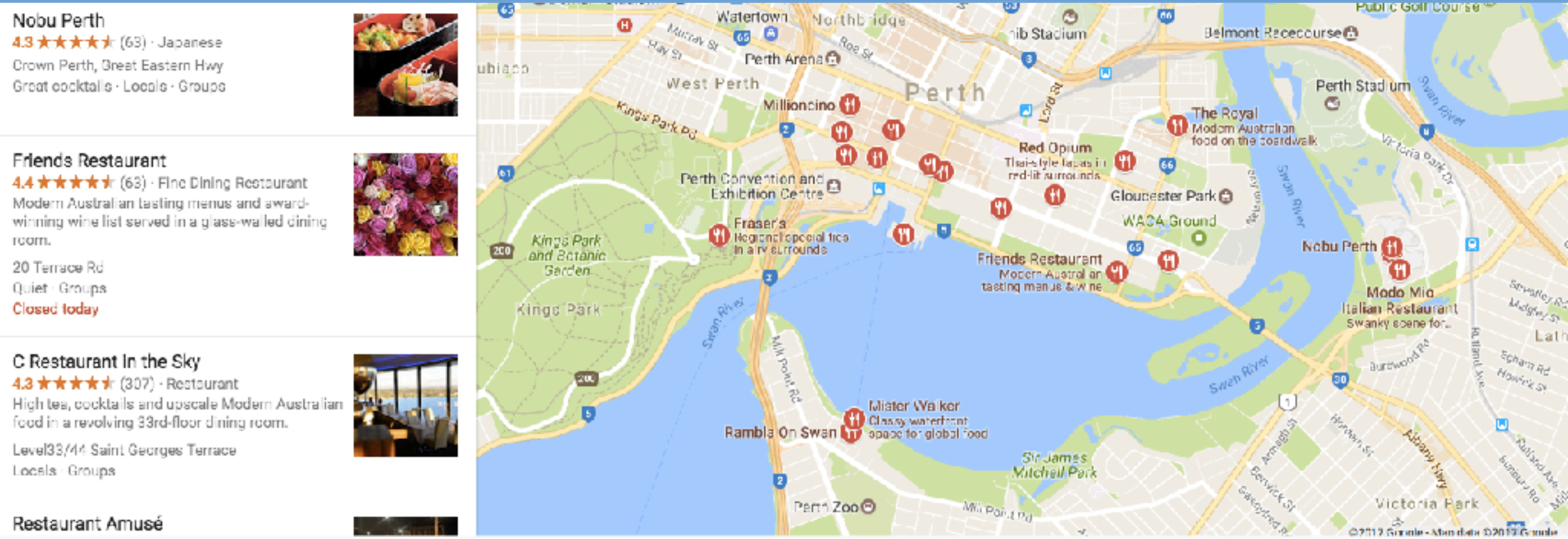
When Big Data Meets Discrete Choice

Andrew Tomkins
Google, Inc.

Parts of this work are joint with Ashton Anderson, Austin Benson, Flavio Chierichetti, Ravi Kumar, Mohammad Mahdian, Bo Pang, and Sergei Vassilvitskii

Example of Discrete Choice: Dinner in Perth



Nobu Perth
4.3 ★★★★★ (63) · Japanese
Crown Perth, Great Eastern Hwy
Great cocktails · Locals · Groups

Friends Restaurant
4.4 ★★★★★ (63) · Fine Dining Restaurant
Modern Australian tasting menus and award-winning wine list served in a glass-walled dining room.
20 Terrace Rd
Quiet · Groups
Closed today

C Restaurant In the Sky
4.3 ★★★★★ (307) · Restaurant
High tea, cocktails and upscale Modern Australian food in a revolving 33rd-floor dining room.
Level33/44 Saint Georges Terrace
Locals · Groups

Restaurant Amusé

Must choose exactly one restaurant for dinner.

Which one to choose?

Example of Stationary Rational Choice

John likes quirky places with great food and authentic atmosphere, mostly vegetarian, but sometimes eats fish if it's a specialty.

Values per place: $\langle \underline{7,} 3, \underline{1,} \underline{3,} 14, 6, 23\rangle$

Beth eats mostly fast food, likes burgers, doesn't enjoy "fancy" food. Sometimes like bagel sandwiches.

Values per place: $\langle \underline{3,} 5, \underline{11,} \underline{16,} 14, 12, 2\rangle$

Steve appreciates good decor, great service, and fun fusion menus. He enjoys all Asian cuisines, especially Thai and Cambodian.

Values per place: $\langle \underline{9,} 1, \underline{1,} \underline{4,} 14, 17, 6\rangle$

Modeling User Preferences

Basic model:

Each user has vector of values for each place

User then selects available alternative of highest value

Many reasons this basic model doesn’t hold

1. Factors beyond the item may influence decision

   Example: User’s choice may be influenced by presentation order

2. Even if rational, choice behavior may not be stationary

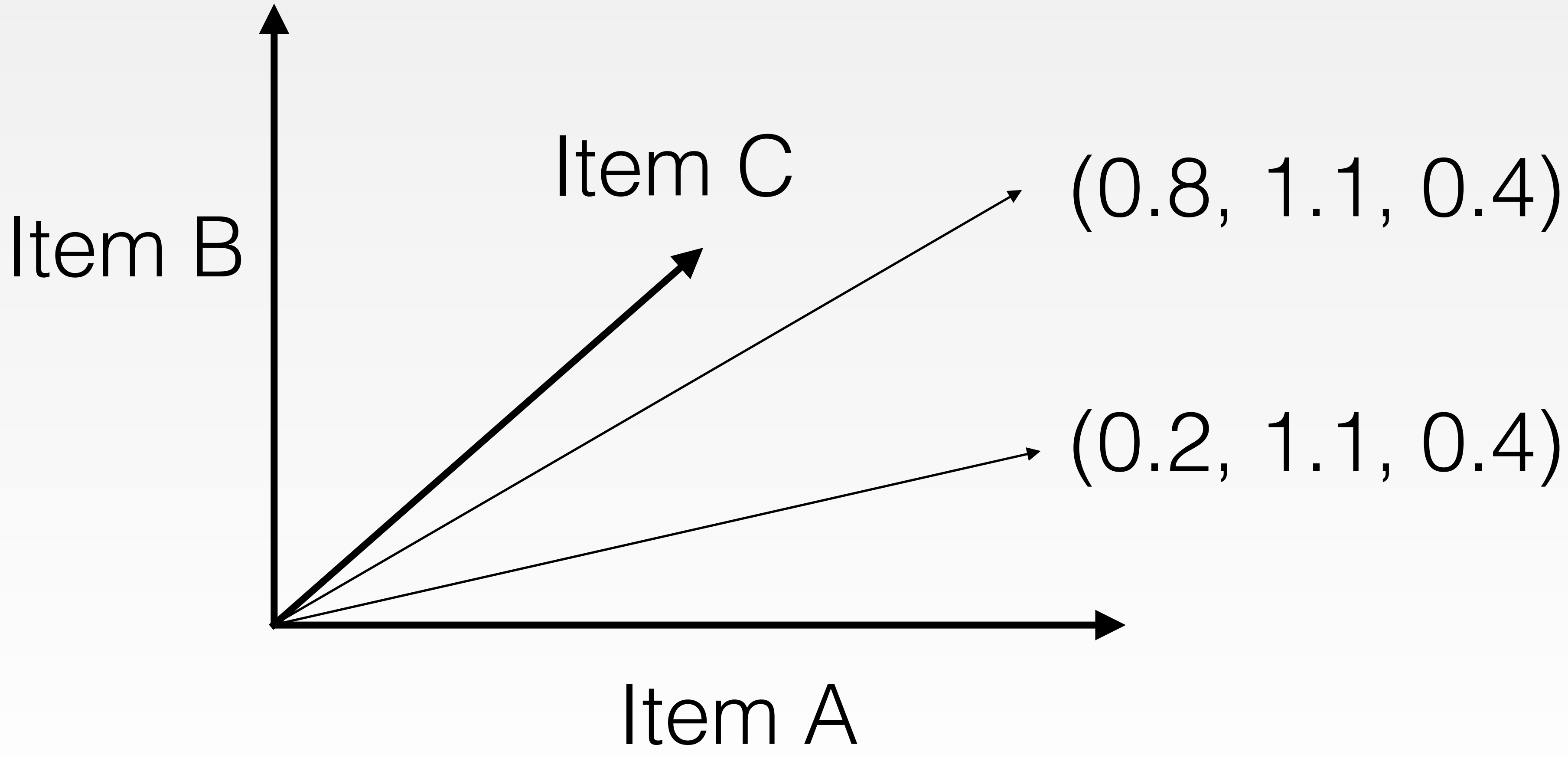
   Example: User may be “in the mood” for Chinese food, or bored with a favorite option

This talk focuses on rational choice — even here behavior is surprisingly rich!

Modeling General Behavior of Discrete Choice

Standard model: Randomized Utility Model (RUM):

- Utilities for a user are drawn from a distribution



User selects the item of max utility from the slate

Discrete Choice: The Technical Problem

Let $\mathcal{X}$ be the universe of items.

Let $X \subseteq \mathcal{X}$ be a slate of choices.

Output a distribution $D_u(X)$ over elements $x \in X$ indicating u's likelihood to choose x.

Simplest Setting: Global “Quality” Scores

In general, joint distribution of utility is arbitrary

So let’s simplify. . . .

Assume each alternative j has absolute “quality” V_j

Each user n deviates from this by some random noise ϵ_{nj}.

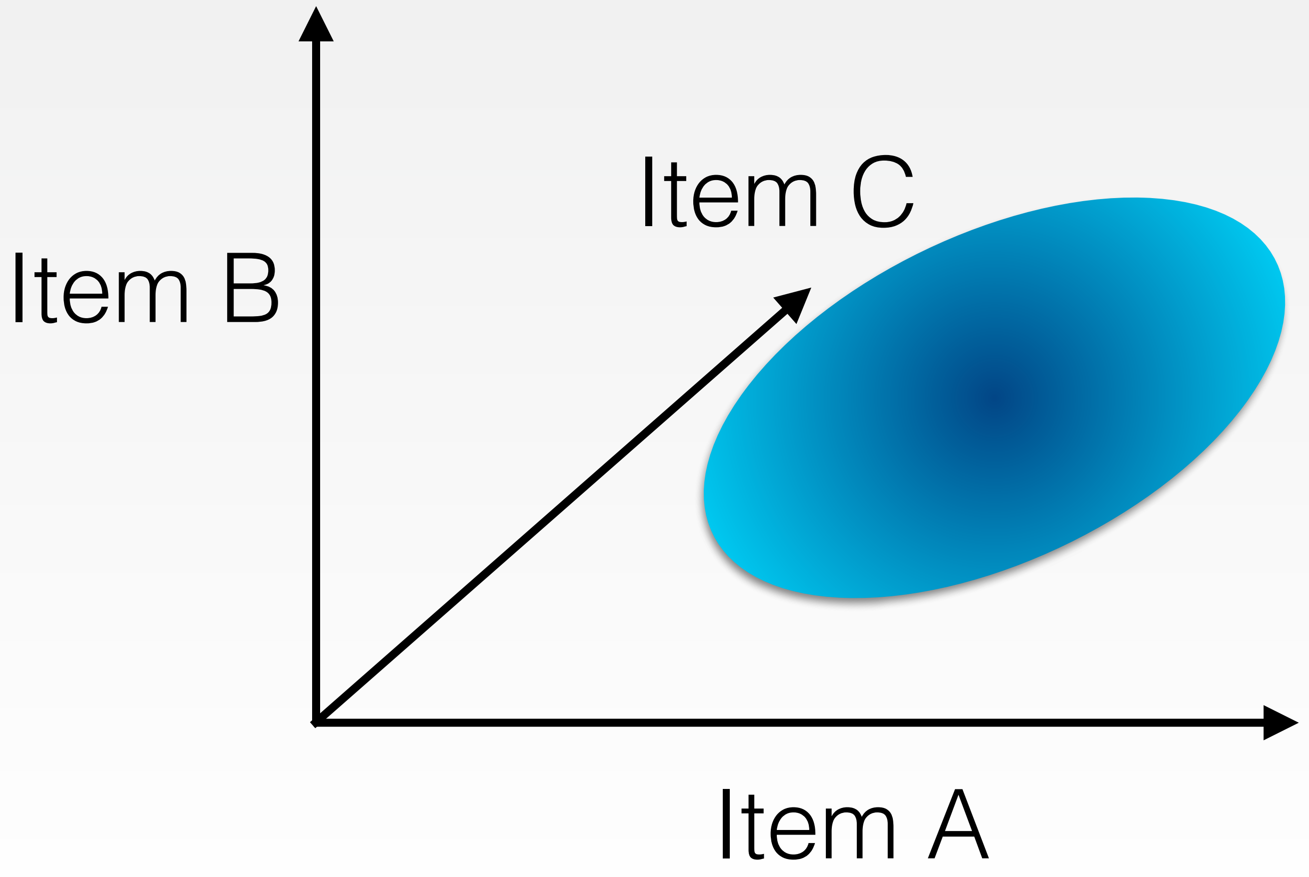
So user’s actualy utility is $U_{nj} = V_j + \epsilon_{nj}$

So $\Pr[\text{choose } j] = \Pr[\forall k : V_j + \epsilon_{nj} > V_k + \epsilon_{nk}]$

Simplifying more, let ϵ_{nj}’s all be chosen i.i.d.

This way, maybe the choice probability has a closed form

New Utility Distribution



A Convenient Choice of Noise Distribution

Let's carefully choose a distribution for ϵ:

$p(\epsilon) = e^{-\epsilon} e^{-e^{-\epsilon}}$

Conveniently, this gives: $\Pr[\text{choose } j] \propto e^{V_j}$

Probability of V_1 "bypassing" V_2 is $e^{V_1} / (e^{V_1} + e^{V_2})$

This is identical to multinomial logistic regression!

Multinomial regression gives identical choice probabilities to RUM with Gumbel-distributed noise!

Including Features in Choice

Recall we simplified to global qualities V_j

Let's remove this, and let V_{nj} be user-specific.

In choice, often $V_{nj} = w^T x_{nj}$ is linear in features

Again: $\Pr[\text{choose } j] \propto e^{V_{nj}} = e^{w^T x_{nj}}$

Two identical formulations:

Choice MNL: $\Pr[\text{choose } j] \propto e^{w^T x_{nj}}$

ML multinomial: $\Pr[\text{choose } j] \propto e^{w_j^T x_n}$

Understanding This Approach

This is “Multinomial Logit” (MNL), the most common approach to Discrete Choice

1. Convex formulation, easily optimized, highly scalable
2. Natural extension of logistic regression — incorporates normalization in the denominator
3. MNL captures one form of stationary rational choice.

Why Not Use This Model?

What restrictions does Multinomial Logit impose?

For convenience, let $w_a = e^{w^T x_{na}}$ for fixed user n

$$\text{Options:}\{a, b\} \qquad Pr[a|a \text{ or } b] = \frac{w_a}{w_a + w_b}$$

$$\text{Options:}\{a, b, c\} \quad Pr[a|a \text{ or } b] = \frac{w_a}{w_a + w_b}$$

Relative likelihood of *a* versus *b* does not depend on other alternatives. Choices are *Independent of Irrelevant Alternatives*, or “IIA.” Also called *Luce’s Axiom of Choice*.

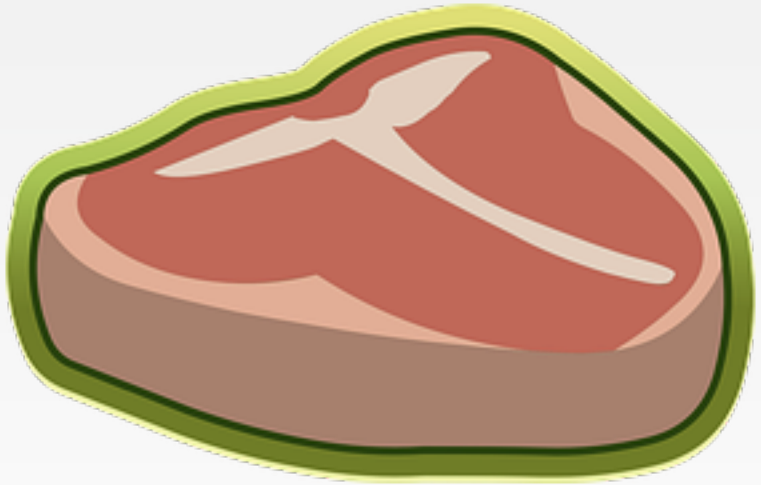
In fact, MNL == choice with IIA.

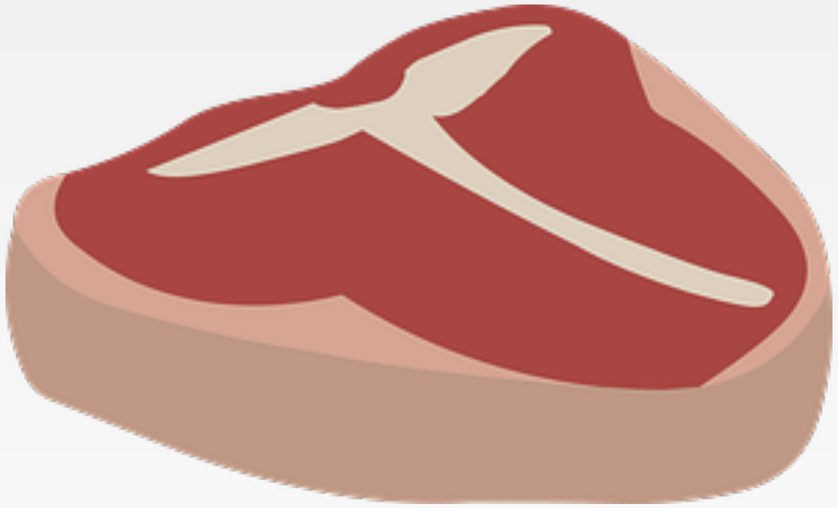
Luce’s Axiom of Choice

$Pr[a|a \text{ or } b]$ does not change when c is added to slate

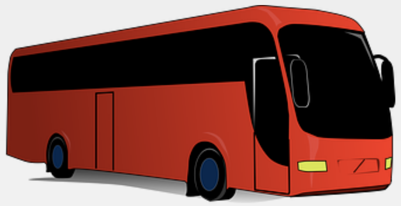
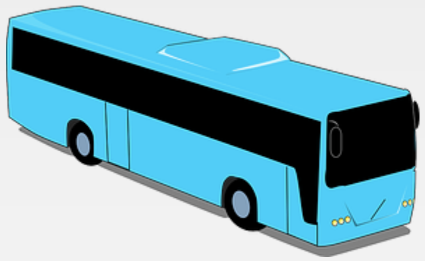
Luce’s Axiom of Choice

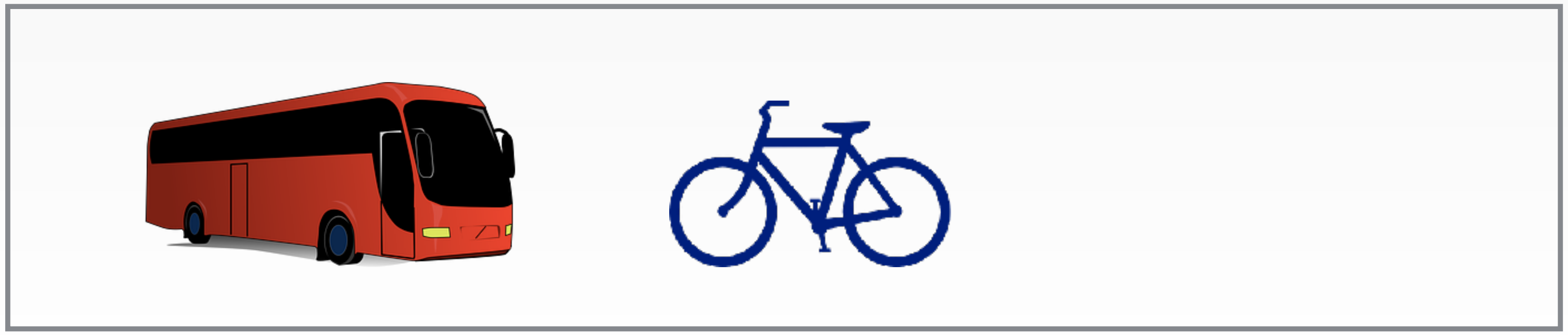
$Pr[a|a \text{ or } b]$ does not change when c is added to slate



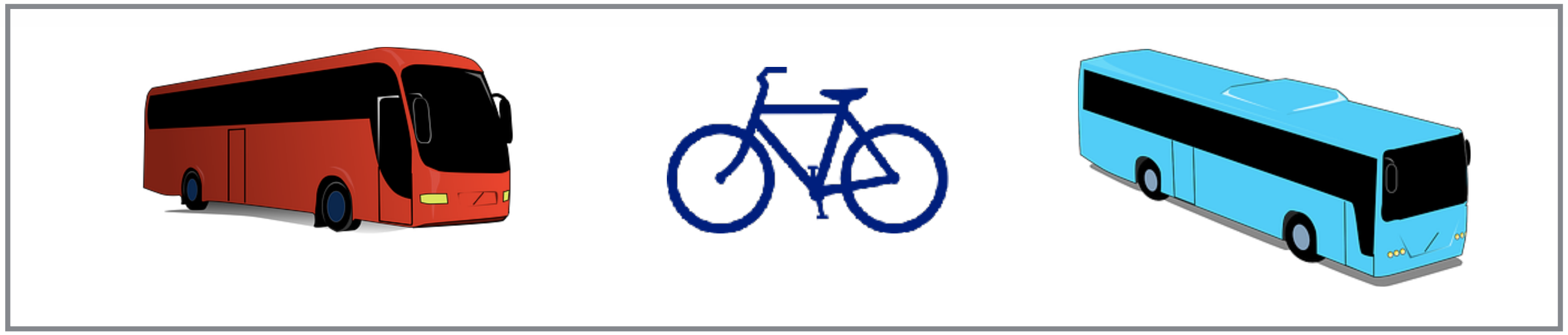


Stationary Rational Choice May Not Follow IIA

			
User Type 1: 50%	5	100	15
User Type 2: 25%	100	1	75
User Type 3: 25%	75	1	100

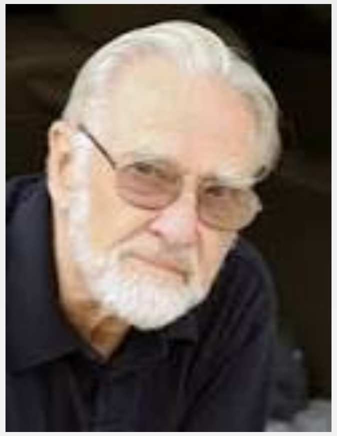


50/50 split



25/50/25 split

A Brief History of IIA



R. Duncan Luce
1925 — 2012

R. Duncan Luce: Mathematical Psychologist

(and “coiner” of the term *clique* for complete subgraph)

Formulated “Luce’s Axiom of Choice” in 1959 (IIA)

But earlier work had set the stage:

In 1951, Kenneth Arrow proved his Impossibility Theorem showing that IIA was one of several mutually incompatible properties of a social choice function.

By 1952, Bradley and Terry had introduced a 2-element variant of the choice model (also studied by Zermelo earlier)

—> multiple choice model often called “BTL” model

Over the two decades after Luce(59), many authors, notably Daniel McFadden, completed the story extending BTL to MNL.

What if IIA is violated?

Situation is much more complex….

2000 Nobel Prize to Daniel McFadden for:

> “development of theory and methods for analyzing discrete choice”

Most powerful models are:

- Mathematically complex
- Computationally intractable
- Require sophisticated external representations of dependence

Practitioners with non-IIA data typically employ “Nested Logit”

Rich body of algorithms in this area, but not fully integrated into computer scientist toolkit

So…Does IIA Hold?

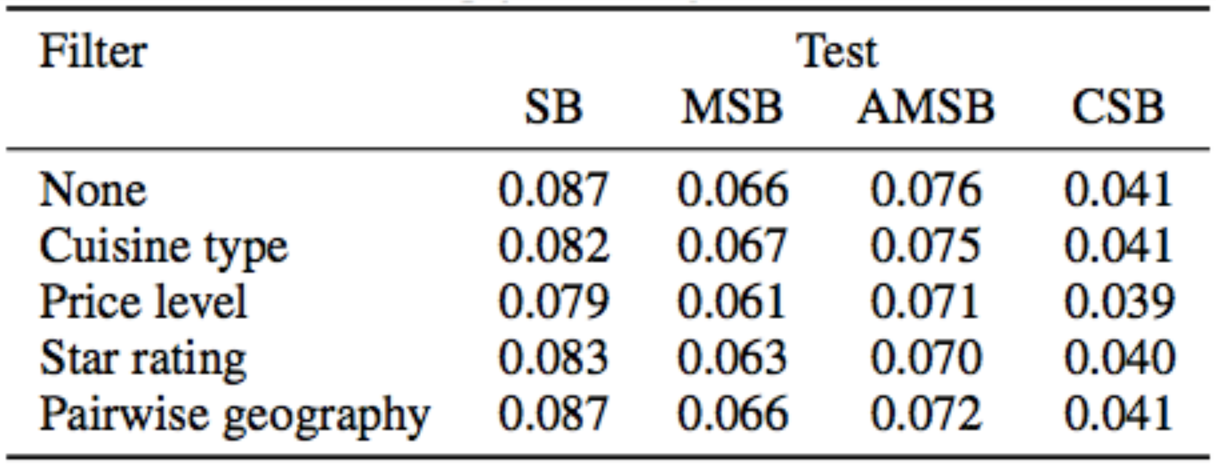
Recent results say:

> “Sometimes yes, sometimes no”

Daniel McFadden [McFadden 1977] says:

> “…it is clear from many experiments that the conditions under which the choice axiom holds are surely delicate.”

Example. Given click data on restaurants, with different slates of choices and conditioned on a single click:



Filter	Test			
	SB	MSB	AMSB	CSB
None	0.087	0.066	0.076	0.041
Cuisine type	0.082	0.067	0.075	0.041
Price level	0.079	0.061	0.071	0.039
Star rating	0.083	0.063	0.070	0.040
Pairwise geography	0.087	0.066	0.072	0.041

The Talk So Far

RUM

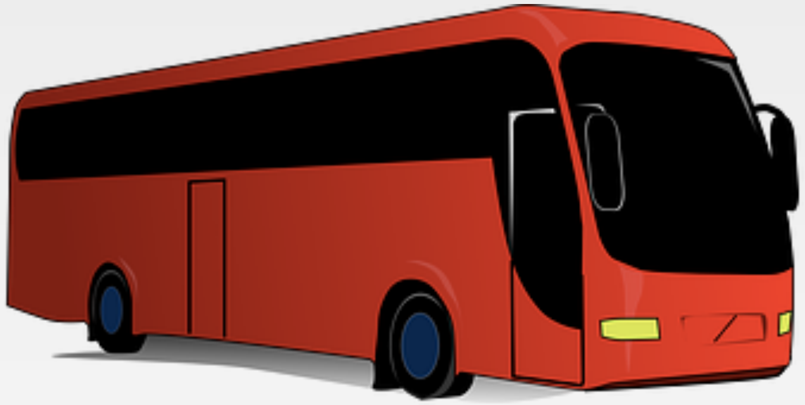
- General approach to characterize choice
- Any RUM approximated by a Mixed Logit
- Model difficult to learn

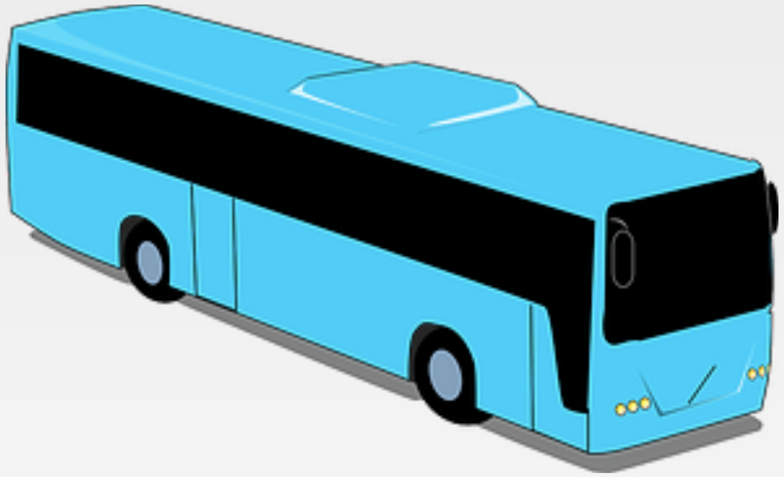
MNL

- Captures only RUMs with IIA
- Easy and fast to optimize
- Easy to interpret

Is there anything in-between?

Problems With IIA Revisited





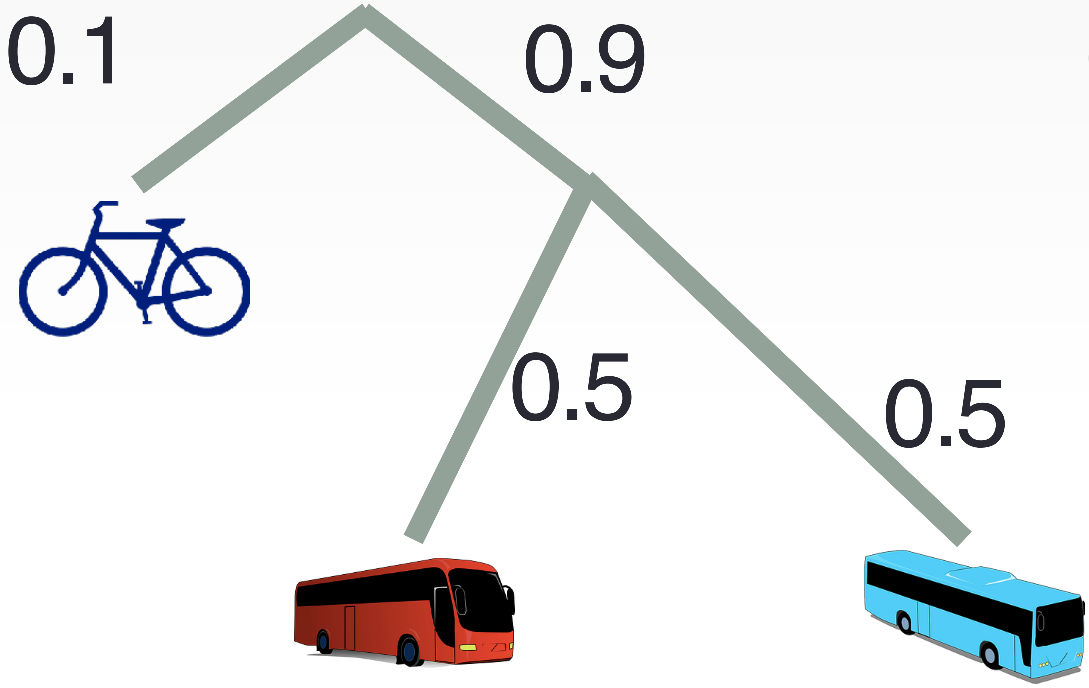
0.1

~~0.9~~ 0.45

0.45

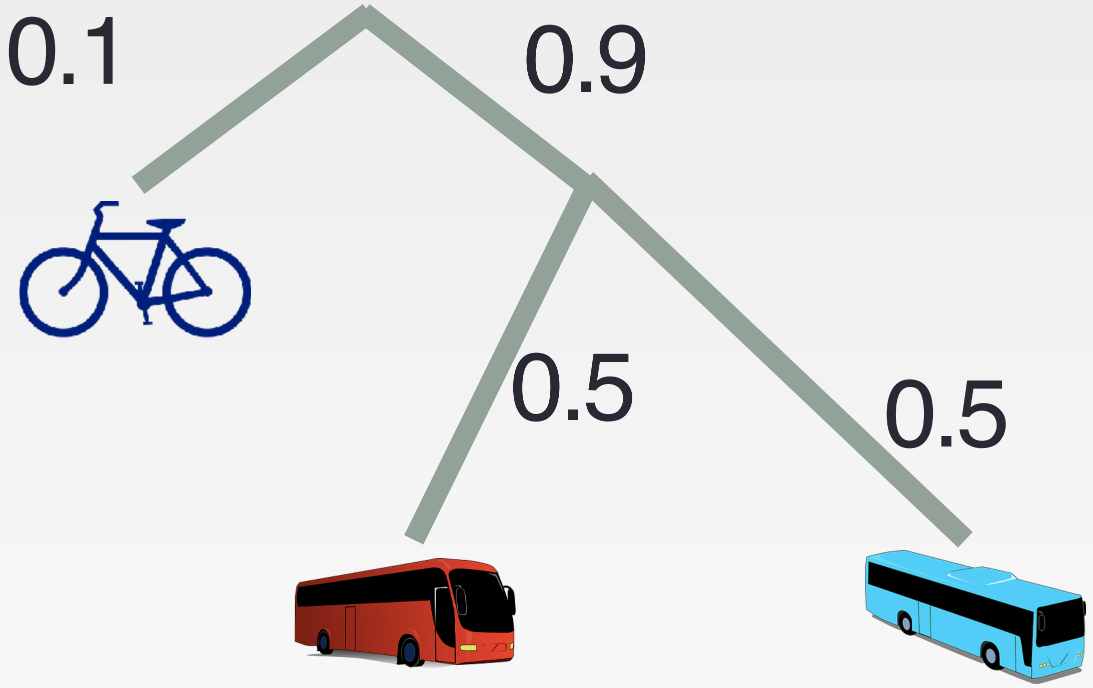
Likelihood(Bus) / Likelihood(Bike) = ~~0.9 / 0.1~~

0.45 / 0.1



Nested logit

Modeling the decision as a tree is a *nested*, *sequential*, or *hierarchical* logit model. It looks like a sequence of multinomial logits. [McFadden 78]



Nested Logit Connections to RUM and MNL

Model behavior: Nested Logit selects an item by traversing tree from root, applying MNL at each level

Casting NL as RUM:

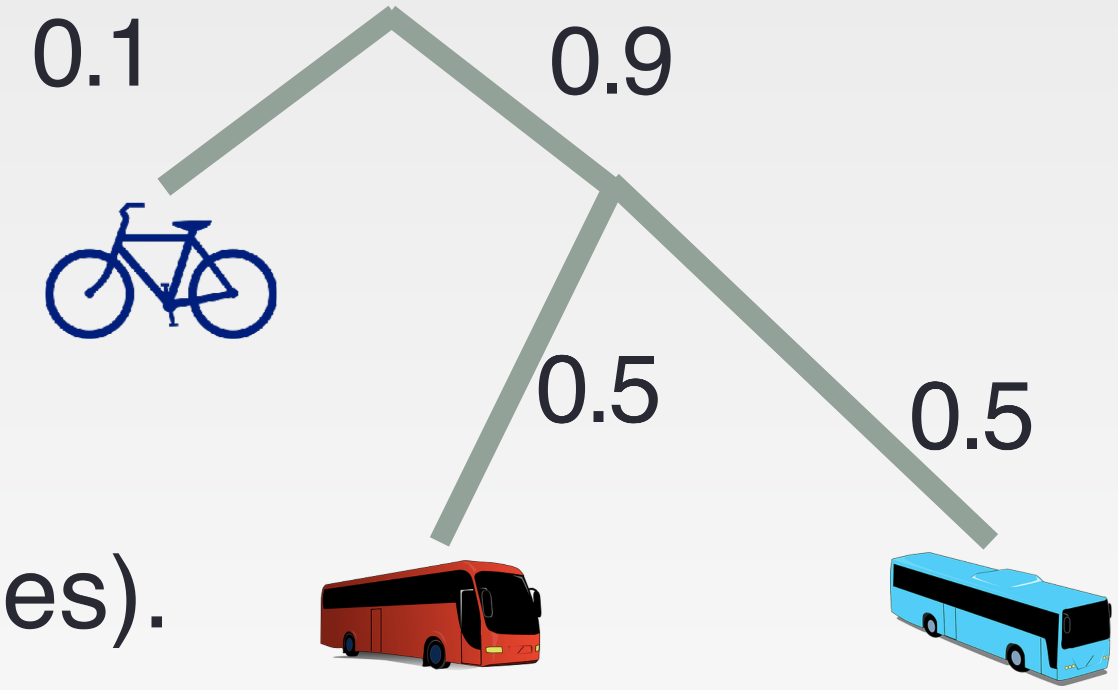
- Utility of each item is fixed in advance.
- Each user's utilities are perturbed slightly from the underlying values
- Perturbation for each item is drawn from specific joint distribution

Power of NL:

- Pros: Captures hierarchical cannibalization cleanly; generalizes MNL
- Cons: Choices must separate cleanly into nests

Problem: Can we solve NL efficiently?

Nested logit



The current approach:

1. Propose tree from intuition.
2. Train model (find probabilities).
3. Measure likelihood or test for significant improvement over multinomial logit. [Hausman & McFadden 84, Small & Hsiao 85]

Problem: does not scale to modern datasets with many different choice sets and alternatives.

Our contributions

1. New algorithm for automatically constructing the nested logit tree under an oracle model.

   Provably optimal in this setting!

2. Modified algorithm for real-world, sparse datasets.

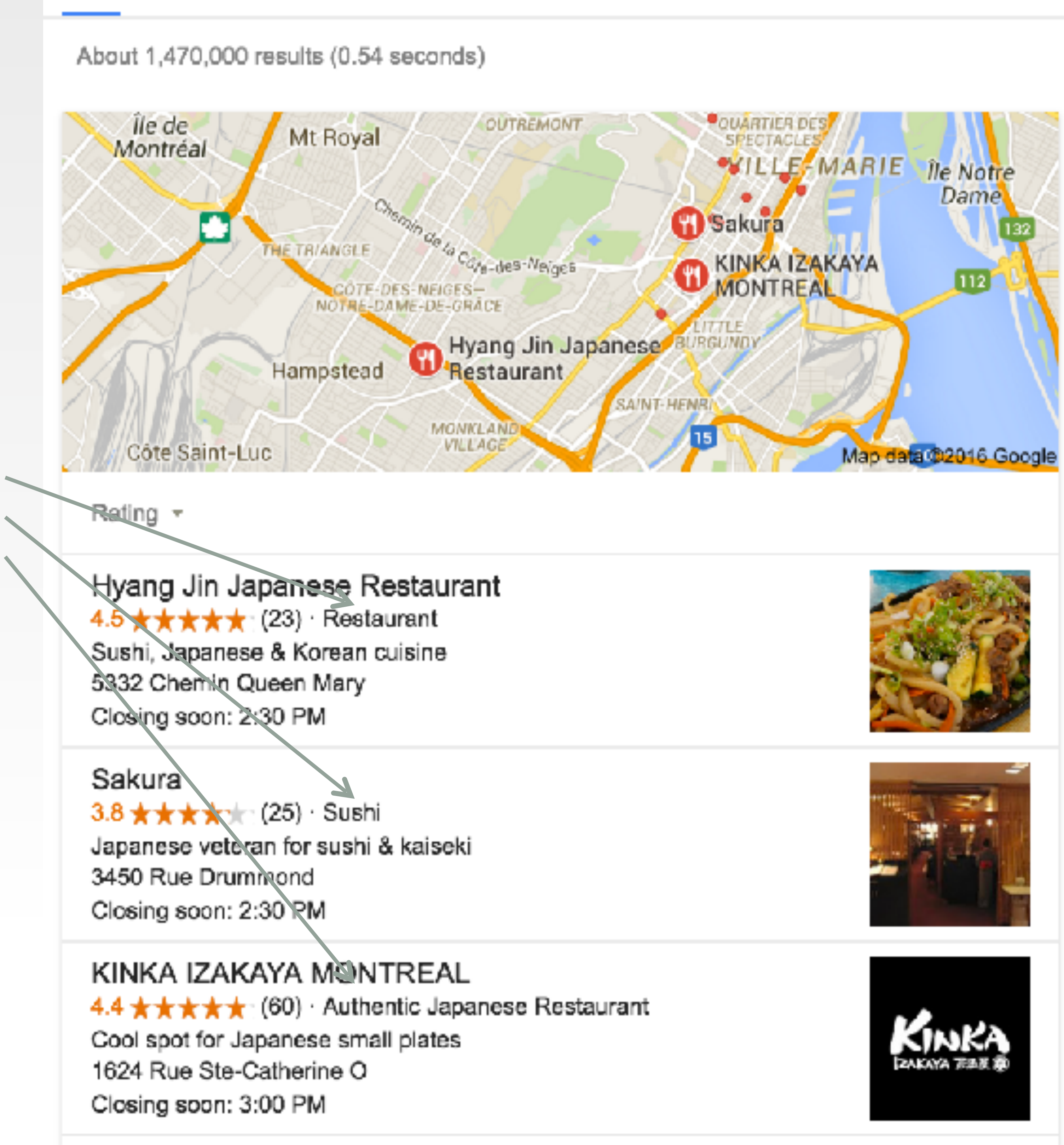
   Restaurant ads, music taste, transportation choices.

3. Several statistical tests to measure the number of IIA violations in choice datasets.

Application I: Japanese food

Choice sets: Japanese restaurant types that appear in search

Selections: clicks



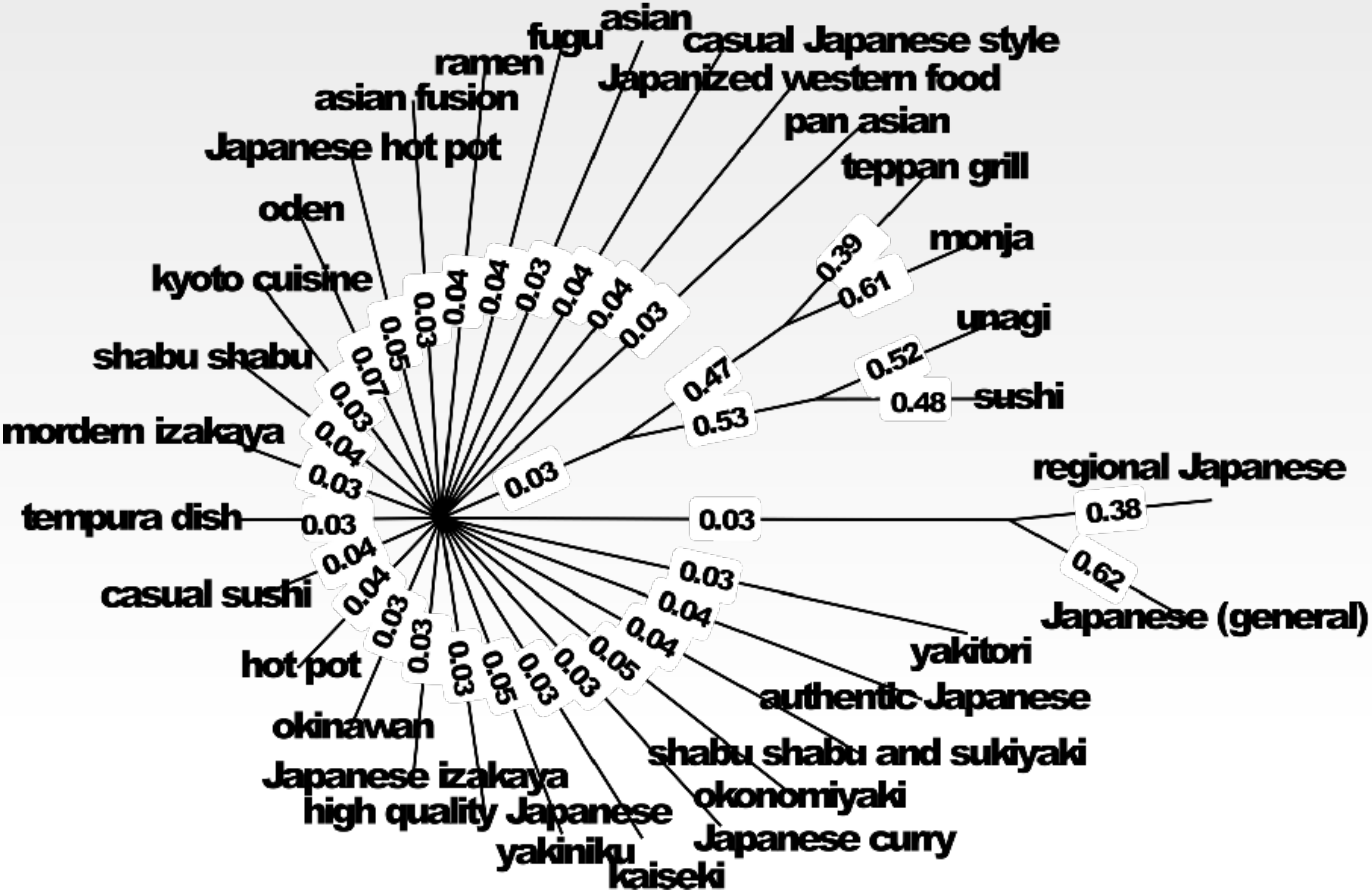
About 1,470,000 results (0.54 seconds)

Hyang Jin Japanese Restaurant
4.5 ★★★★★ (23) · Restaurant
Sushi, Japanese & Korean cuisine
5332 Chemin Queen Mary
Closing soon: 2:30 PM

Sakura
3.8 ★★★★☆ (25) · Sushi
Japanese veteran for sushi & kaiseki
3450 Rue Drummond
Closing soon: 2:30 PM

KINKA IZAKAYA MONTREAL
4.4 ★★★★★ (60) · Authentic Japanese Restaurant
Cool spot for Japanese small plates
1624 Rue Ste-Catherine O
Closing soon: 3:00 PM

Application I: Japanese food



Application II: music choice

Choice sets: last 3 genres played on last.fm, accounting for position

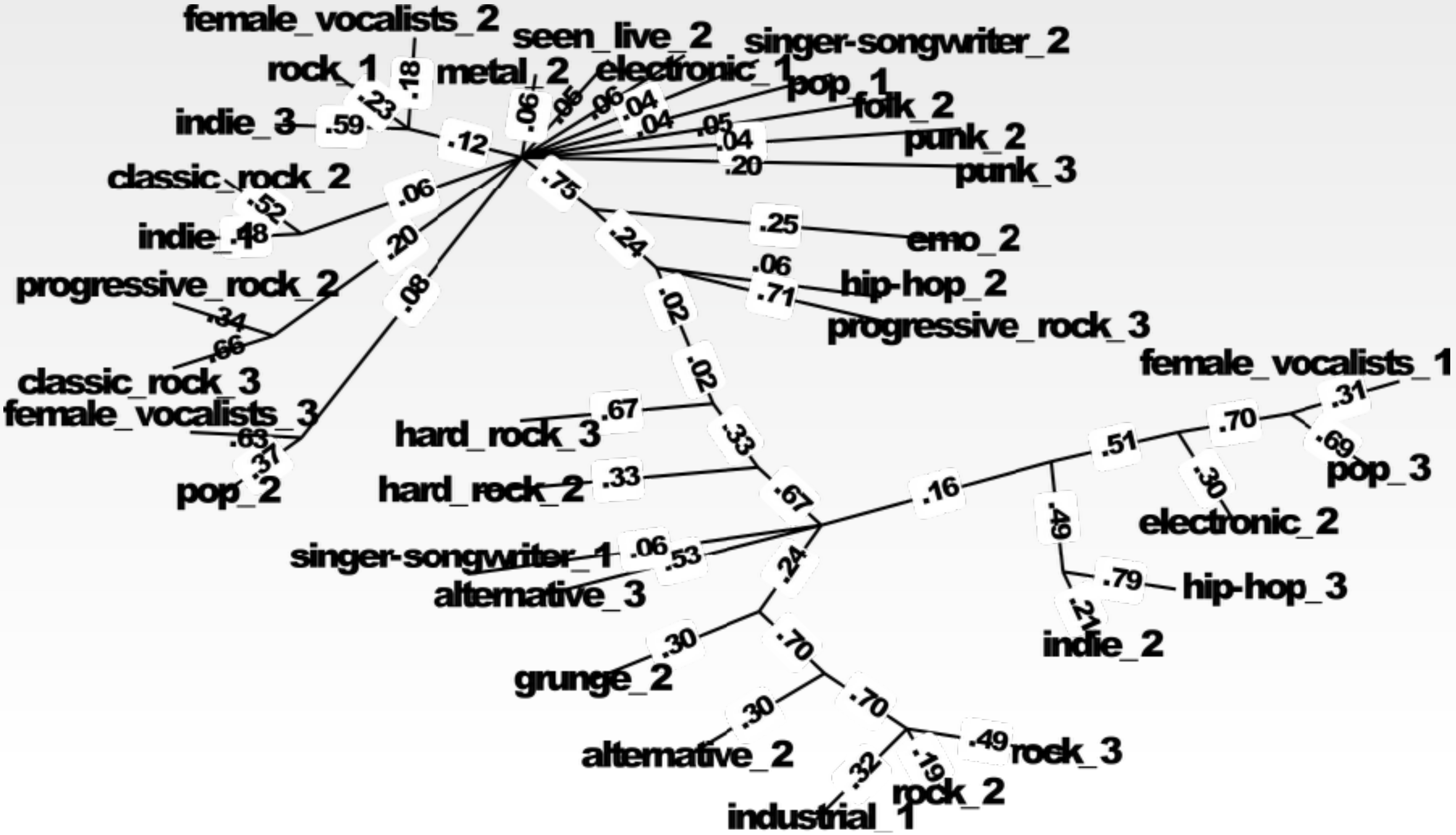
female-vocalist_3

pop_2

hip-hop_1

Selection: next genre played (if from last 3)

Application II: music choice



Filling Out the Discrete Choice Landscape

So far, we've seen:

Families of decision models:

- General choice: any function from slate —> choice distribution
- RUM: only functions consistent with joint utility distribution

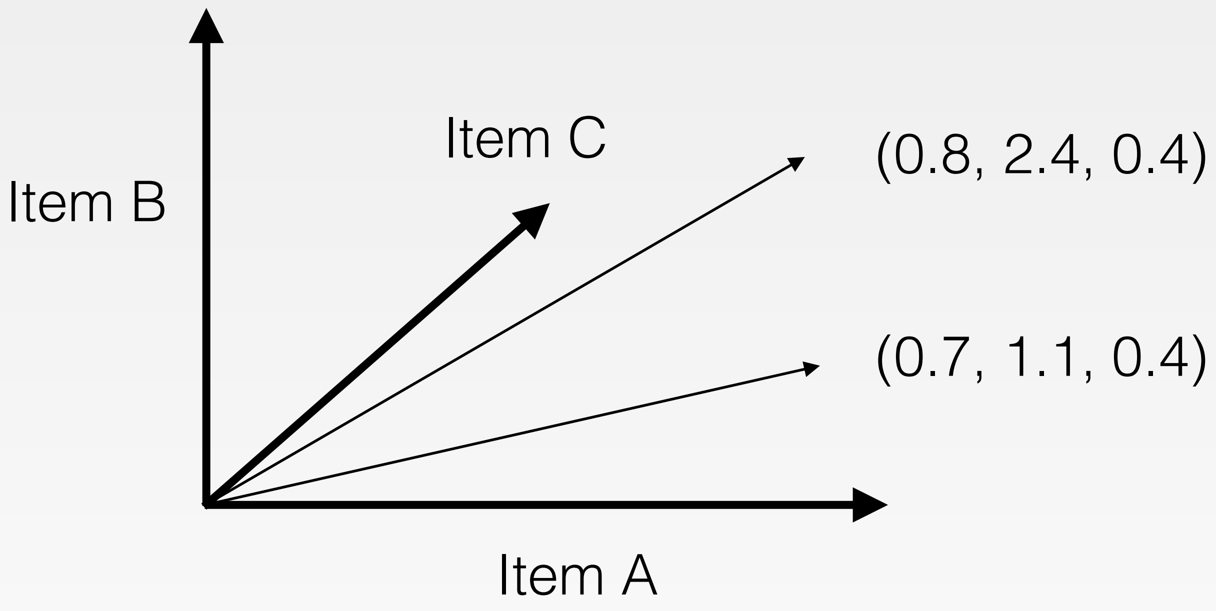
Classes of choice models

- NL: nested multinomial logit — many extensions
- MNL: single multinomials, consistent with Luce's Axiom

Are there more powerful classes of choice models?

Mixed Logit Models

Let’s revisit RUM to form one last choice model



Permutation: $[a, c, b]$

Approximation: $w_a \gg w_c \gg w_b$

Discussion:

1. Only order of utilities matters
2. RUM == distribution over permutations
3. One MNL can approximate any one permutation
4. Mixture of MNLs can approximate mixture of permutations
5. Called Mixed Logit (ML)

ML has many nice properties, including arbitrary substitution patterns

Captures (approximately) every randomized utility model

Summary and Open Questions

Summary:

$$\text{Full choice} \gg \text{RUM} \approx \text{ML} \gg \text{NL} > \text{MNL}$$

General problem: Which classes of RUM can be learned?

- From an oracle
- From a distribution

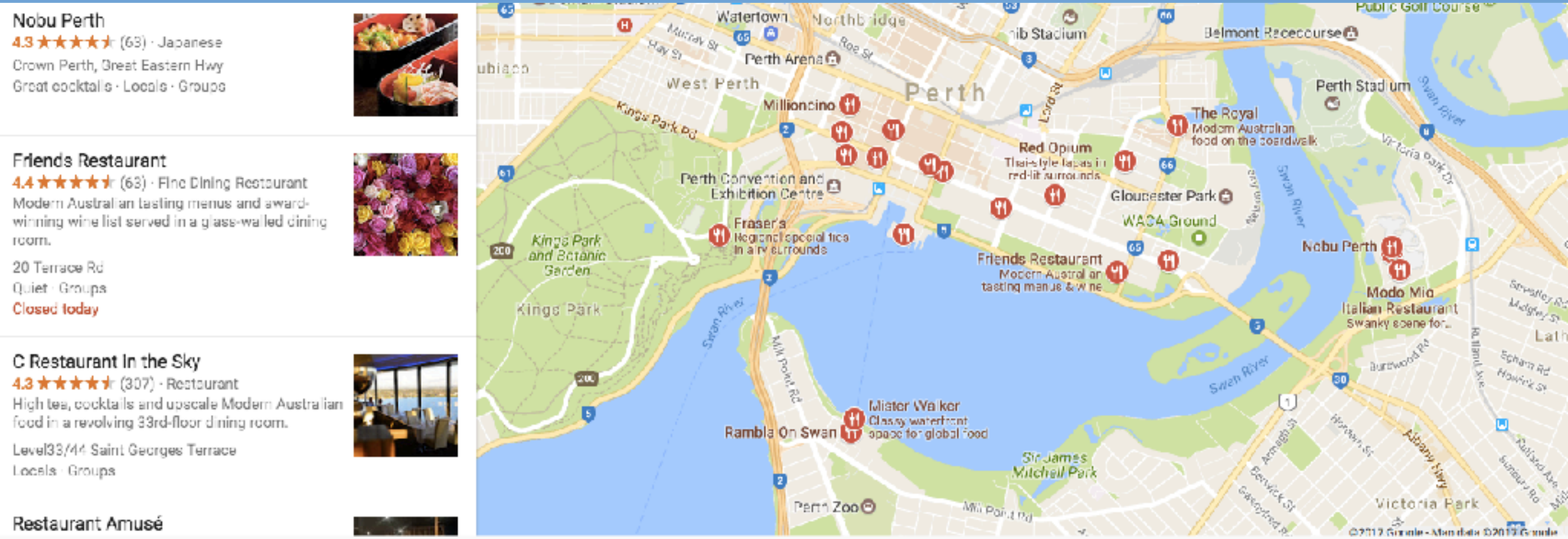
Model extensions:

- Is it possible to learn with noise?
- Are there meaningful personalized models?
- Are there meaningful models that account for well-known irrationalities?

Application 1: Geographic Choice

(or: where should we have dinner tonight?)

Where shall we eat tonight, revisited….



Nobu Perth
4.3 ★★★★★ (63) · Japanese
Crown Perth, Great Eastern Hwy
Great cocktails · Locals · Groups

Friends Restaurant
4.4 ★★★★★ (63) · Fine Dining Restaurant
Modern Australian tasting menus and award-winning wine list served in a glass-walled dining room.
20 Terrace Rd
Quiet · Groups
Closed today

C Restaurant In the Sky
4.3 ★★★★★ (307) · Restaurant
High tea, cocktails and upscale Modern Australian food in a revolving 33rd-floor dining room.
Level33/44 Saint Georges Terrace
Locals · Groups

Restaurant Amusé

Some Factors in Restaurant Choice

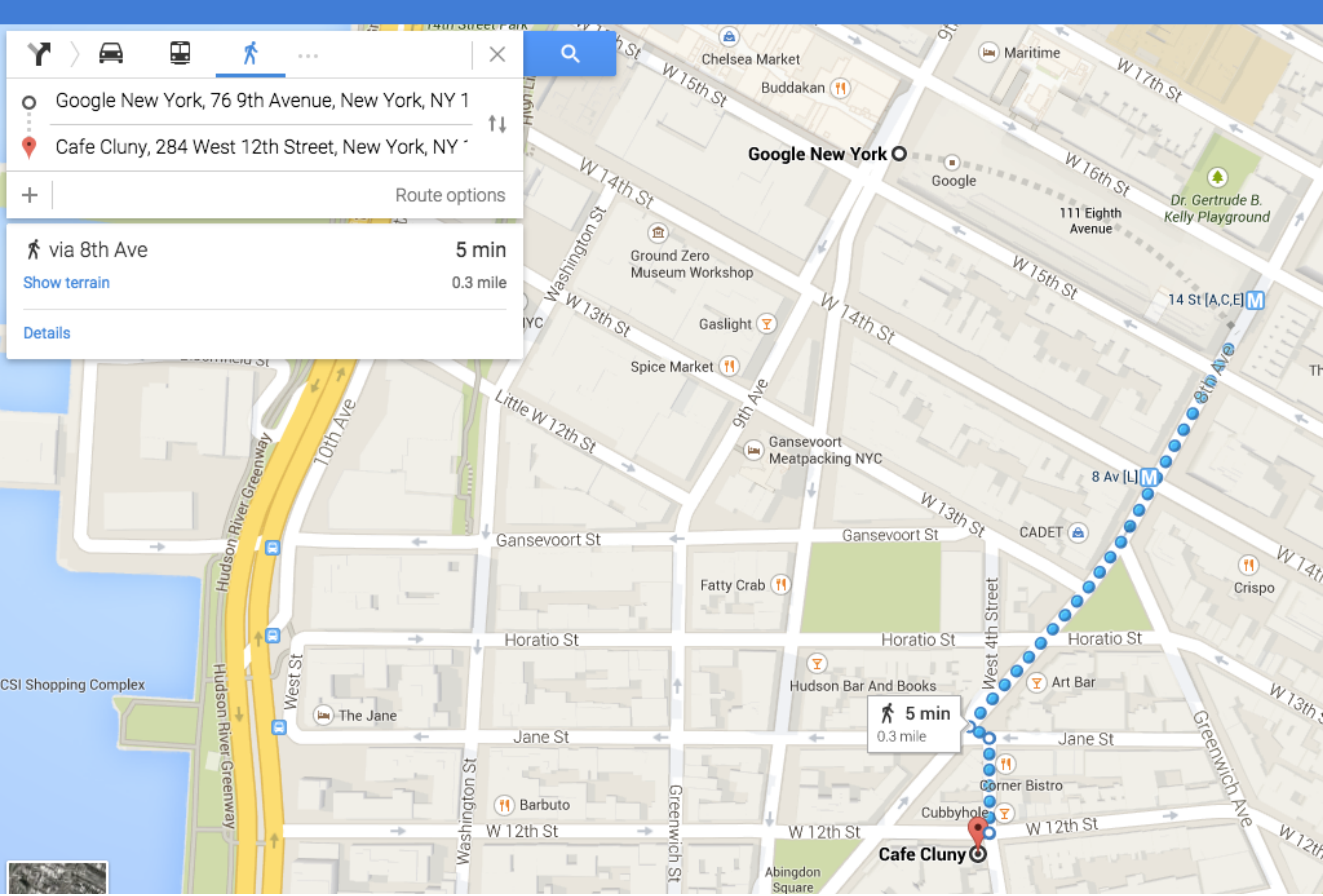
Deciding where to go for dinner:

- Quality of the restaurant
- Distance to the restaurant
- Price
- Cuisine type
- Time since last visit
- Opinions of dining companion(s)
- ...

Dataset

Directions queries:

- Number of directions queries to US/Canadian restaurants in Google Maps
- Random sample of 15.5M queries to ~400K restaurants



Google New York, 76 9th Avenue, New York, NY 1

Cafe Cluny, 284 West 12th Street, New York, NY 1

Route options

via 8th Ave 5 min

Show terrain 0.3 mile

Details

Dataset

Directions queries:

- – Number of directions queries to a US/Canadian restaurants in Google Maps
- – Random sample of 15.5M queries to ~400K restaurants

Caveats:

- – Not all visits have an associated directions search
  - Familiar locations
  - Spontaneous decisions
- – Not all searches result in visits
  - Aspirational searches
  - Traffic & time estimates

Classical Discrete Choice Models

Recall our basic discrete choice model:

- Assign a score to each alternative
- Select with probability proportional to score

$$\Pr[x|A] = \frac{w_x}{\sum_{y \in A} w_y}; w_x = e^{V_x}$$

Goal:

- Better understand the score

Score function

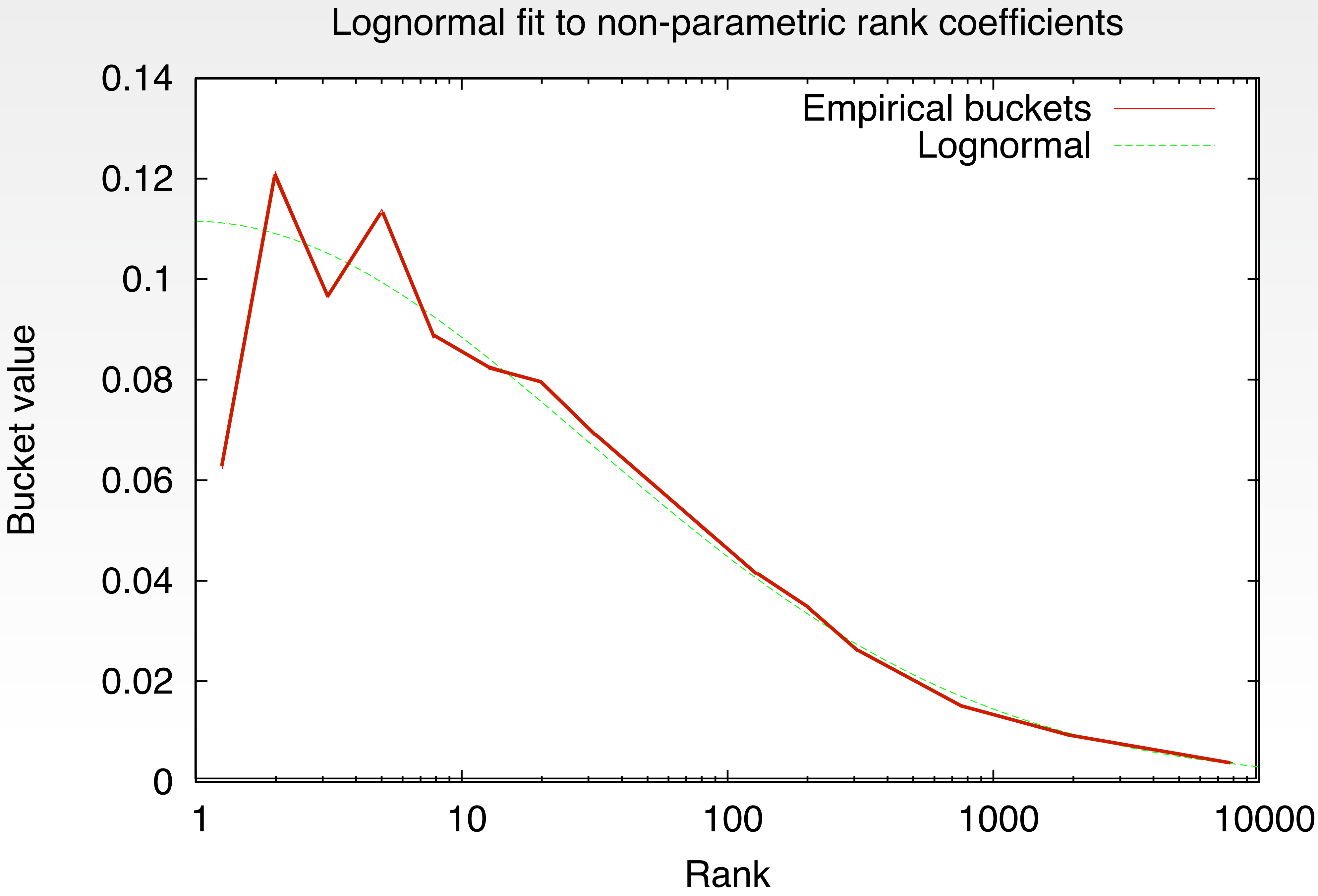
Today:

- Distance to the restaurant d
- Number of closer restaurants, rank: r
  - Captures density of restaurants
  - Acts as a proxy for the amount of competition
- Quality of particular restaurant: q
- Assume utility is linear in these features $V_x = d_x + r_x + q_x$

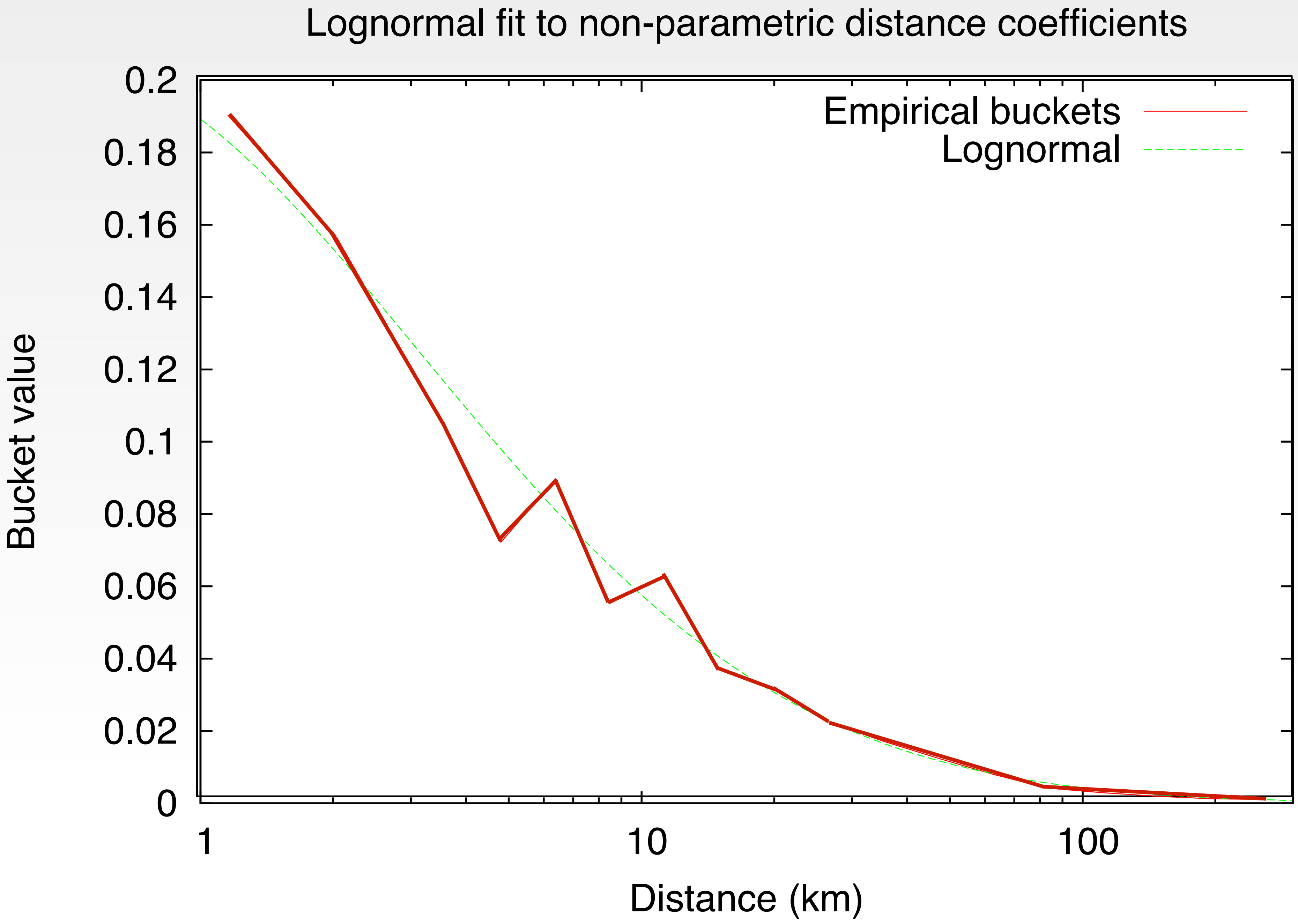
Not Today:

- Personal (user specific) preference
- Time since last visit
- Companions’ desires

Imputed Rank Function



Imputed Distance Function



Results

Predict Likelihood on a held out test set:

Method	**Likelihood**
Uniform choice	1.1
Distance only model	3.9
Rank only model	4.6

Model

Fit both rank and distance functions by log-normals

- Four parameter model: $\mu_{\text{rank}}, \sigma^2_{\text{rank}}, \mu_{\text{distance}}, \sigma^2_{\text{distance}}$

$$s_i = \frac{1}{r_i \sigma_{\text{rank}}} \exp\left(-\frac{(\ln r_i - \mu_{\text{rank}})^2}{2\sigma^2_{\text{rank}}}\right) \cdot \frac{1}{d_i \sigma_{\text{distance}}} \exp\left(-\frac{(\ln d_i - \mu_{\text{distance}})^2}{2\sigma^2_{\text{distance}}}\right)$$

Results

Predict Likelihood on a held out test set:

Method	**Likelihood**
Uniform choice	1.1
Distance only model	3.9
Rank only model	4.6
Lognormal coefficient fit (4 parameters)	5.1

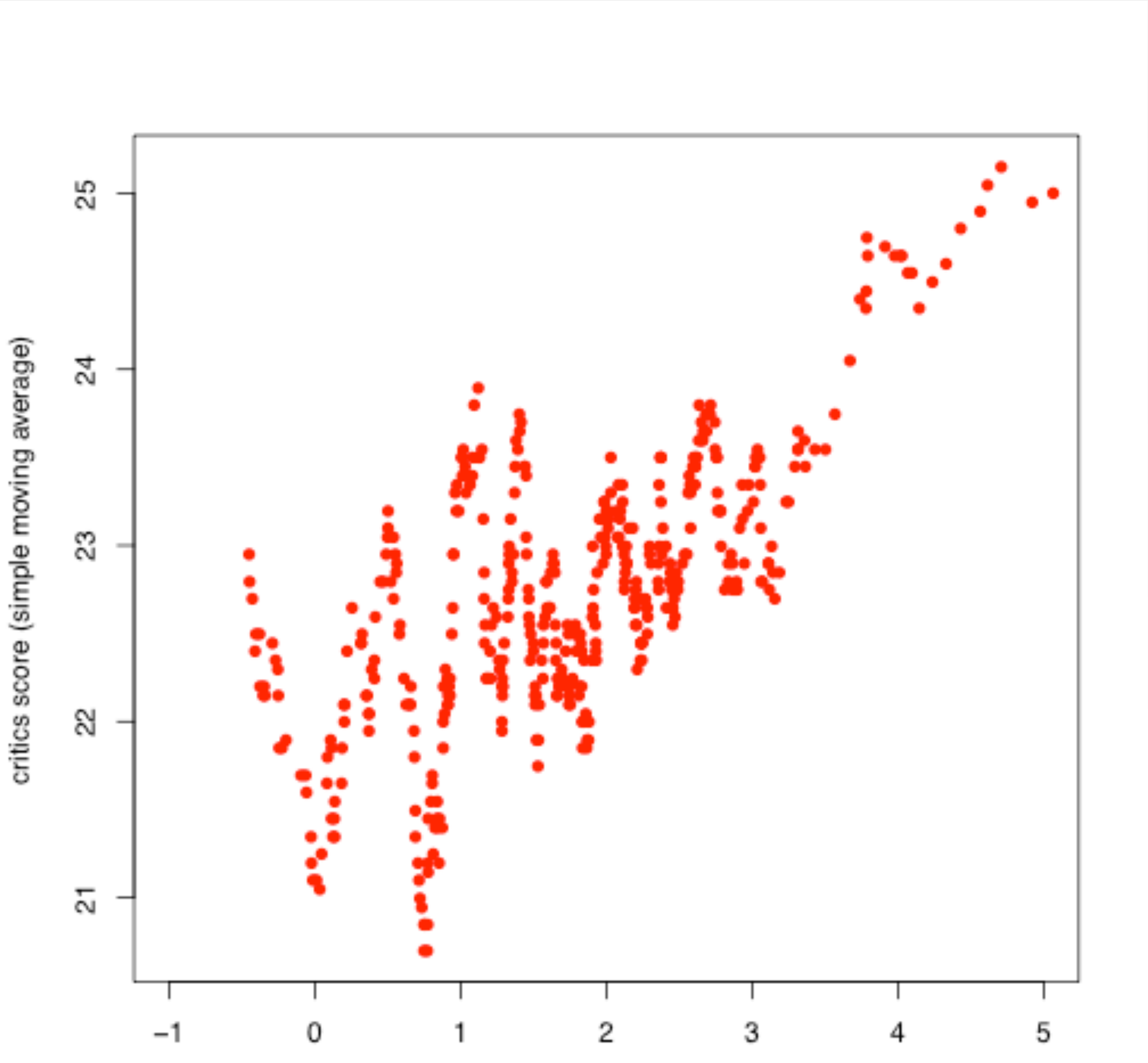
Results

Predict Likelihood on a held out test set:

Method	Likelihood
Uniform choice	1.1
Distance only model	3.9
Rank only model	4.6
Lognormal coefficient fit (4 parameters)	5.1
Non-parametric factored model	5.3

Quality Factor

- – Quality is restaurant specific, makes the model much richer
- – Learn it as the residual on ranks, distances
- – Evaluation: correlation with critics’ scores



Geographic Choice: what have we seen?

Multinomial Logistic Regression with buckets is a powerful technique to assess influence of features based on intensity

Captured interactions may give significantly different influence weights than feature correlations

Given the output of such models, it is possible to observe deeper structure

From this structure, we may find models that are far more parsimonious (why lognormal?)

These new models are much easier to fit when data is sparse

Application 2: Sequential Choice

Repeat consumption

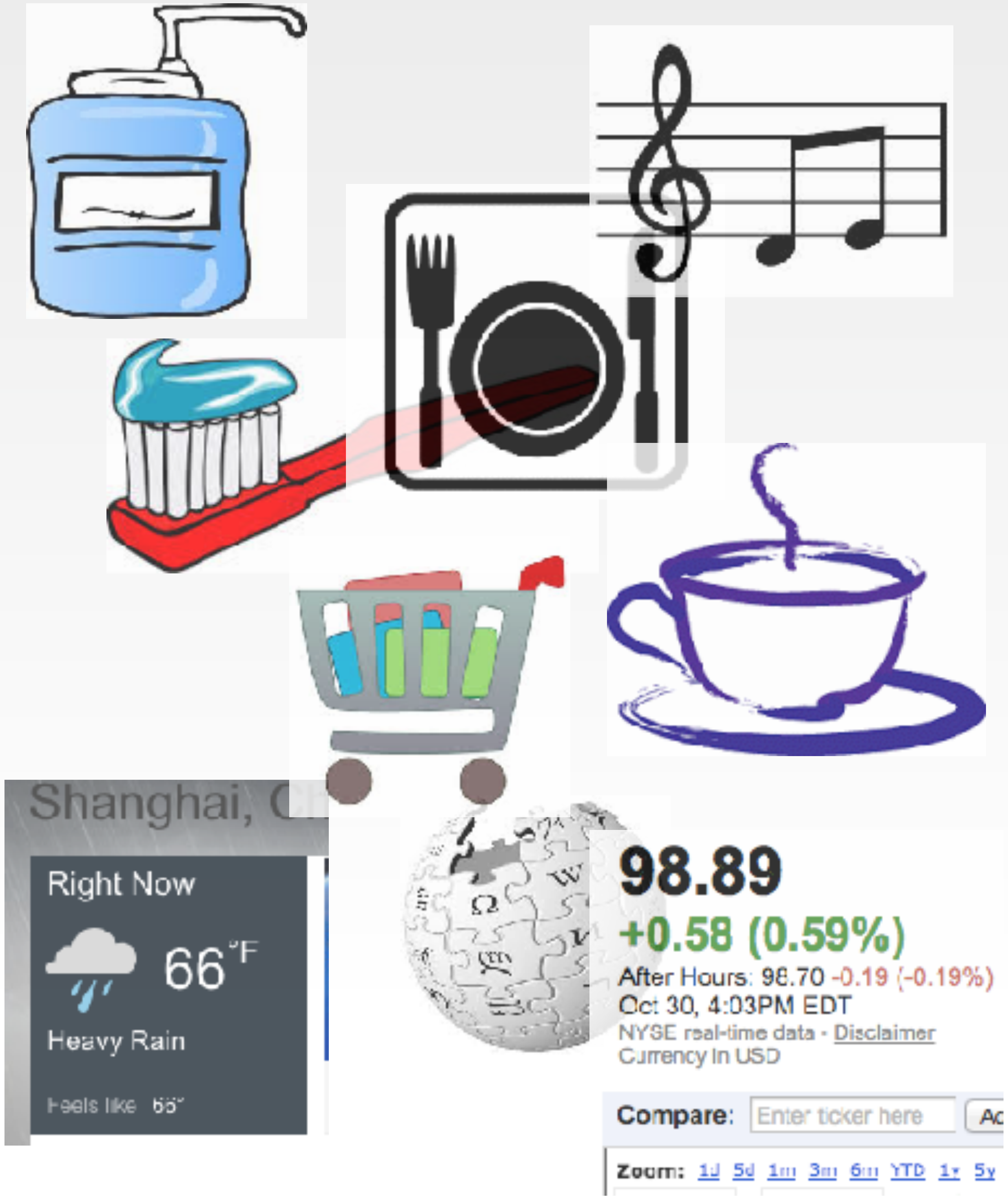
Most of the items we consume are not for the first time

Sometimes go for reliability

Sometimes go for novelty

- Boredom
- New options

We focus on the repeated consumption, not the novel choice.


Shanghai, Ch
Right Now
66°F
Heavy Rain
Feels like 65°
98.89
+0.58 (0.59%)
After Hours: 98.70 -0.19 (-0.19%)
Oct 30, 4:03PM EDT
NYSE real-time data - Disclaimer
Currency in USD
Compare: Enter ticker here Ad
Zoom: 1d 5d 1m 3m 6m YTD 1y 5y

Repeat consumer choice

Marketing studies

Consumer behavior

Music listening experiment [Kahnx et al 97]

- Melioration/overconsumption: listen to favorite on each trial
- Maximization: preserve the high level of enjoyment

Possible explanations

- Difficulties in prediction of taste
- Users try to create the best memory (five flavors vs one flavor LifeSavers)
- Zen principles (pain vs pleasure)

Re-searching

Repeat queries in search logs [Teevan et al]

40% of queries are re-finding queries

Navigational queries are more likely to be repeated

- Information re-finding

Repeat behavior leads to easier prediction of which results will be clicked

Re-visiting web pages

Web page revisitation using browser logs [Adar et al]

50-80% of the web pages are revisited

Revisitation reasons

- Bookmarks/use as hub
- Track content change
- Backbutton

Types of revisitation

- Fast: shopping pages, references, traffic
- Medium: mail, forums, news, ...
- Slow: weekend activity, software updates, ...

Domains of reconsumption

Location checkins

- BrightKite
- Google+

Clicks

- Businesses on maps
- Restaurants on maps
- Wikipedia

Media

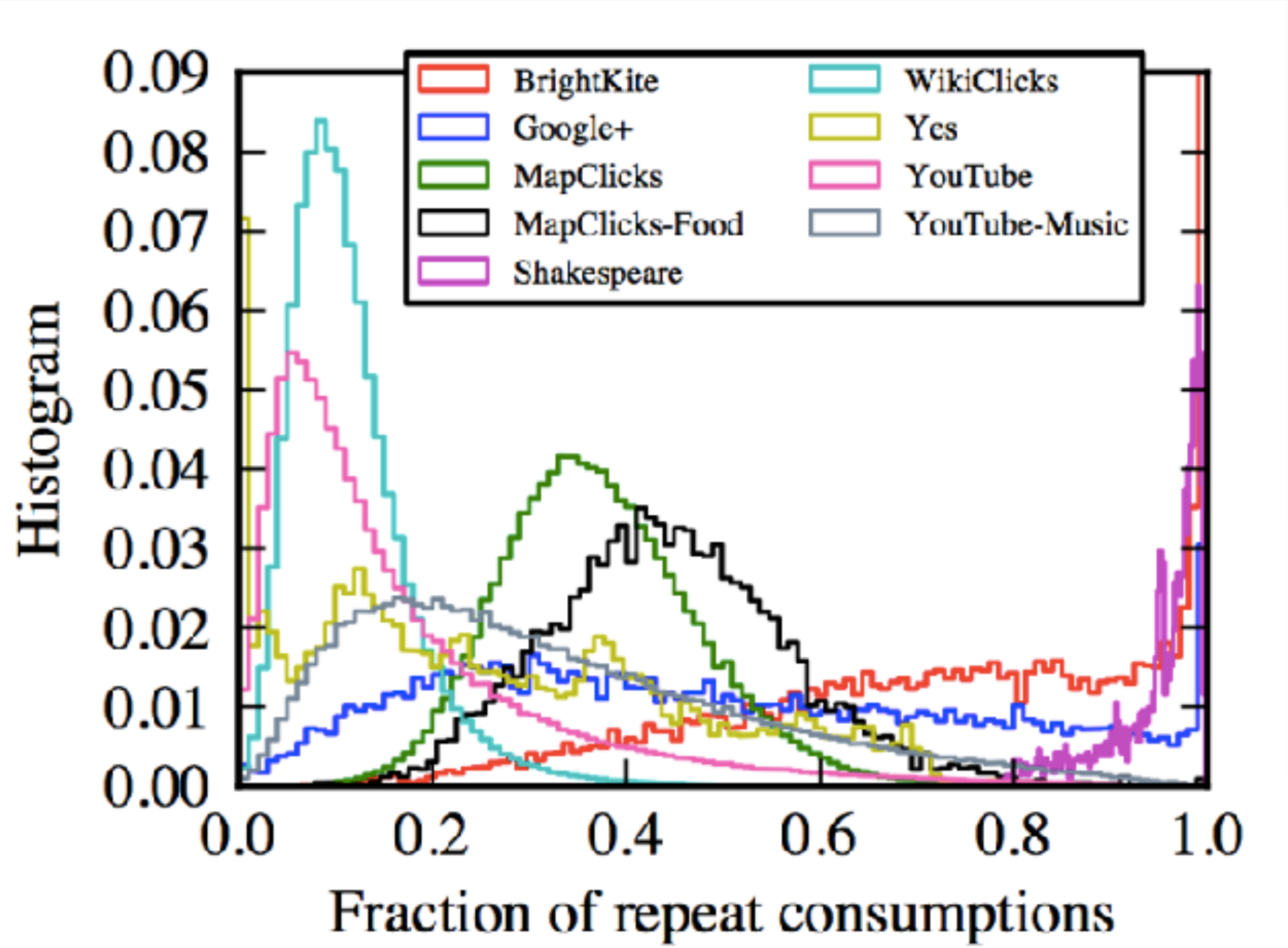
- Youtube
- Music videos
- Playlists from a radio station

Shakespeare!

Characterizing Reconsumption

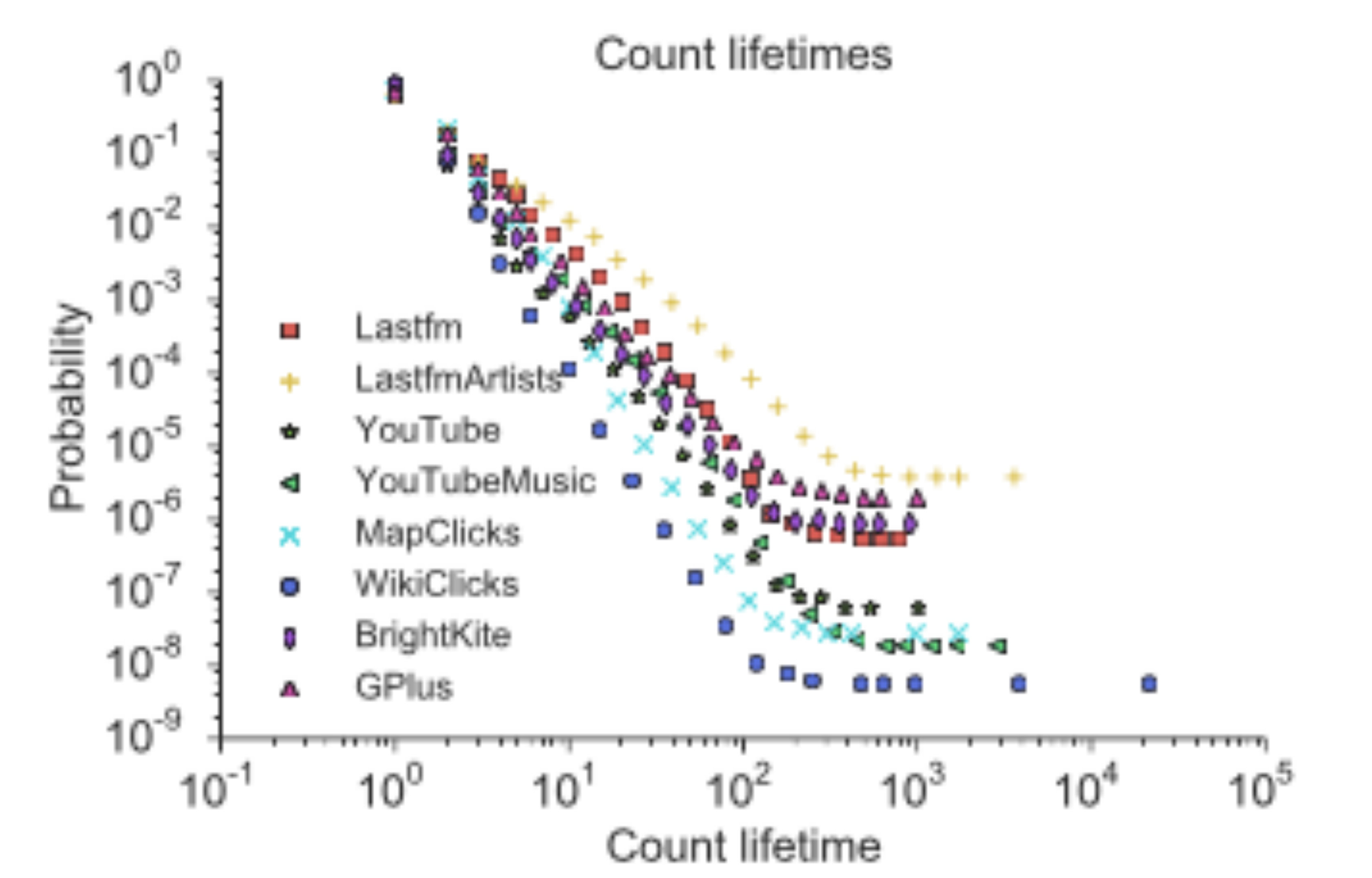
Does it exist?

Distribution of the fraction of repeat consumption



Lifetime distributions

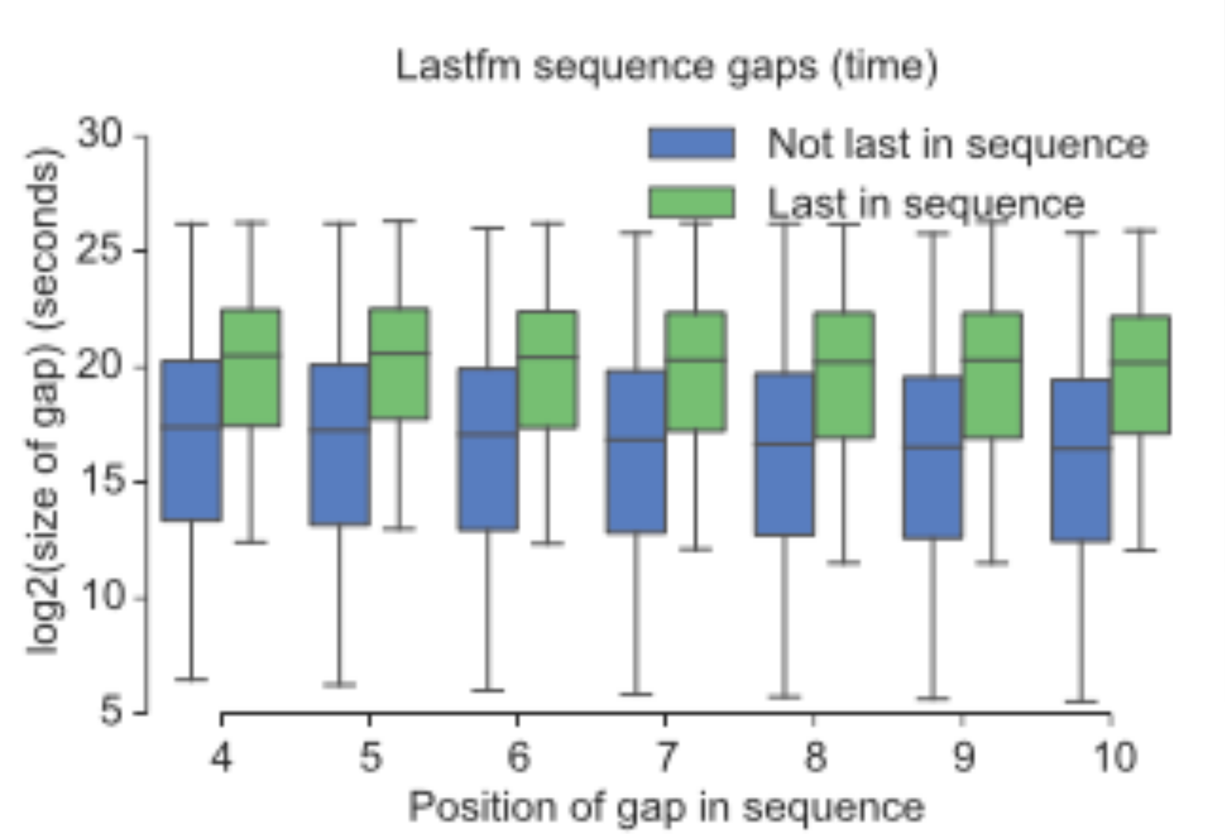
Do items have finite lifetimes?

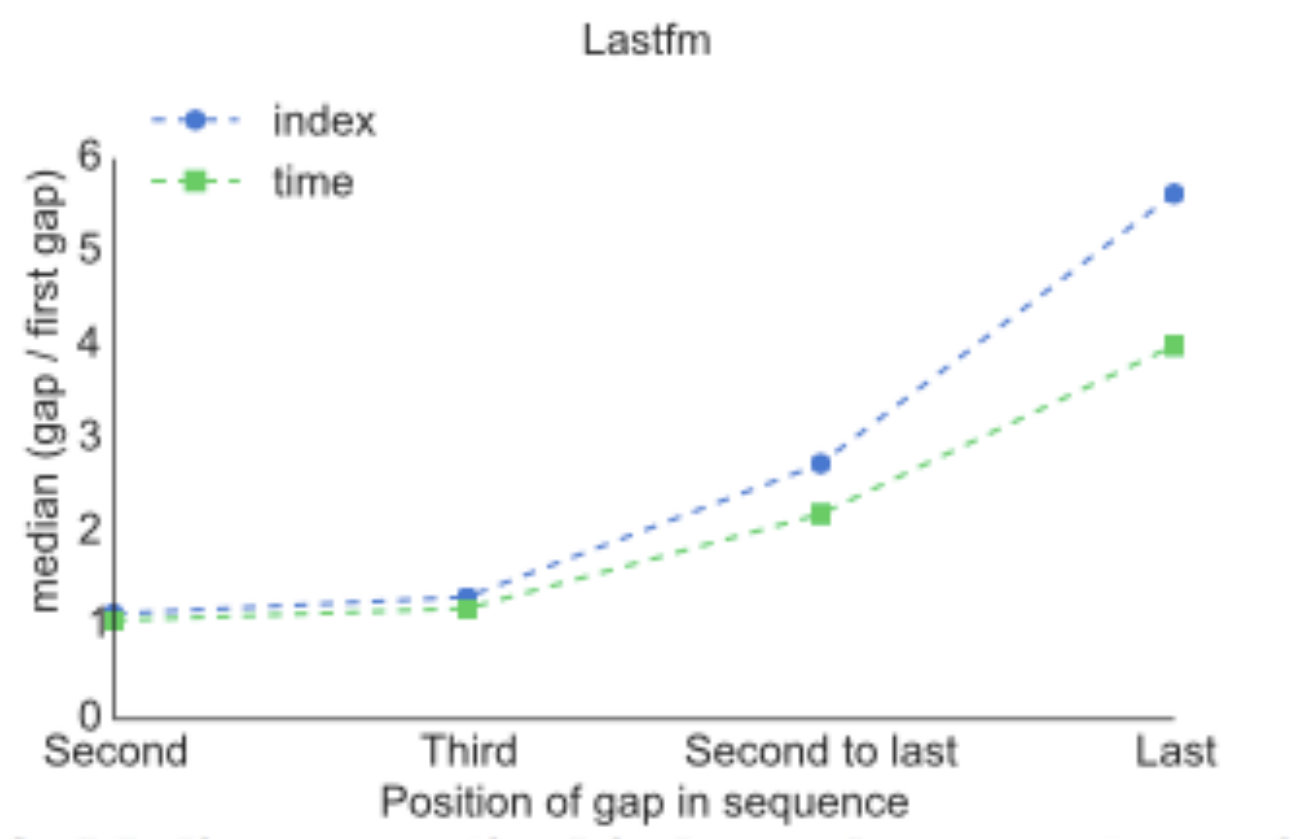


Boredom

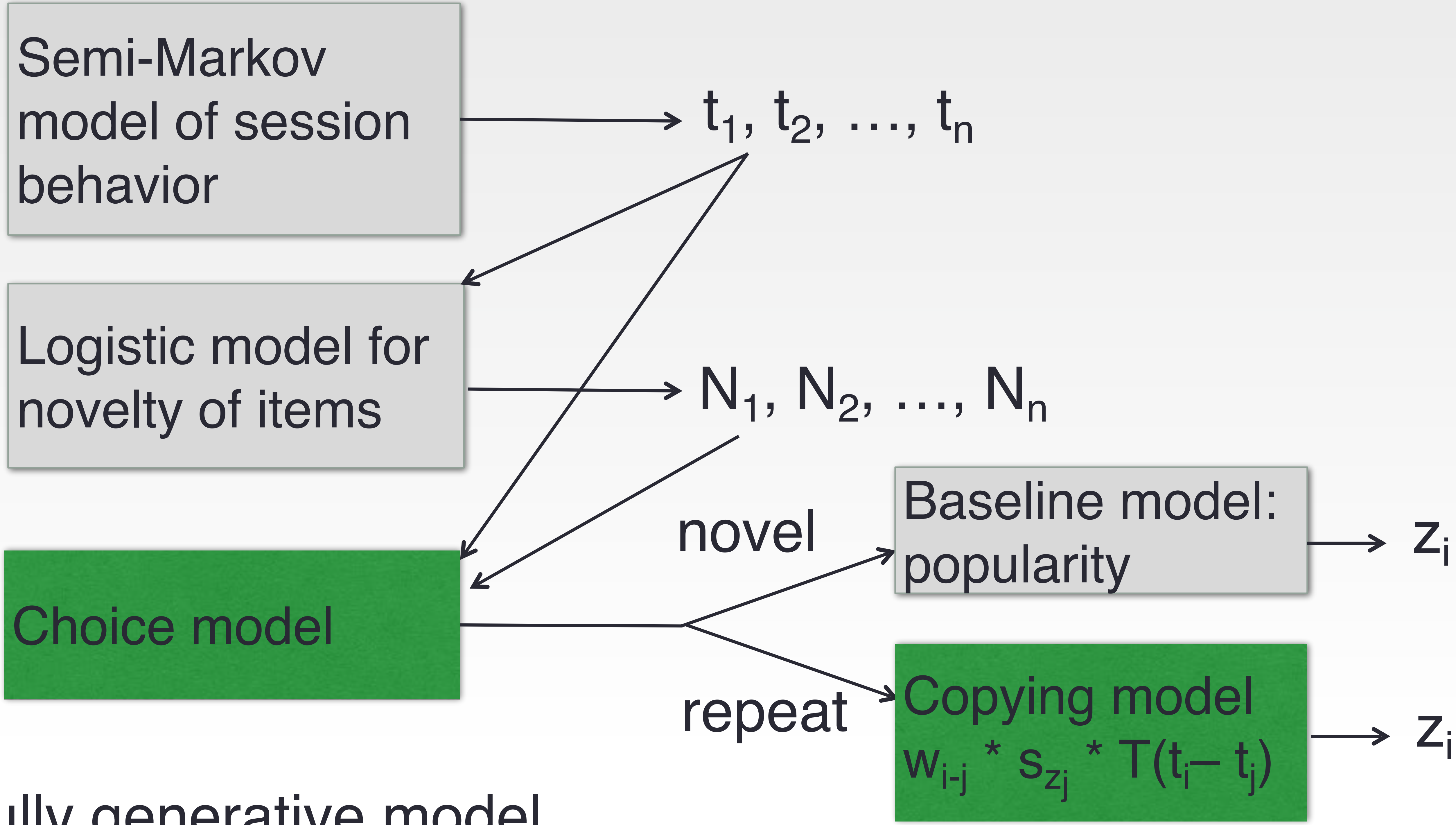
Do users get bored with repeat consumption?

- – Marketers, advertisers care about this
- – Churn/variety-seeking behavior





Summary of model



Semi-Markov model of session behavior

$t_1, t_2, \ldots, t_n$

Logistic model for novelty of items

$N_1, N_2, \ldots, N_n$

Choice model

novel

repeat

Baseline model: popularity

z_i

Copying model $w_{i-j} * s_{z_j} * T(t_i – t_j)$

z_i

Fully generative model.
Also matches macroscopic properties (up next!)

Three key factors

- How popular is the item?
- Time gap since it was last consumed
- How recently was it consumed?

- Can we develop a holistic mathematical framework powerful yet simple enough to explain patterns of reconsumption we observe in real data?

Recency model

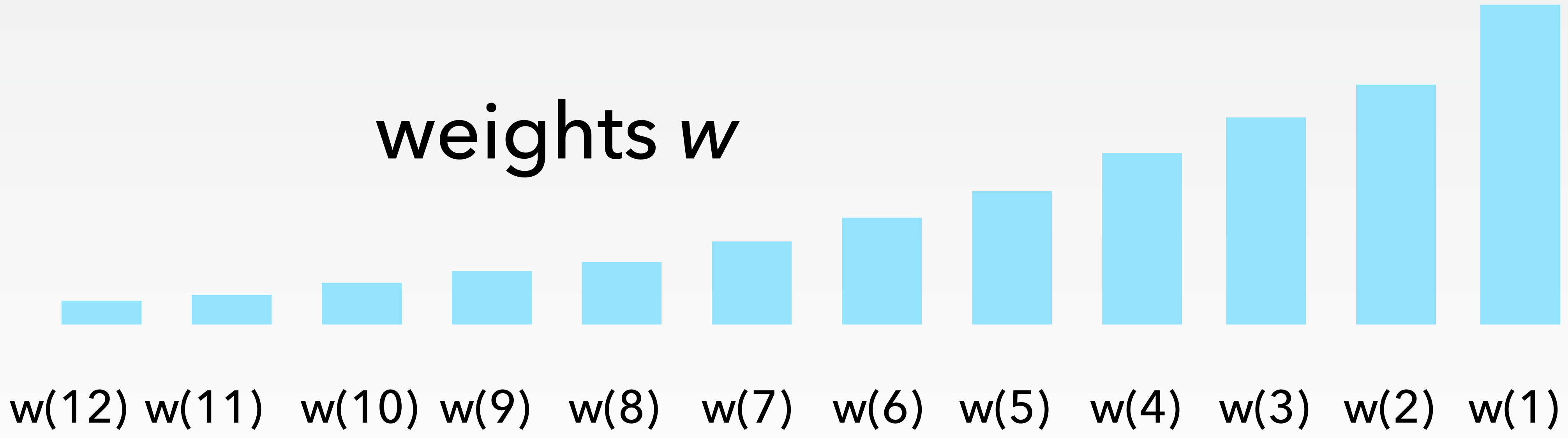
Empirically, recency seems to play a strong role in reconsumption

Technical approach: Combine discrete choice model with “copying model” [Simon, 55] based on recency

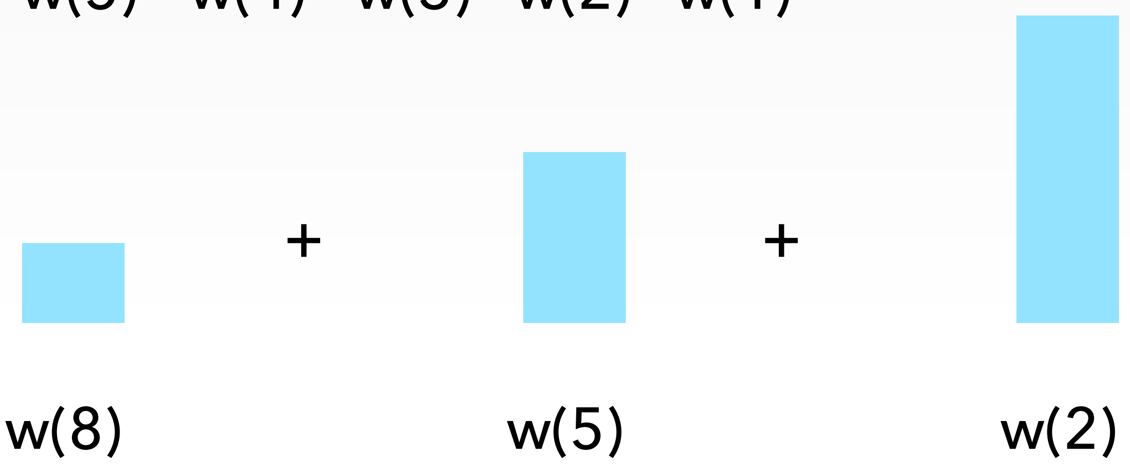
Example

consumption history

a	b	b	c	d	e	b	d	a	c	d	c	?


weights *w*

Pr[d is consumed next] ~



Score-based model

Each item *x* has a score s_x

The score reflects the quality of the item

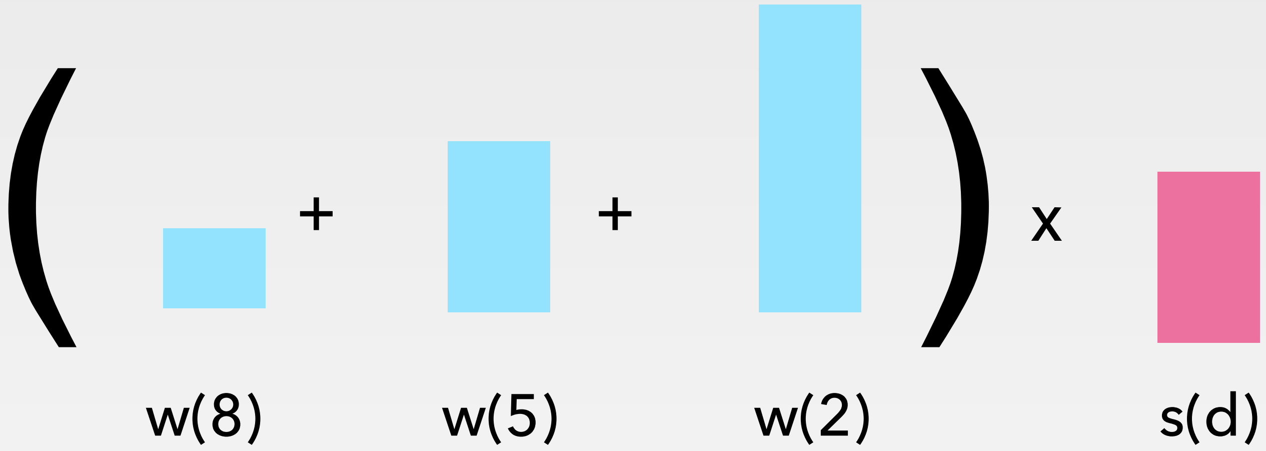
The score dictates the reconsumption pattern

Pick next item *x* using discrete choice, with probability:

$$\Pr[x|X] = \frac{\boldsymbol{s}_x}{\sum_{y \in A} \boldsymbol{s}_y}$$

Combining Recency and Quality

Pr[d consumed next] ~



At position i, pick item *x* with probability:

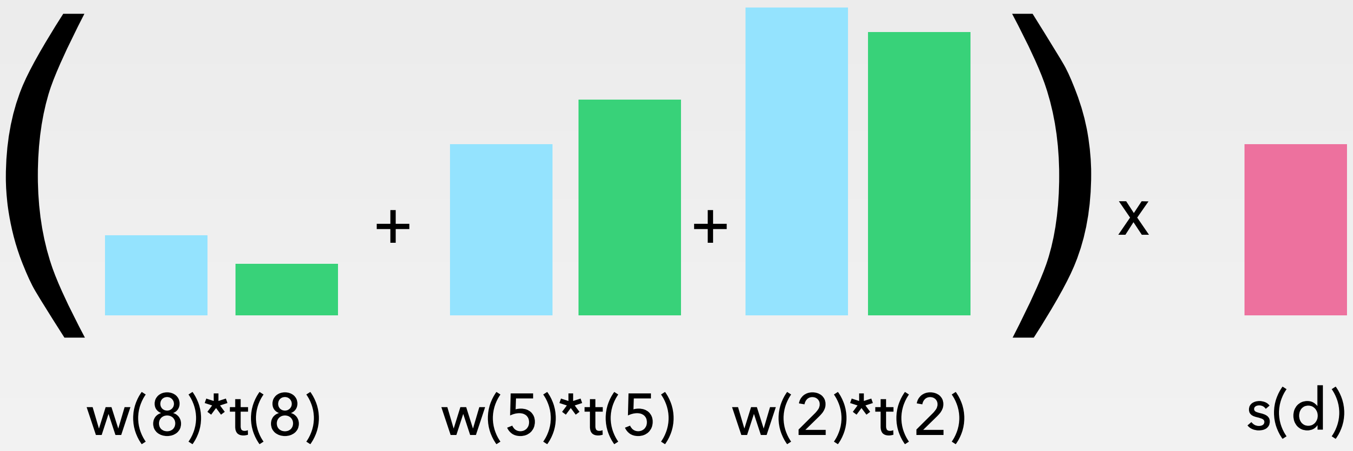
$$\frac{\sum_{j<i} I(x_j = x) w_{i-j} s_{x_j}}{\sum_{j<i} w_{i-j} s_{x_j}}$$

Stochastic gradient ascent

Alternating updates to scores and weights

Combining Recency, Quality, and Time

Pr[d consumed next] ~



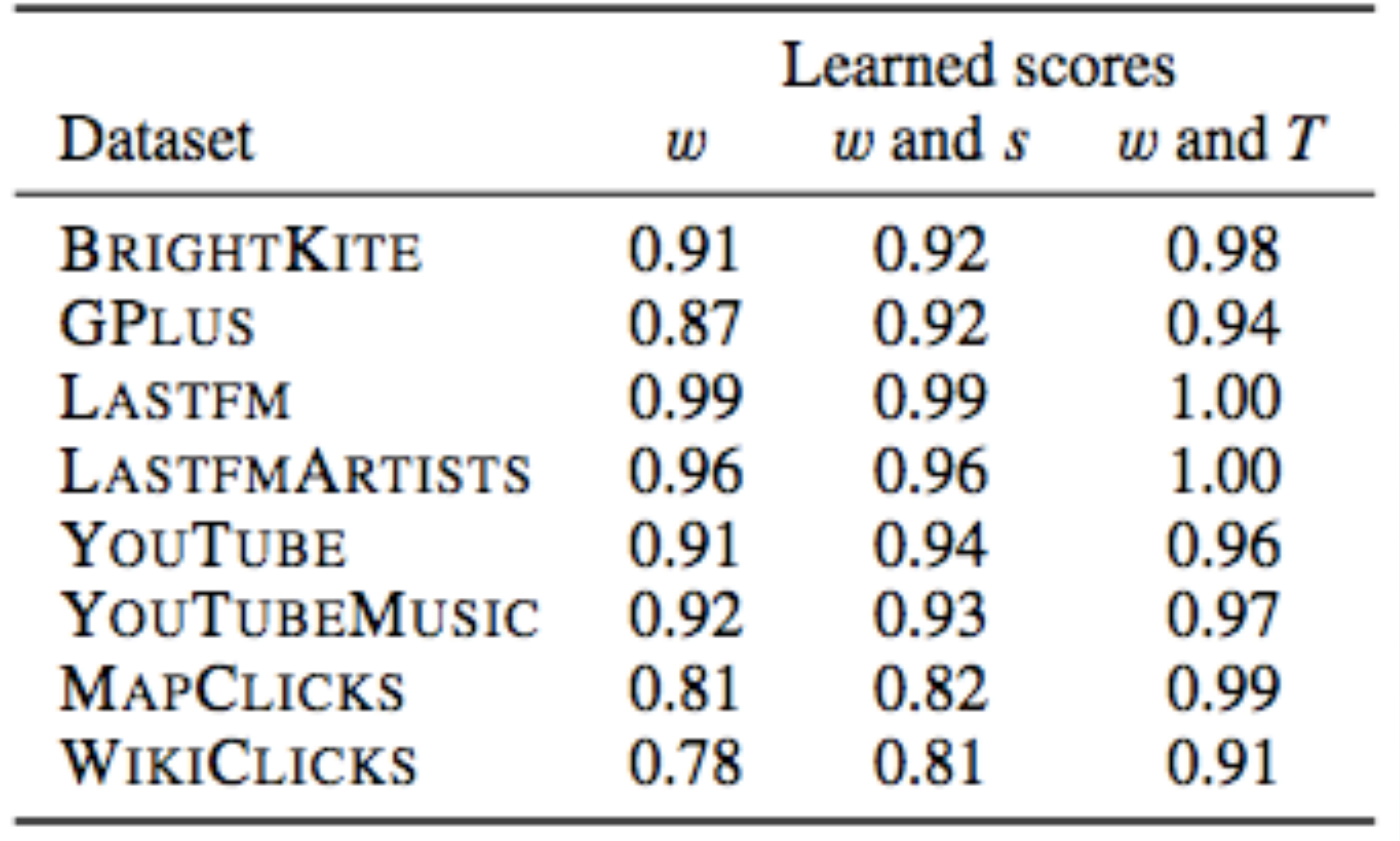
At position i, pick item *x* with probability:

$$\frac{\sum_{j<i} I(x_j = x) w_{i-j} s_{x_j} t_{t_i - t_j}}{\sum_{j<i} w_{i-j} s_{x_j} t_{t_i - t_j}}$$

Stochastic gradient ascent

Alternating updates to scores and weights

Model Quality



	Learned scores		
Dataset	w	w and s	w and T
BrightKite	0.91	0.92	0.98
GPlus	0.87	0.92	0.94
LastFM	0.99	0.99	1.00
LastFMArtists	0.96	0.96	1.00
YouTube	0.91	0.94	0.96
YouTubeMusic	0.92	0.93	0.97
MapClicks	0.81	0.82	0.99
WikiClicks	0.78	0.81	0.91

- Score-only and Popularity-only not competitive
- Recency is most important feature
- Time is more important than item quality
- All model components bring some gain

Item Lifetimes Theoretical Analysis

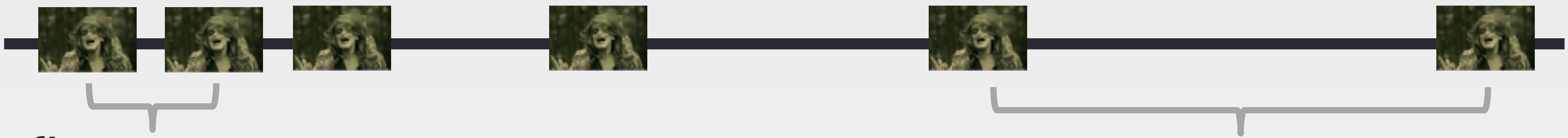
For simple “copying” model with recency only, we can analyze conditions in which an item lives forever:

Theorem:

$$\text{Let } \alpha \text{ be probability of novel item}$$

$$\text{If } \sum_{i=1}^{\infty} w_i < 1/\alpha \text{ then } \Pr[\text{lifetime}(x) < \infty] \to 1$$

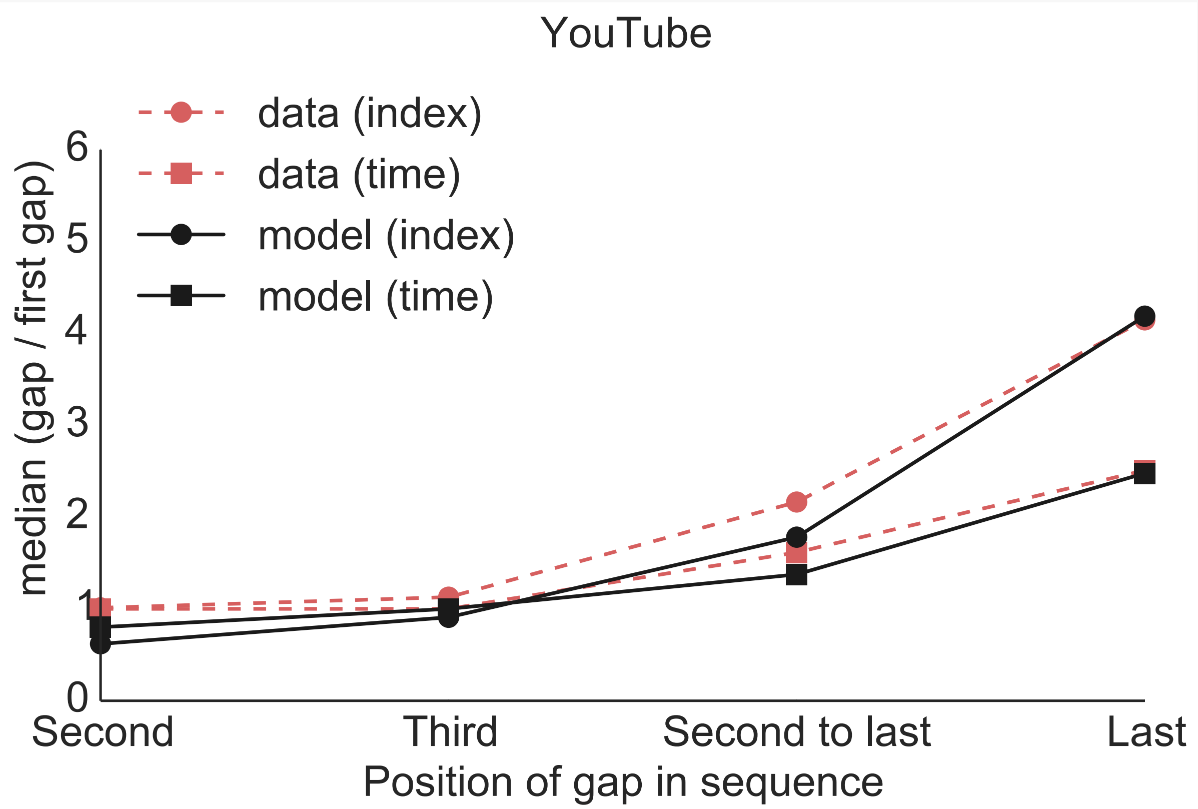
Boredom

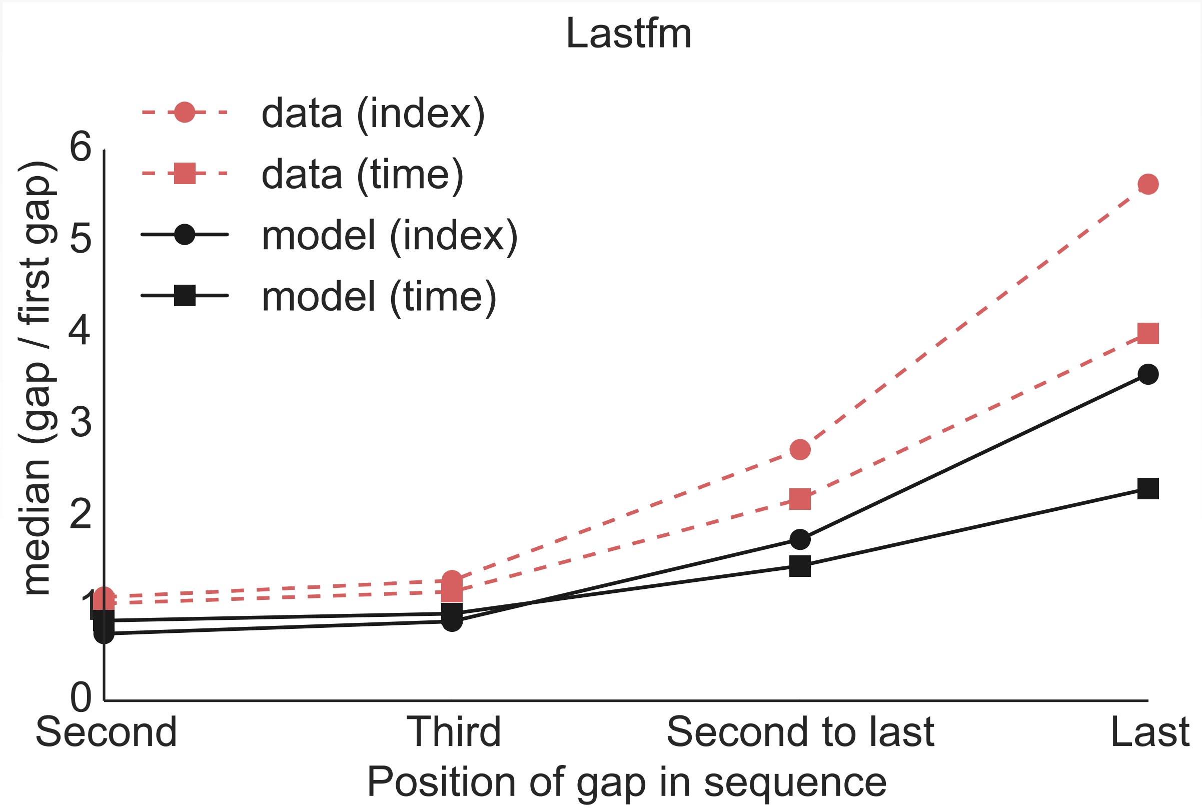


first gap

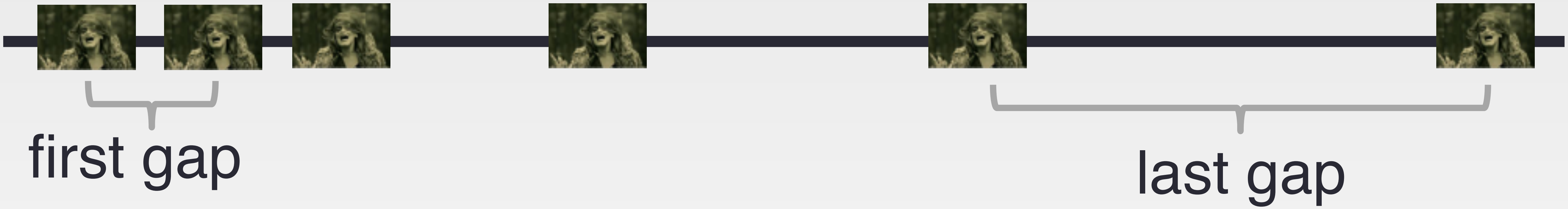
last gap

Before items are abandoned, the gap between consumptions of that item grows in both “index” and “real” time.





Boredom



Consider a simplified choice model with uniform time and item quality scores.

Theorem: Suppose that the weights *w* are monotonically decreasing. Then:

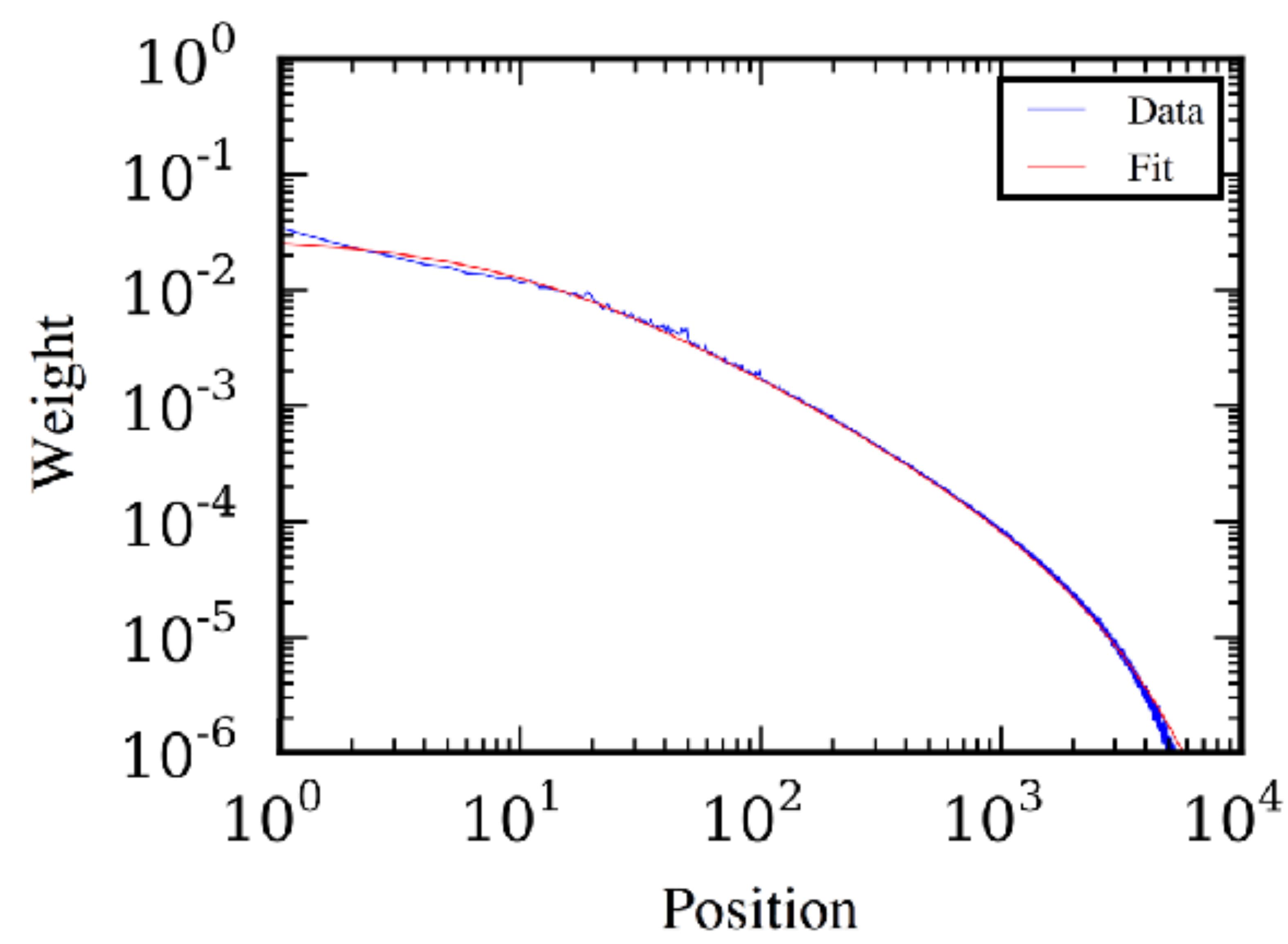
1. $E[j^{th}\text{gap}] < E[(j-1)^{st}\text{gap}]$
2. $E[j^{th}\text{gap}|\text{last occurrence}] > E[j^{th}\text{gap}]$
3. $\forall j > J_0 : E[j^{th}\text{gap}|j^{th}\text{ is last}] > E[j-1^{st}\text{gap}|j^{th}\text{ is last}]$

Parsimonious model

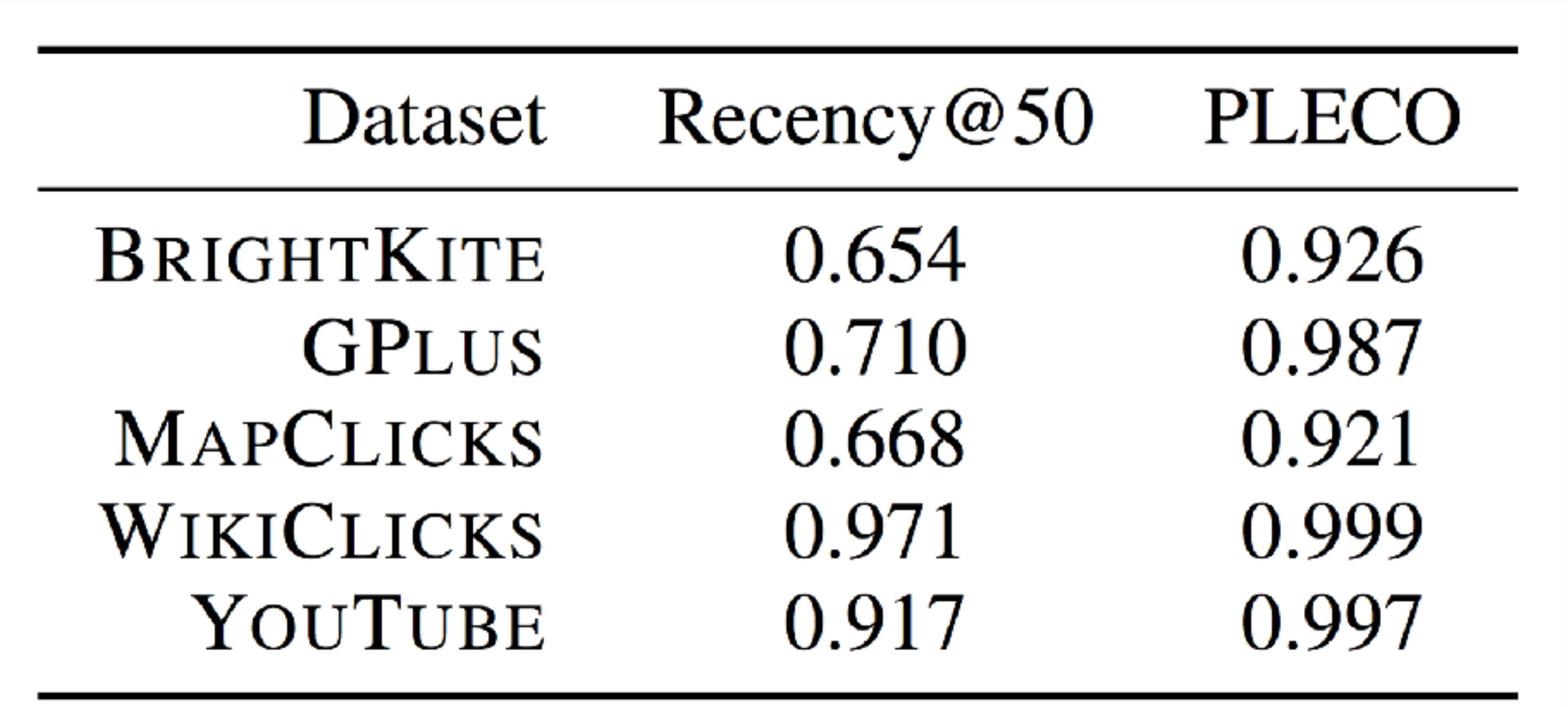
- Recency weights can be compressed

- Good fit: power law with exponential cutoff:

$$\Pr[x] \propto (x+c)^{-a} e^{-bx}$$



Parsimonious model



Dataset	Recency@50	PLECO
BrightKite	0.654	0.926
GPlus	0.710	0.987
MapClicks	0.668	0.921
WikiClicks	0.971	0.999
YouTube	0.917	0.997

Recency model can be expressed using just three parameters!

Conclusions

Conclusions

Discrete choice problems are *everywhere*

The first question is always: does IIA apply — needs empirical validation

If IIA applies, this with some lightweight assumptions implies MNL accurately reflects data

MNL is efficient and extensible, covers a range of choice problems.

It is often possible to sparsify resulting models by analyzed the learned model parameters rather than the marginals

If IIA does not hold, many alternatives exist, but may be difficult to run at internet scale

Thank you!

atomkins@gmail.com