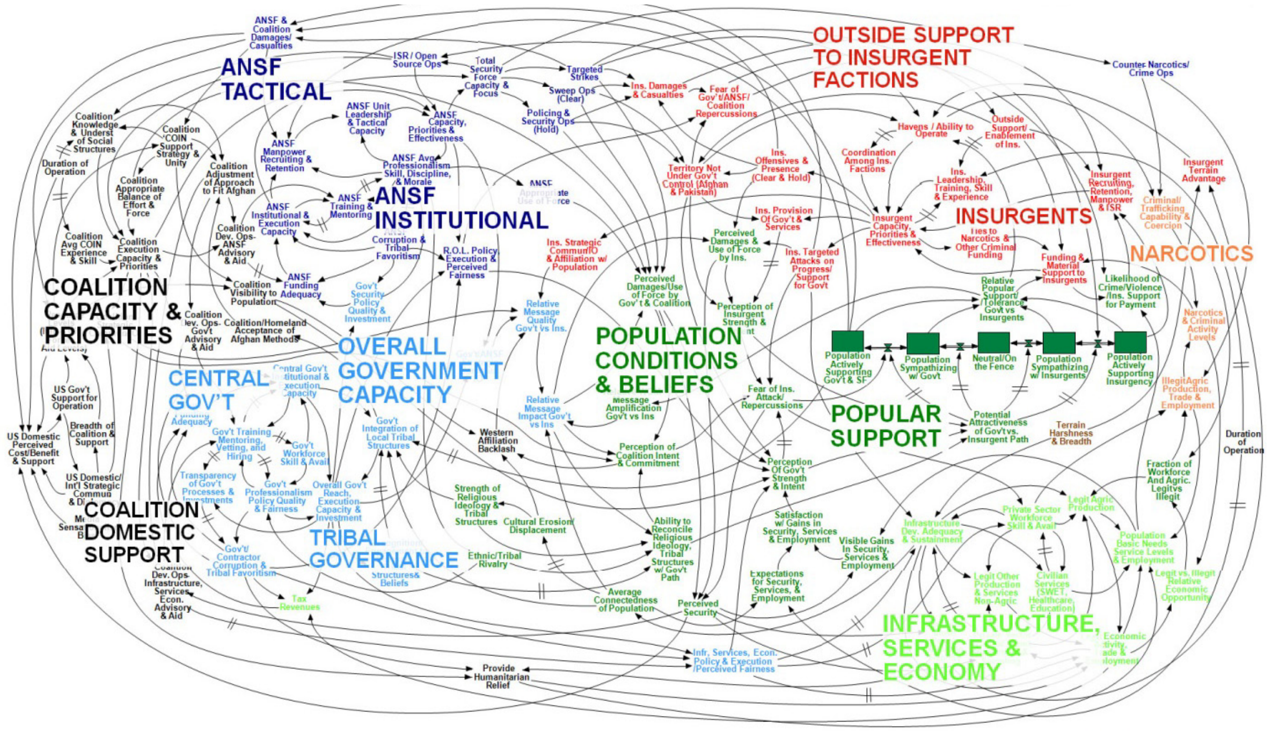
Dynamics: Everything (people, neighborhoods) changes

Dyads and Networks: Leveraging connections

Noncompenatory Behavior: “Deal Breakers”?

Oh, You Mean Machine Learning! Well… No

"Everything causes everything else"



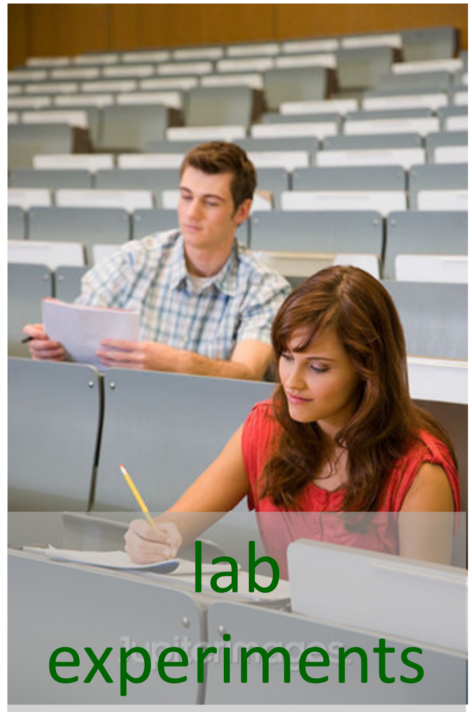
Problem with machine ("deep") learning view:

Models reproduce reality without describing it in "human accessible" terms

MICHIGAN
ROSS SCHOOL OF BUSINESS

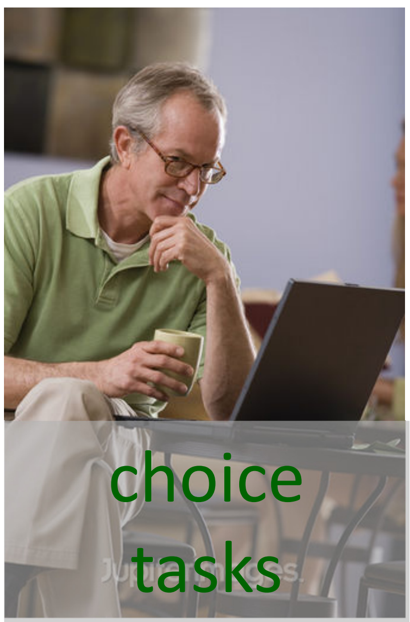
Examples: Individual-Level “Sociological” Data

Purchases



lab experiments

Surveys



choice tasks

GeoDemographics

Housing

MICHIGAN
ROSS SCHOOL OF BUSINESS

Data Fusion Example: Limitations of EXISTING Data for Empirical Social Science

No information about **preferences** for **new** social programs, businesses, transportation, local institutions…

Limited information about **preferences** for **existing** attributes

Should it have a pool? Parking?	Entrances? Hours?	Tuition? Location? Multilingual?

Limited information on **heterogeneity** in **preferences**

WHY Fuse Data?

Reality! But…

No info about **new** possibilities

Limited information about:

- **Existing** attributes (collinearity)
- Heterogeneity (few or no repeated measures for individuals / households)

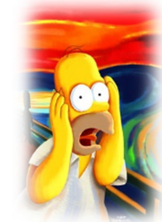
- Control
- Experimental design

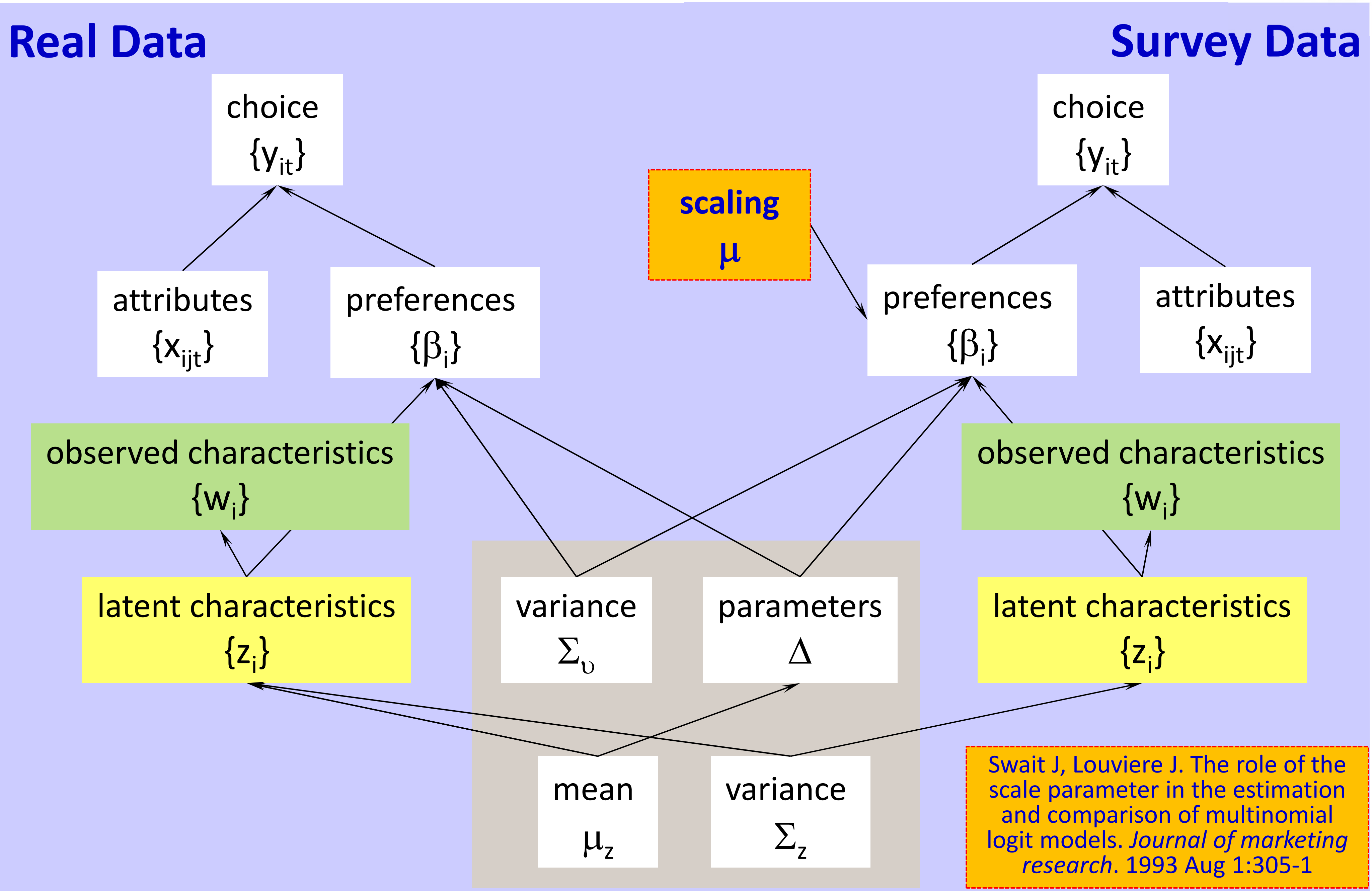
But.. Not "reality"

[Various biases: status quo, social desirability, conformity,…]

MICHIGAN ROSS SCHOOL OF BUSINESS

Hierarchical Bayes Modeling Framework: Fusion with Missing Data





Fancy! But… how about a REAL example?

"Public school choice"

Ample **actual choice data** (ranked preferences, actually)

Some **survey data**

Many (aggregated) covariates on both schools and neighborhoods: incomes, ethnicity, distance to schools, quality metrics, household composition, etc.

Big Question: **How** do families decide which school(s) they prefer for their child?

This is a question about both PROCESS and CHOICE

Has This Been Done?

Extracting multistage screening rules from online dating activity data

Elizabeth Bruch, Fred Feinberg, and Kee Yeun Lee

Dating Data (Bruch, Feinberg, Lee, PNAS 2016)

A "realistic" 2-stage model of mate choice behavior

- **Browsing** (1st stage) / **Writing** (2nd stage)

Identifying (heterogeneous) decision rules

AND (homogeneous) "human universals"

Allow for **non-compensatory rules**:

"deal-breaker" / "deal-maker"

"Questions from Teddy"

a) Your background

b) Your toolkit of computational methods

c) How you learned this material

d) What you are working on

e) Inspirational words of wisdom for beginners!

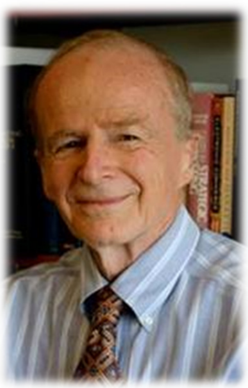
MIT-Sloan, 1984-88

NO idea what I'm doing. Never took a business course before!

<u>CORE Award Citation</u>: "… Professor Feinberg's *unique and wide-ranging methodological expertise* has made him an extraordinarily valuable colleague and mentor to faculty and PhD students…"

1984: Took my one-and-only stats course ever. Loathed it.

1985: Asked to TA it for a cool guy named Tony Wong. Finally got it!



Got to know John Little, of "Little's Laws" fame. Read papers on optimal control of advertising models… *which had lots of math*.

I ask him to Chair my dissertation on that topic. He says Yes!

Started to learn choice modeling,

which he'd brought into the field.

But what about the “Computational Social Science” stuff, huh?

Elizabeth Bruch
Sociology

QMP Quantitative Methodology Program

Fred Feinberg
Ross-Business

Gives talk on discrete choice models at QMP

“Do you know about uses of this in Sociology?”

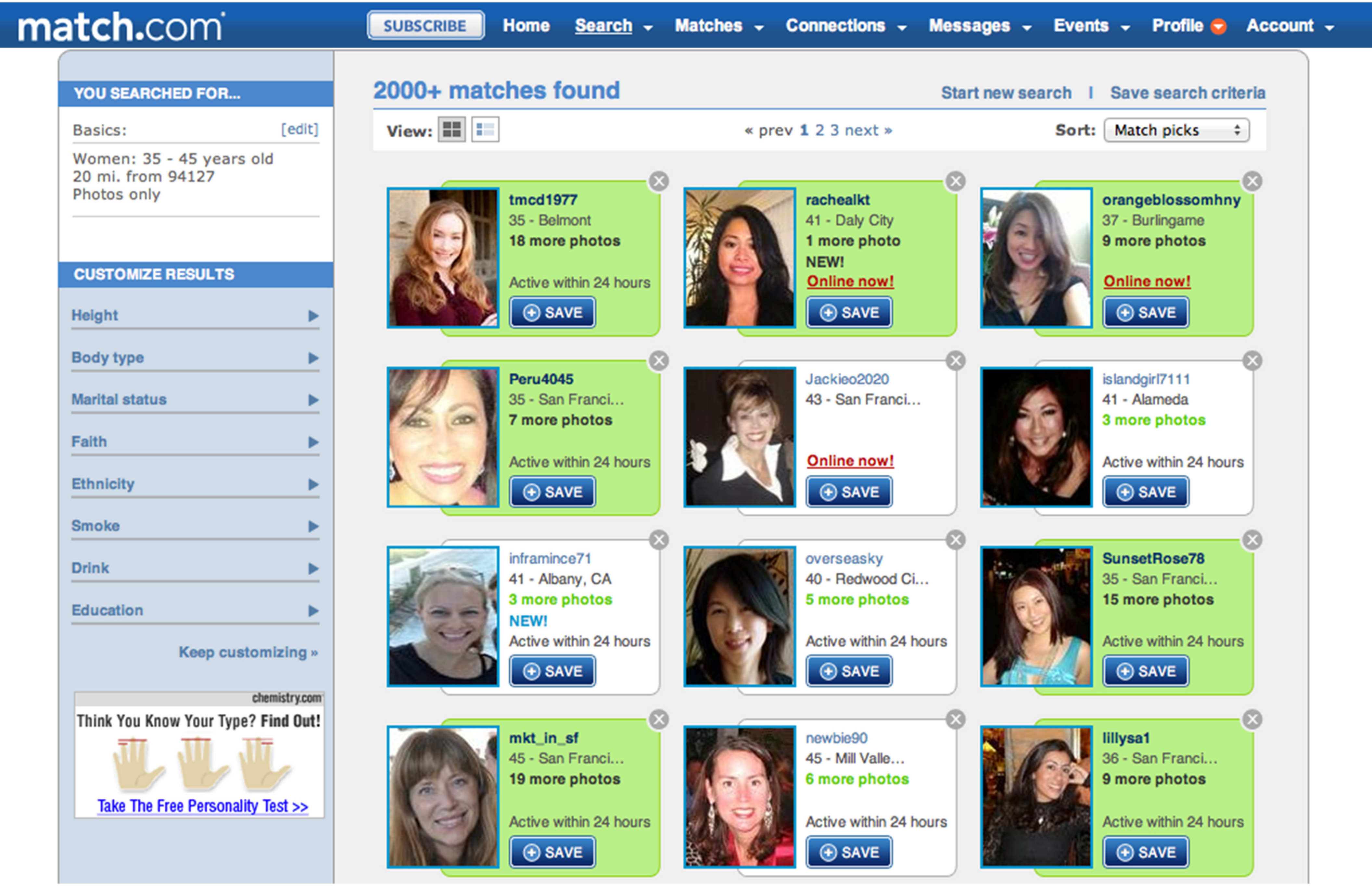
“Nope.”

“I think there are uses for this in Sociology. Can we chat about it?”

“Sure!”

In 2014, both are at Stanford / CASBS, work intensively on these data

"Mate Search"


match.com
SUBSCRIBE
Home
Search
Matches
Connections
Messages
Events
Profile
Account
YOU SEARCHED FOR...
Basics: [edit]
Women: 35 - 45 years old
20 mi. from 94127
Photos only
CUSTOMIZE RESULTS
Height
Body type
Marital status
Faith
Ethnicity
Smoke
Drink
Education
Keep customizing »
chemistry.com
Think You Know Your Type? Find Out!
Take The Free Personality Test >>
2000+ matches found
Start new search | Save search criteria
View:
« prev 1 2 3 next »
Sort: Match picks
tmcd1977
35 - Belmont
18 more photos
Active within 24 hours
SAVE
rachealkt
41 - Daly City
1 more photo
NEW!
Online now!
SAVE
orangeblossomhny
37 - Burlingame
9 more photos
Online now!
SAVE
Peru4045
35 - San Franci...
7 more photos
Active within 24 hours
SAVE
Jackieo2020
43 - San Franci...
Online now!
SAVE
islandgirl7111
41 - Alameda
3 more photos
Active within 24 hours
SAVE
inframince71
41 - Albany, CA
3 more photos
NEW!
Active within 24 hours
SAVE
overseasky
40 - Redwood Ci...
5 more photos
Active within 24 hours
SAVE
SunsetRose78
35 - San Franci...
15 more photos
Active within 24 hours
SAVE
mkt_in_sf
45 - San Franci...
19 more photos
Active within 24 hours
SAVE
newbie90
45 - Mill Valle...
6 more photos
Active within 24 hours
SAVE
lillysa1
36 - San Franci...
9 more photos
Active within 24 hours
SAVE

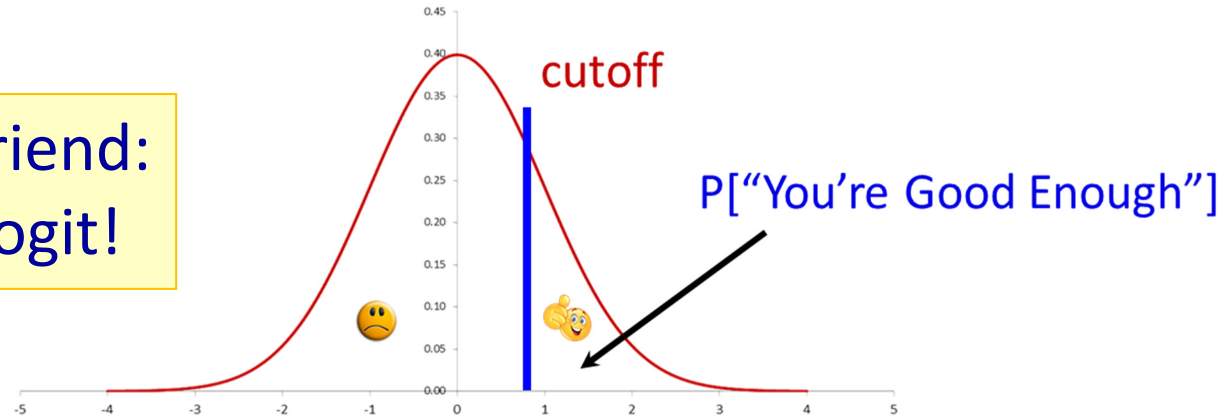
But what do these (Big) Data look like?

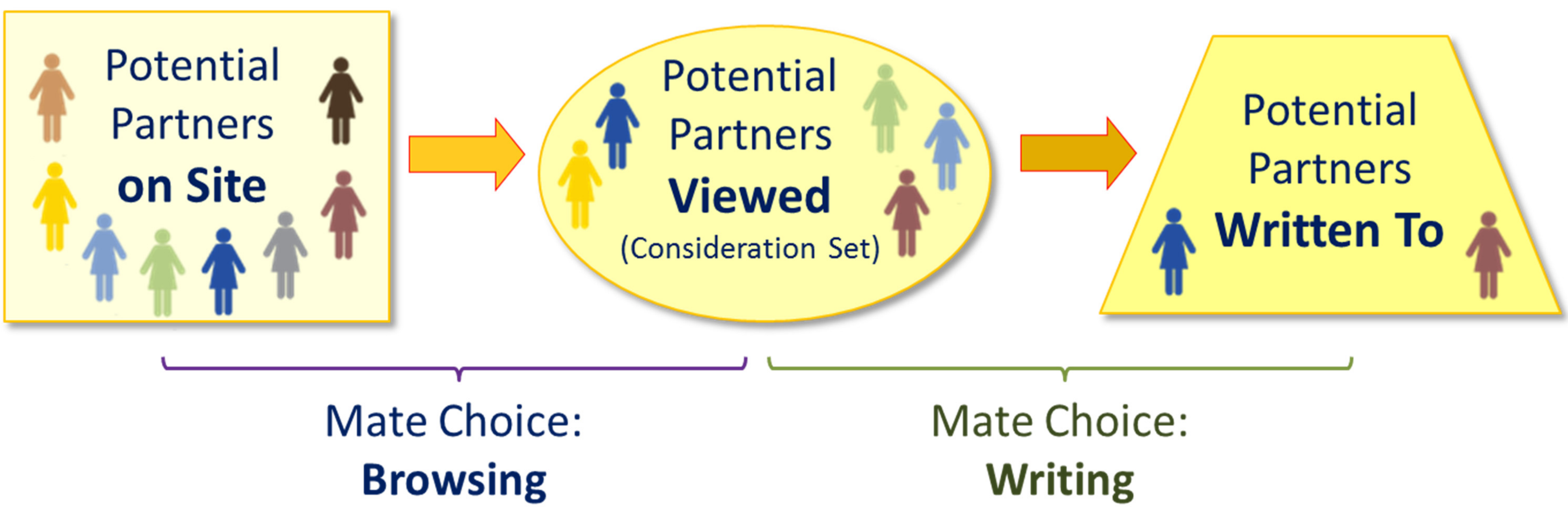
Profile Data	Search Data	Browsing Data	Messaging Data
• **Demographics** (age, income, occupation, height, body type, etc.) • **Attitudes, Desires, & Beliefs** (e.g., monogamy, marriage, deception, willingness to date fat people, etc.) • **Text fields** (words, unique words, words > 6 letters, photos, etc.) • **Account info** (start date, last login, reasons suspended or canceled) • **Attractiveness Ratings** (dyadic; disaggregate)	• **Attributes & values** (age range, distance, race/ethnicity, etc.) • **Sort order** (distance, random, attractiveness, match) • **ID of profiles** (that met search criteria) • **Ordering of results** (discretized)	• **ID of profiles** (that met search criteria) • **Ordering of results** (discretized)	• Words • Unique words • Words > 6 letters • Email address • Phone number • Pos. / Neg. words • Hedge words • Sympathy words • Self references (myself, I, etc.) • Partner references (you, yourself, etc.) • Third person references (he, himself, etc.) • Other keywords from ngram analysis

How Do People Find Others Online?

1. Who's good enough for me to **browse**? ["browsing utility"]
2. Now… of those browsed, who's good enough to **write to**? ["writing utility"]

It's our friend: binary logit!





Key Features of Model

Uses actual behavior: browsing and writing

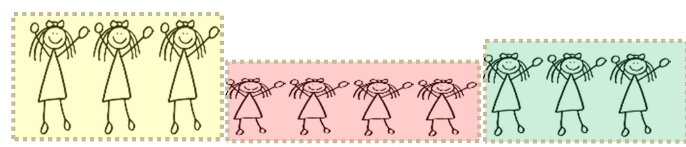
People can have "**deal breakers**" or "**deal makers**":

"I **won't** go out with anyone over 40"

"I **need to** date someone vegan"

"Having a PhD is a **huge plus**"

Users parceled into **groups**



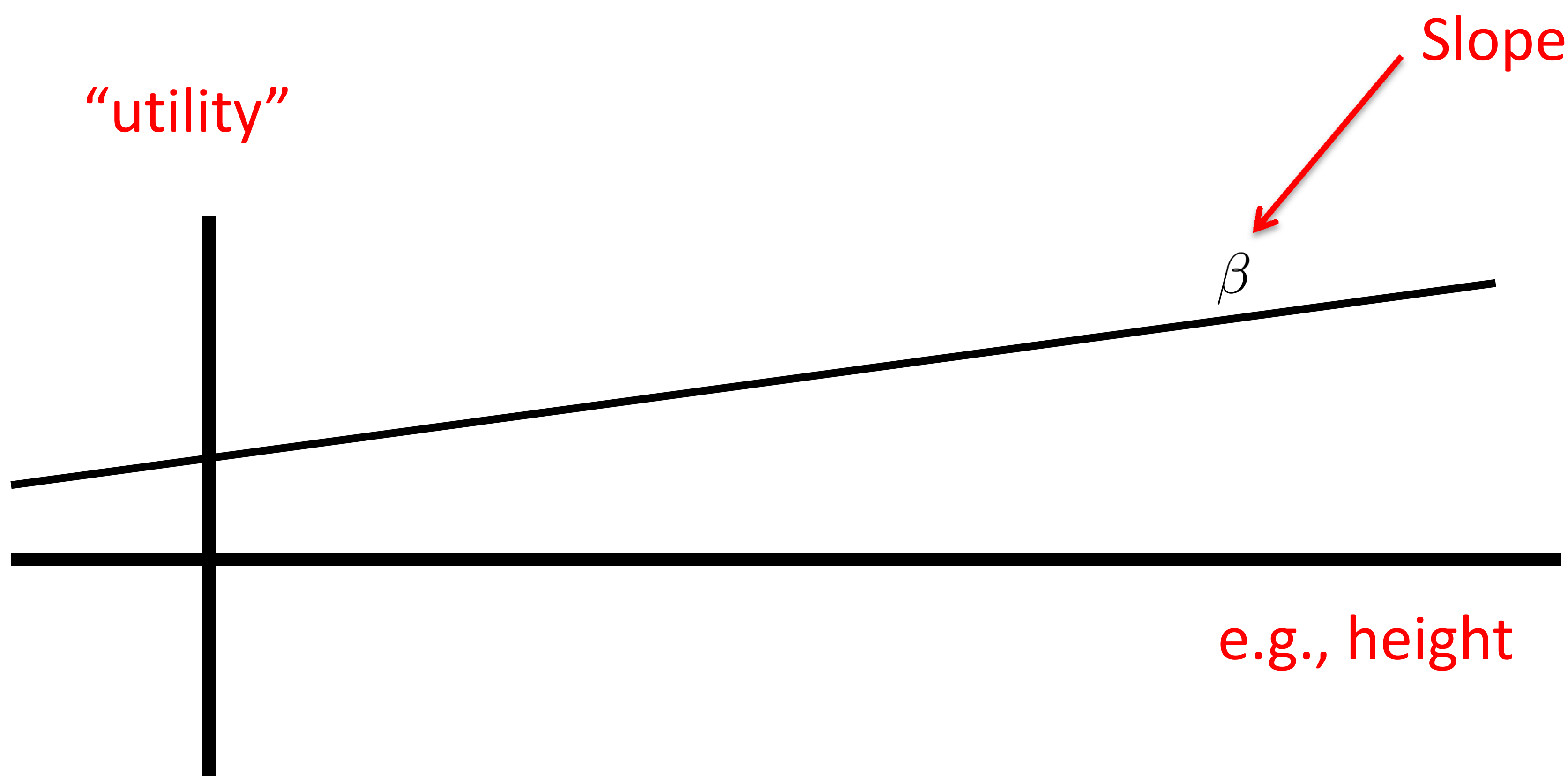
Easy to use as a **predictive model**

Can incorporate **stated preferences**

Massively **multivariate**: dozens of variables possible

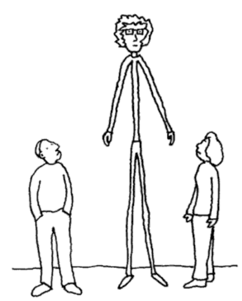
Usual Assumption in "Discrete Choice Models"

Monotonicity: More is Always Better (or Worse)

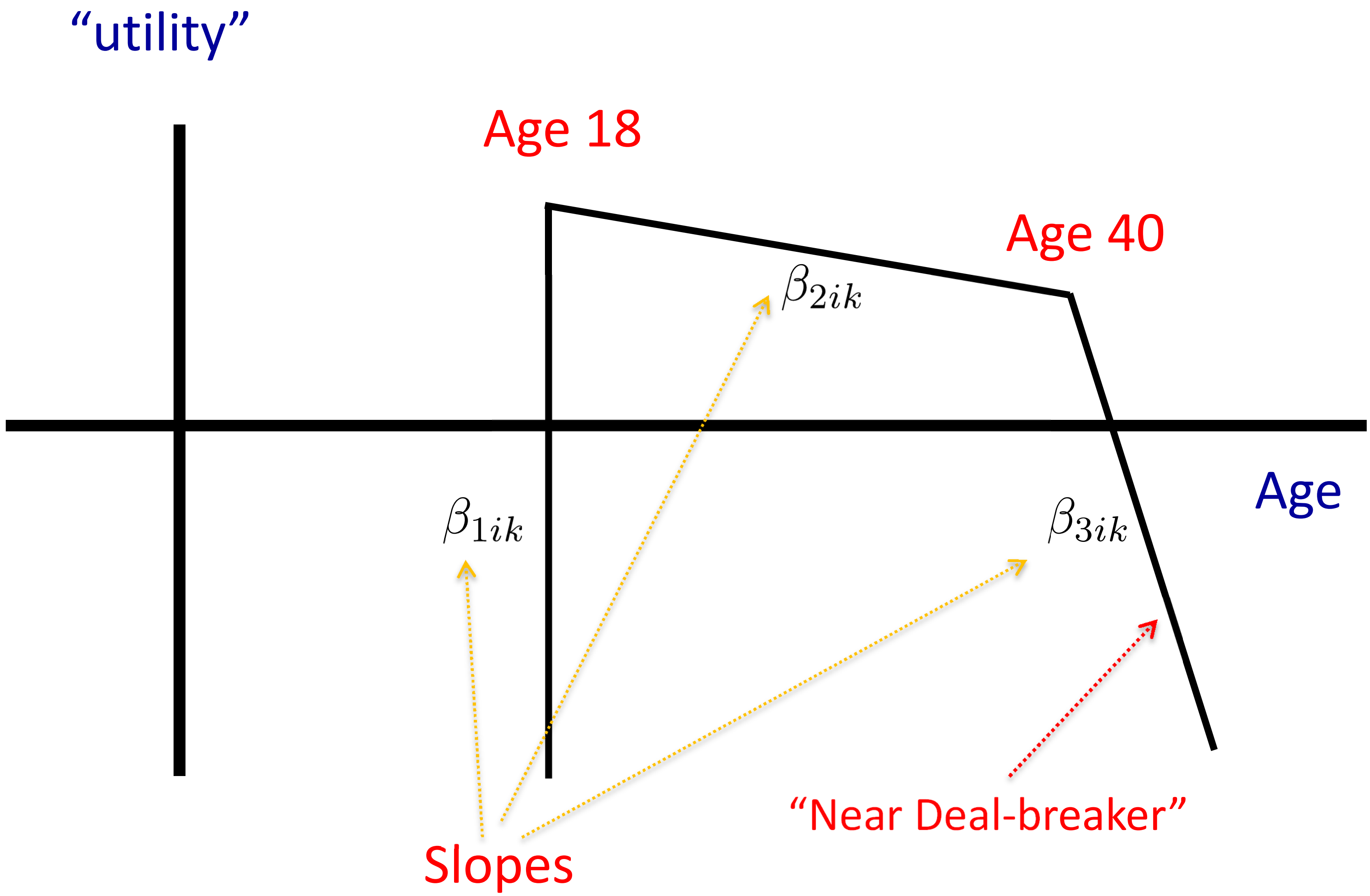


"The taller, the better"

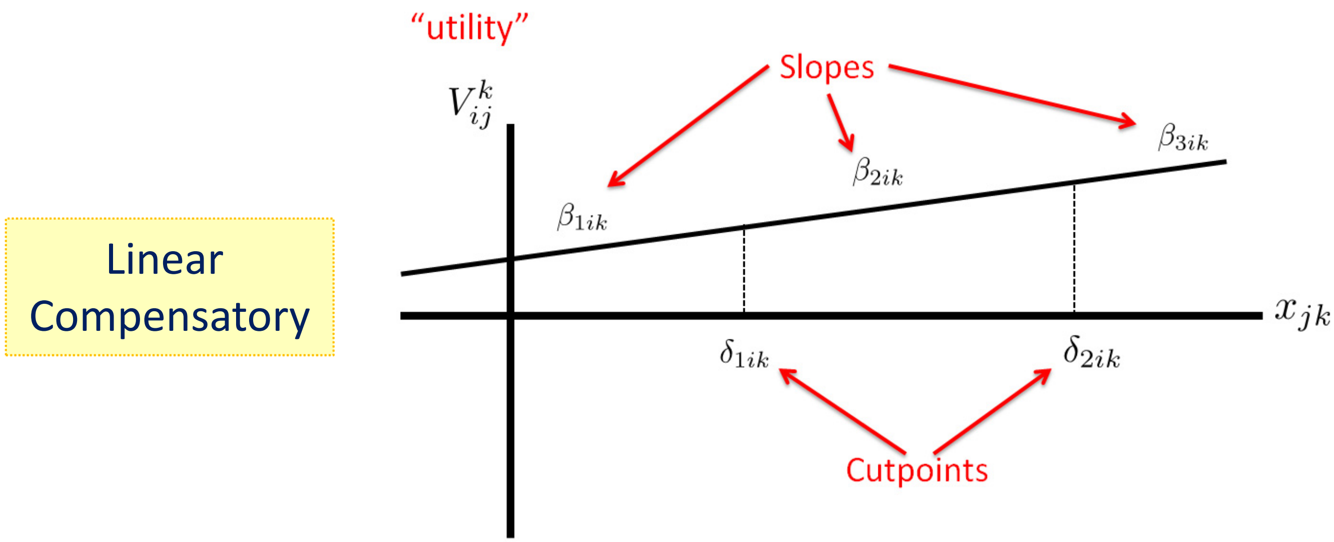
But is this realistic?

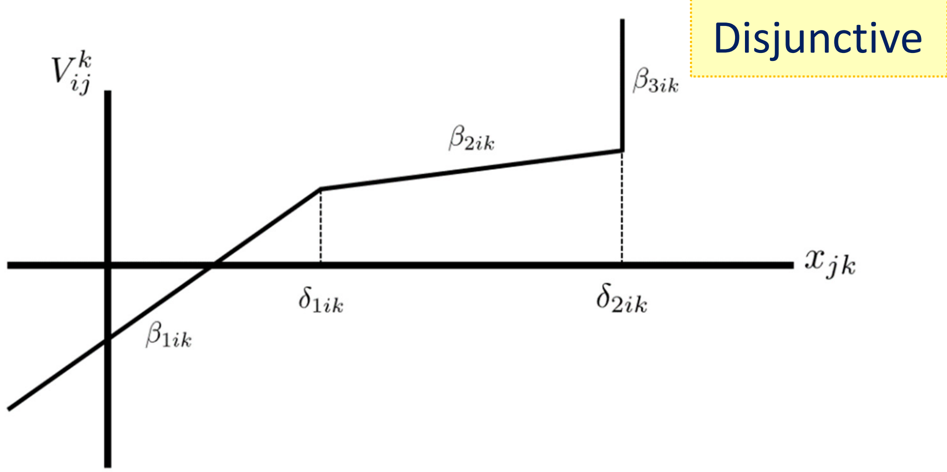


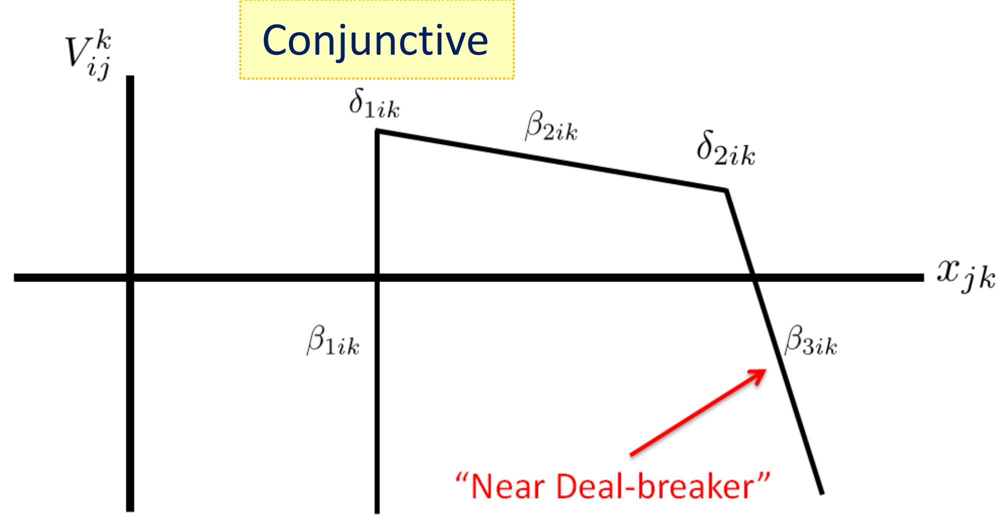
"Deal-breaker" for Age:
Over 40? Unlikely. Under 18? **NEVER!**



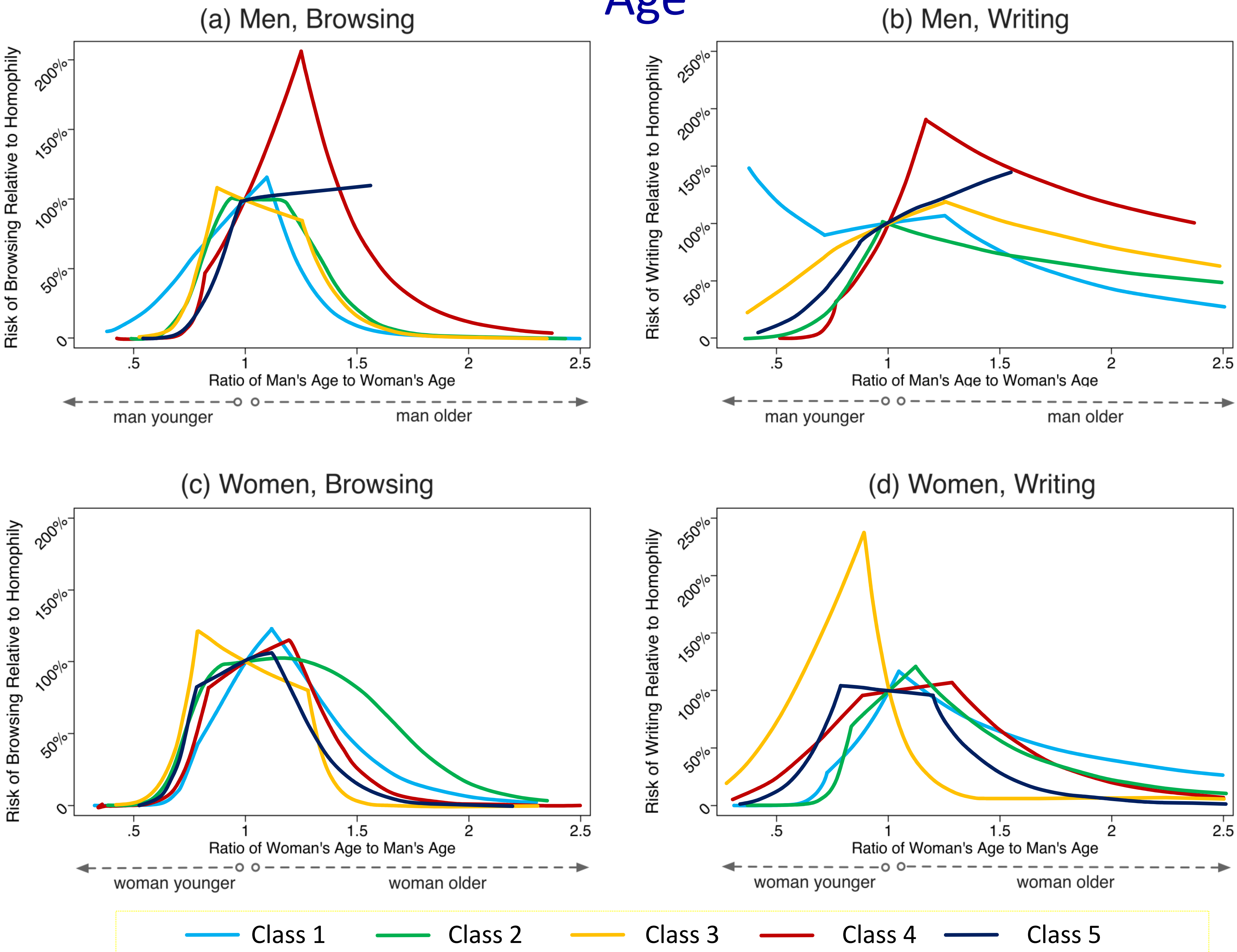
Linear Compensatory, Conjunctive, and Disjunctive Rules… **All from the data!**





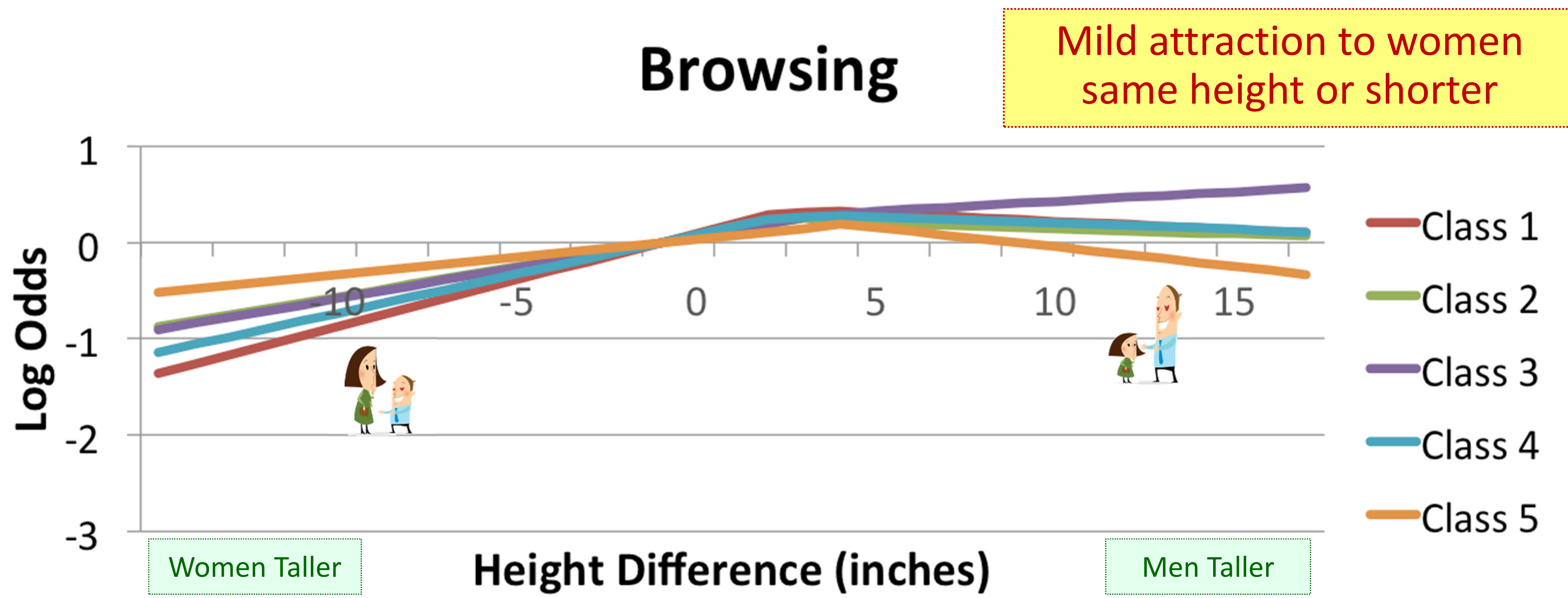


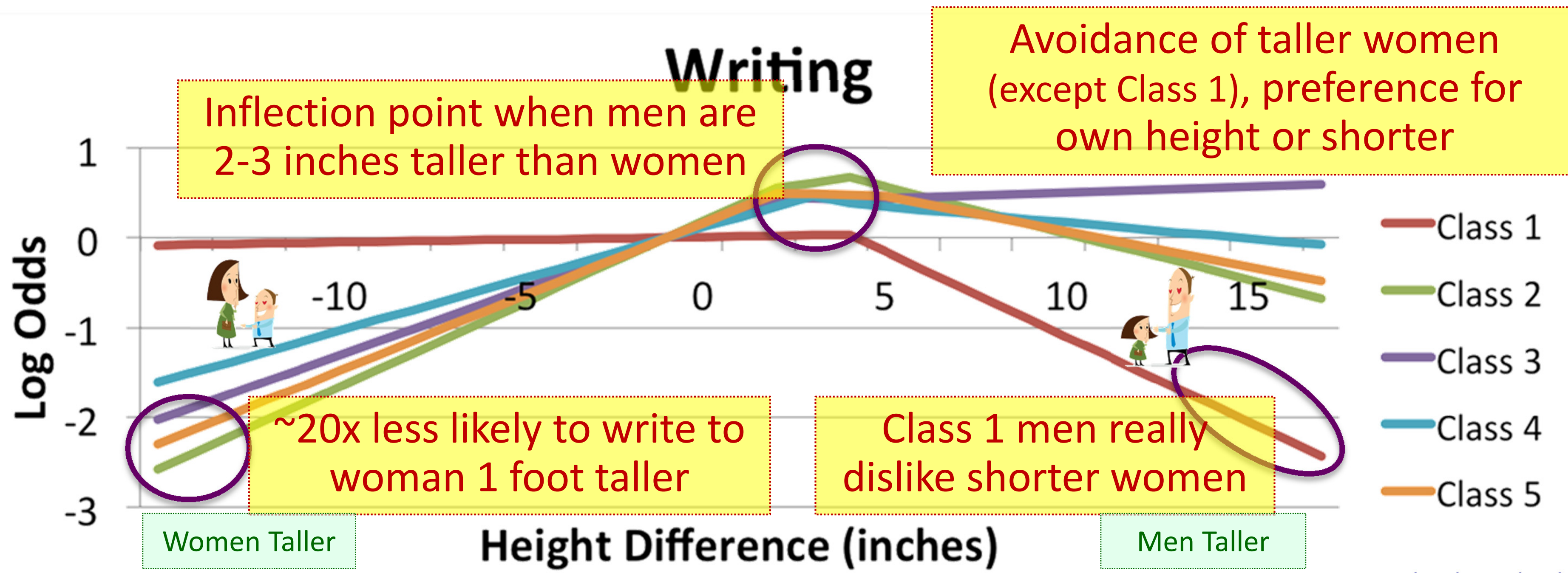
"Age"



(a) Men, Browsing
Risk of Browsing Relative to Homophily
0
50%
100%
150%
200%
.5
1
1.5
2
2.5
Ratio of Man's Age to Woman's Age
man younger
man older
(b) Men, Writing
Risk of Writing Relative to Homophily
0
50%
100%
150%
200%
250%
.5
1
1.5
2
2.5
Ratio of Man's Age to Woman's Age
man younger
man older
(c) Women, Browsing
Risk of Browsing Relative to Homophily
0
50%
100%
150%
200%
.5
1
1.5
2
2.5
Ratio of Woman's Age to Man's Age
woman younger
woman older
(d) Women, Writing
Risk of Writing Relative to Homophily
0
50%
100%
150%
200%
250%
.5
1
1.5
2
2.5
Ratio of Woman's Age to Man's Age
woman younger
woman older
Class 1
Class 2
Class 3
Class 4
Class 5

Height Effects, Men





Tentative General Findings

Group users via **site usage:** M&W each in **5 classes**

Dealbreaker for both Men and Women is… Age

- **Best**: someone near your **own age**
- Men prefer younger; Women somewhat older
- Women over 40 write to much older Men

“No photo”: **20x** less likely to be browsed

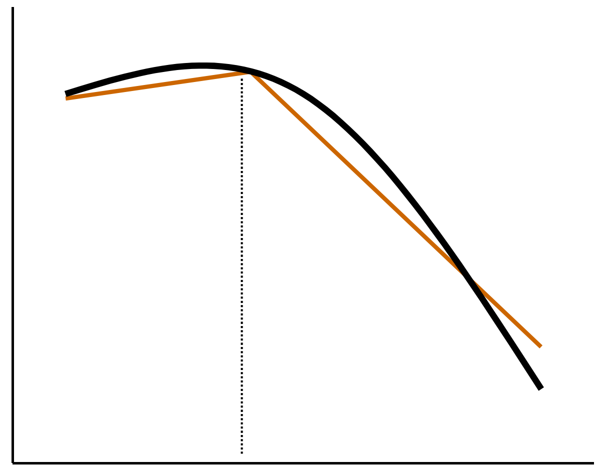
Height preferences vary, but…

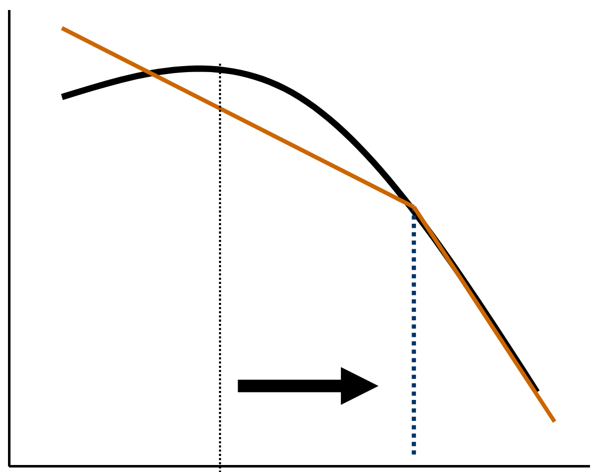
- Taller generally better for men
- **3 inches** minimum gap

[Lots and lots of other findings… read the paper!]

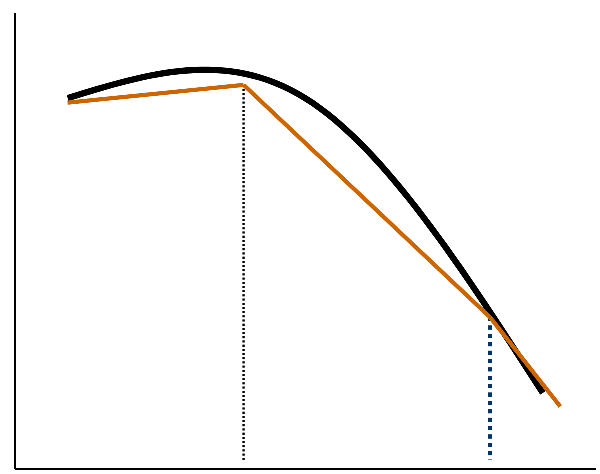
Next Step: **Nonparametric Bayes**

Individual Contours / Nonlinear Utilities





Change in knot location



Change in knot number

Quick Final Points

We are finally seeing a **convergence**:

Bayesian methods to **integrate data sources**

Nonparametrics to **avoid bad assumptions about patterns** and **reliance on linearity**

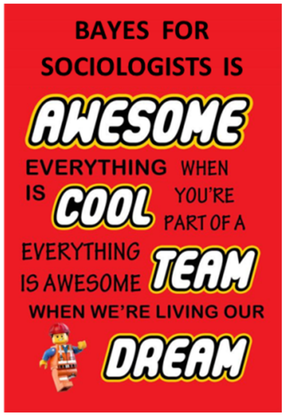
Dynamic models help determine "**did something really important change here**?"

Next 5-10 years: making these **easy to use** for empirical researchers with large data sets

Intrigued / Piqued / Triggered ?

We (EB, FF) are writing a paper and R package on all this **and more**, aimed at "Social Scientists":

- Discrete outcomes (binary, multinomial, ranks, …)
- Multiple stages (e.g., browse then choose)
- Screening / discontinuities (splines; "changepoints")
- "Exploratory behavior" (e.g., just trying it out)
- Dynamics / evolution of behavior



It will be awesome (eventually)

Right now: SAS / STATA have basic Bayes.

STAN gets you started with a fancy / speedy form of Bayes with almost zero technical burden. Totally free; integrates with R (mc-stan.org)

"Questions from Teddy" Redux

~~a) Your background~~

~~b) Your toolkit of computational methods~~

~~c) How you learned this material~~

d) What you are working on

e) Inspirational words of wisdom for beginners!

"What you are working on?"

Tons of stuff: 

Online ad response: Determining the shape of ad response curves [w/ Hernan Bruno, Inyoung Chae]

Data Fusion for Online Promotional Optimization [w/ Longxiu Tian]

Online Dating: Many projects, including language, networks, dyadic choice, "swipe left", …
[w/ Elizabeth Bruch, Jeff Lockhart, Mark Newman, Dan Ariely, Dan Jurafsky…]

Charitable Donations and Scaling: Many projects, in collaboration with Philanthropic organizations in England and France
[w/ Kee Yeun Lee, Jen Shang, Arnaud de Bruyn, Geun Hae Ahn]

Modeling Dishonesty and Data Breaches Online: Uses online dating data from "cheaters" [w/ Bruch, Turjeman]

Credit Score Prediction: Rating consumer credit-worthiness in real-time, using nonparametric Bayes [w/ Linda Salisbury; Longxiu Tian]

Fraud Detection in Medical Claims Data [w/ Jun Li, Dana Turjeman]

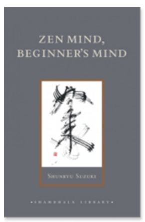
Models of Choice Endogeneity: De-biasing data when we only have data on people who "chose" to provide it [w/ Longxiu Tian]

Consideration set models for auto purchase prediction [w/ Mike Palazzolo]

Interface between Marketing and Engineering Models: Many ongoing projects with Design Science and Mech. Eng.
[w/ Panos Papalambros, Yi Ren, Namwoo Kang]

Bayesian nonparametrics in general [w/ many faculty and students]

“Inspirational words of wisdom for beginners!”



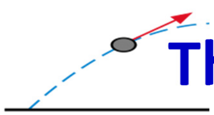
“Let It Be”

Zen Mind, Beginner’s Mind

“Big Data” oversold: quality MUCH more important than quantity

Learn lots of methods, but don’t let them lead you

Put together teams with complementary skills



Think “trajectory”

Work hard early in your career: it will pay you back 1000-fold

Read the best papers, even if they are 40 years old

In the end, you’re only remembered for your best work

Look both ways before crossing ☺

Avoid emoticons

Thank You!

Questions?

Comments?