Machine Learning Choices

Johan Ugander
Management Science & Engineering
Stanford University

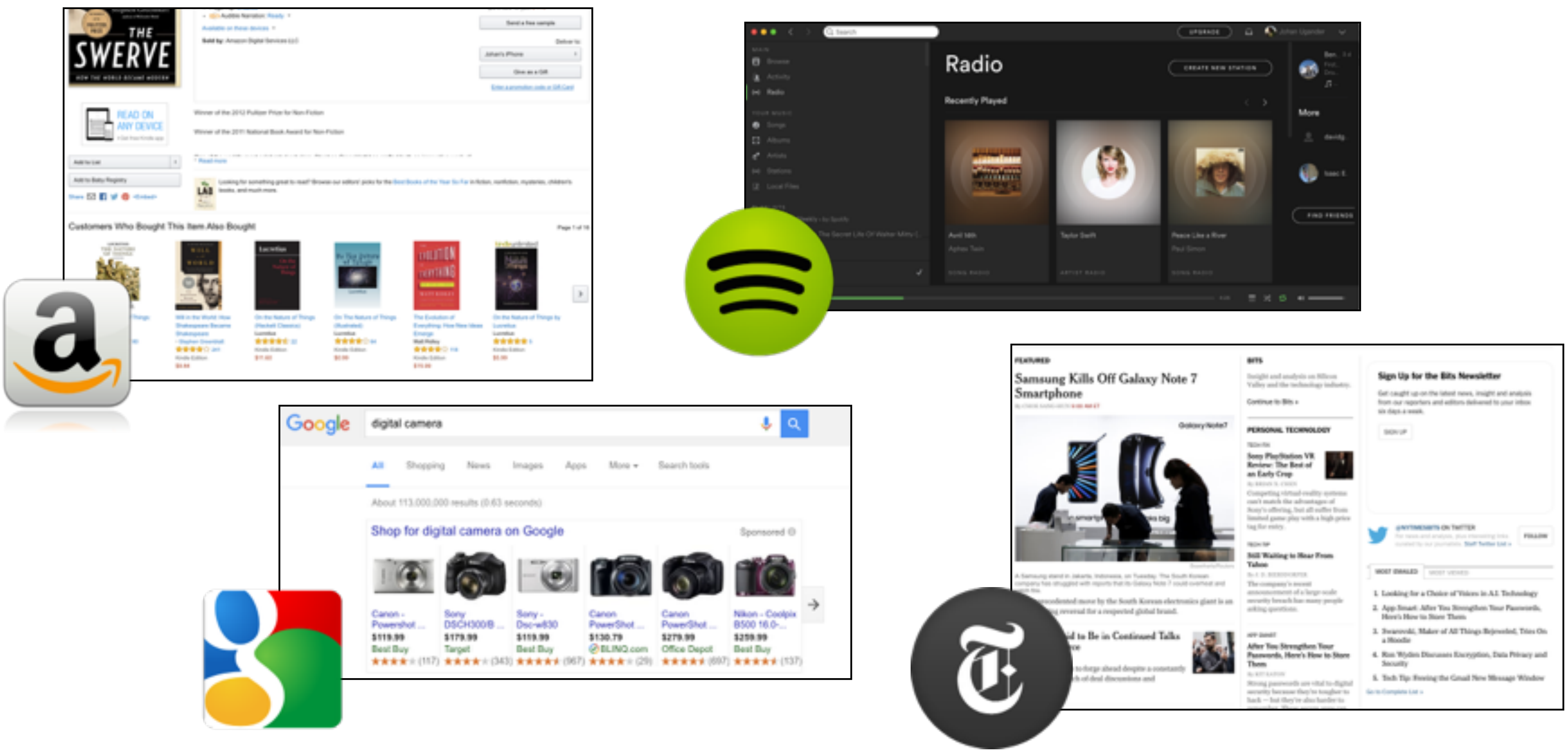
Joint work with:
Stephen Ragain

CODE@MIT
October 15, 2016

Predicting discrete choices

- Classic modeling problem with applications to consumer preferences (Thurstone ’27), commuting (McFadden ’78), and school choice (Kohn-Manski-Mundel ’76)

Predicting digital discrete choices



How well can we learn/predict “choice set effects”?

(a.k.a. “violations of the independence of irrelevant alternatives”)

Learning from comparisons

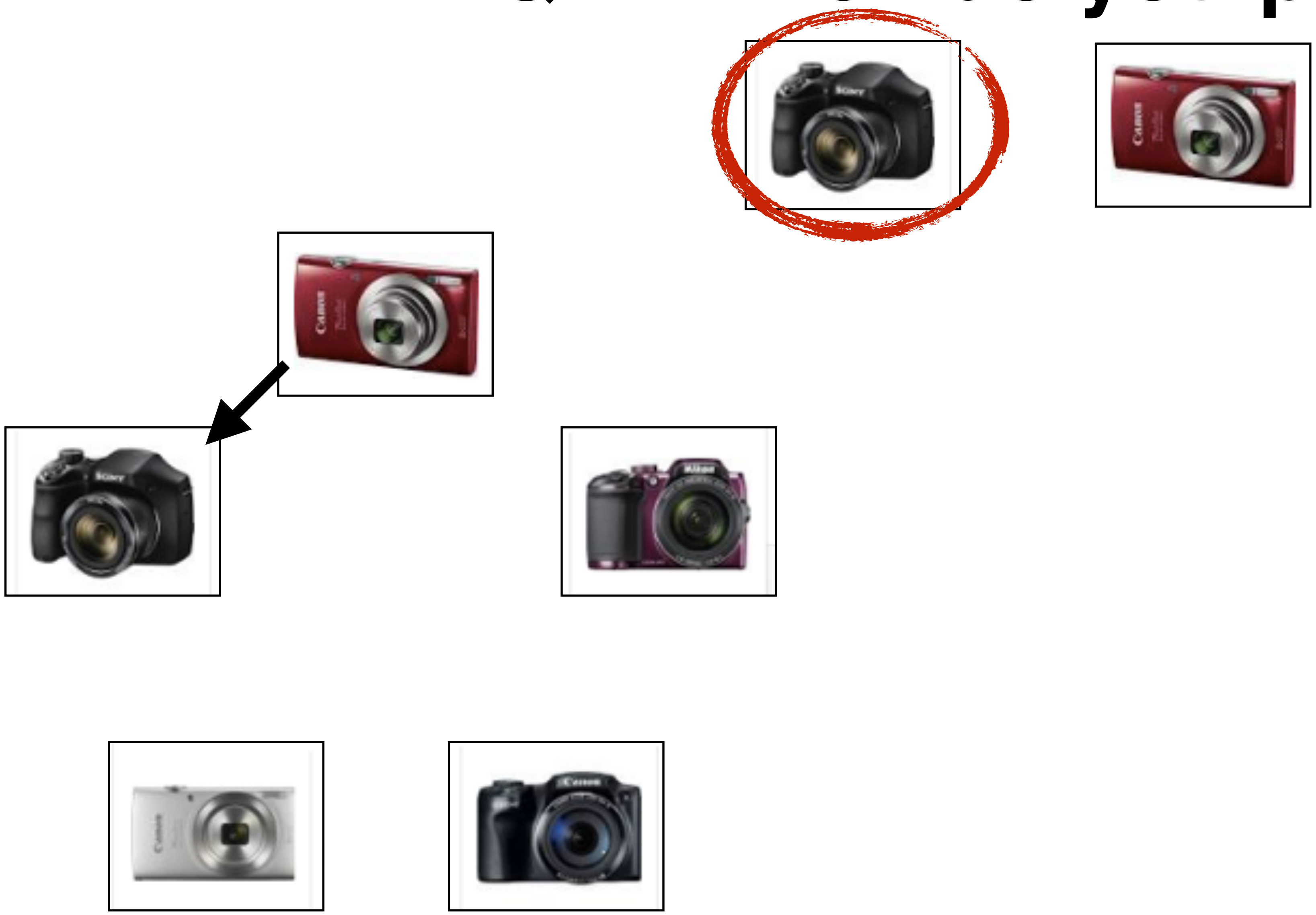
- Comparative Judgement (Thurstone ’27, Bradley-Terry ’57)
  - Learning: “ranking from pairwise comparisons”

Q: “Which do you prefer?”

Learning from comparisons

- Comparative Judgement (Thurstone ’27, Bradley-Terry ’57)
  - Learning: “ranking from pairwise comparisons”

Q: “Which do you prefer?”



Learning from comparisons

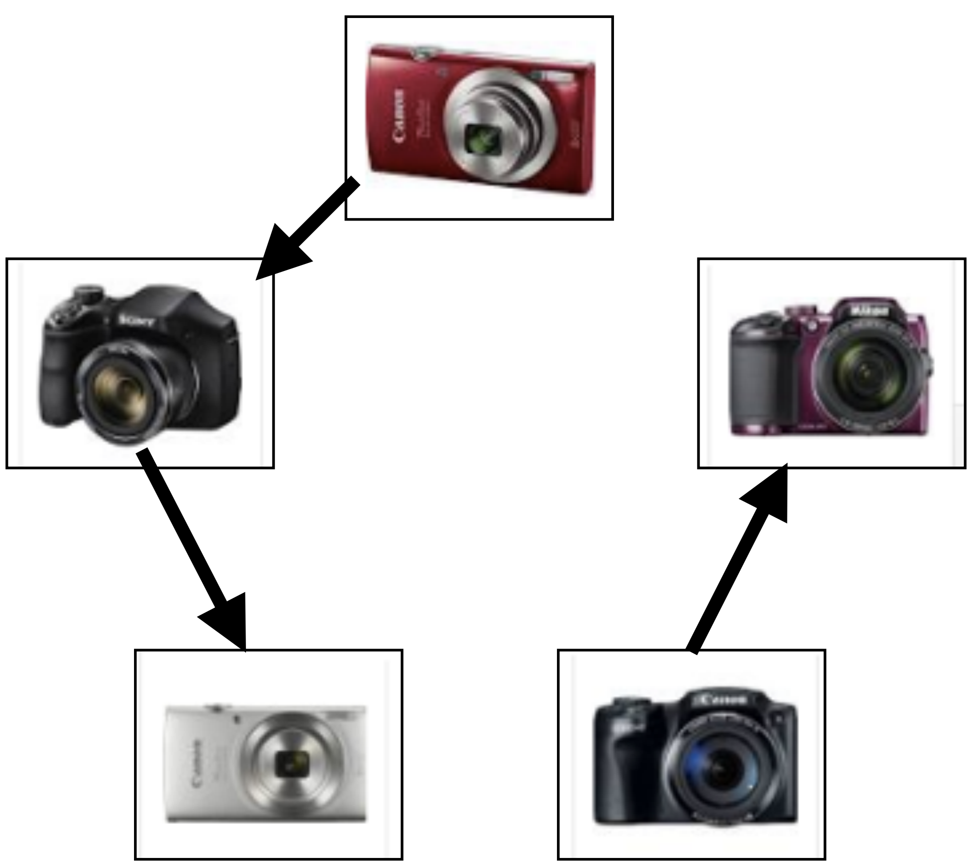
- Comparative Judgement (Thurstone ’27, Bradley-Terry ’57)
  - Learning: “ranking from pairwise comparisons”

Q: “Which do you prefer?”

Learning from comparisons

- Comparative Judgement (Thurstone ’27, Bradley-Terry ’57)
  - Learning: “ranking from pairwise comparisons”

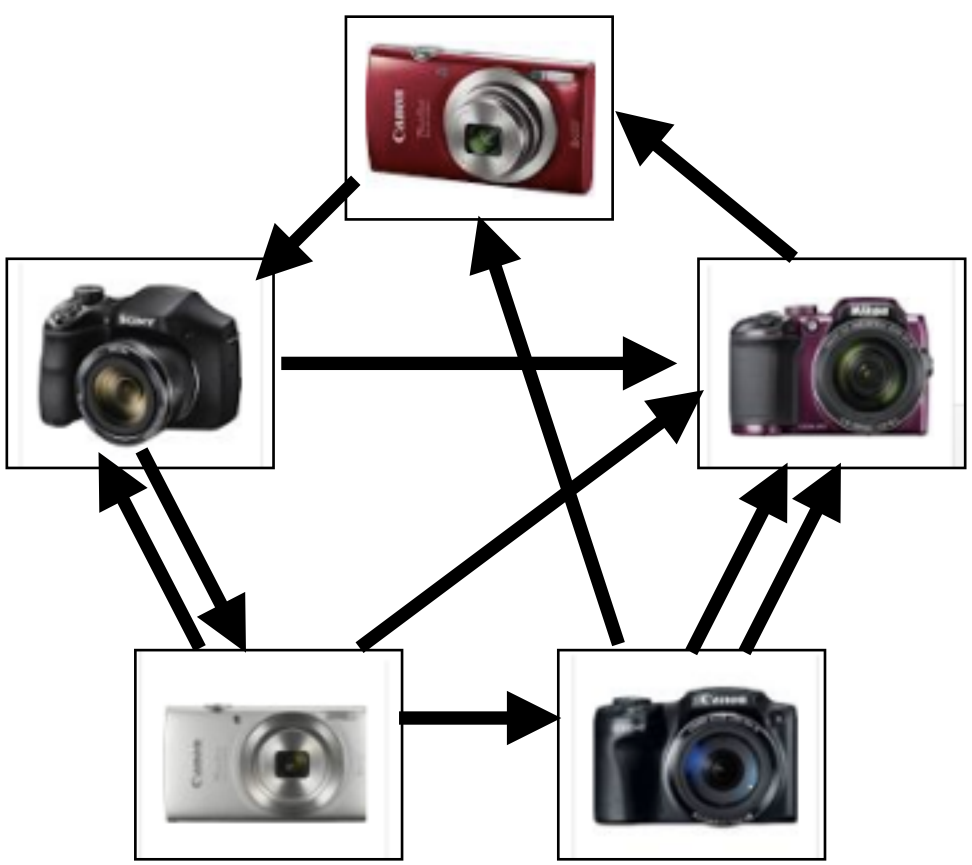
Q: “Which do you prefer?”



Learning from comparisons

- Comparative Judgement (Thurstone ’27, Bradley-Terry ’57)
    - Learning: “ranking from pairwise comparisons”

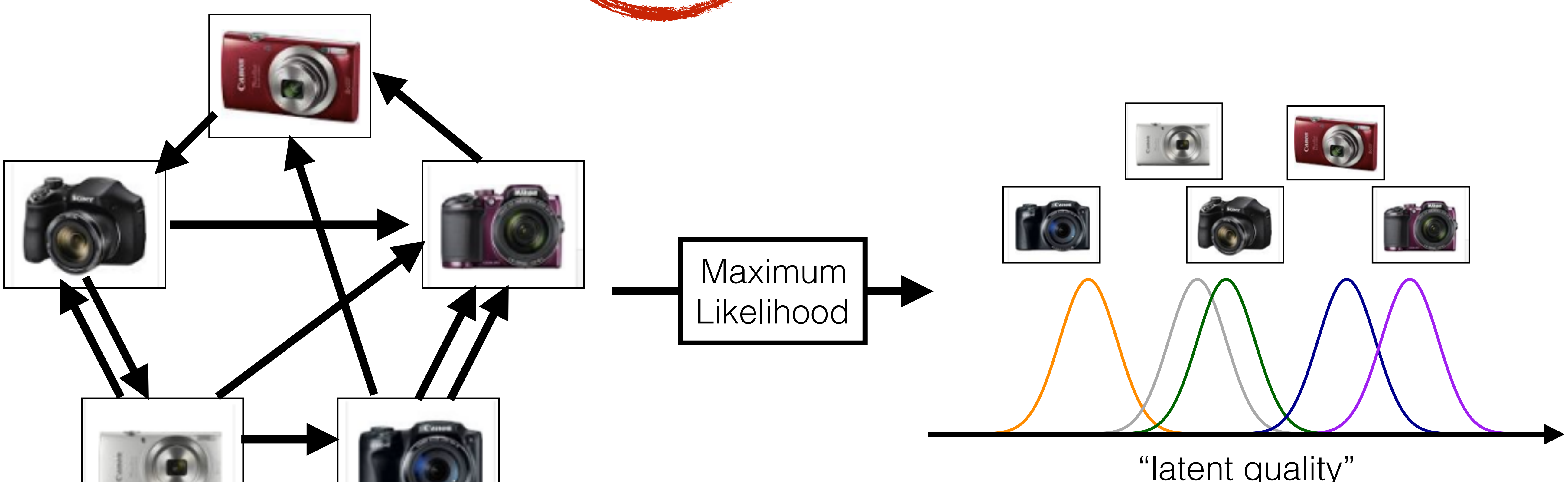
Q: “Which did users click on?”



Learning from comparisons

- Comparative Judgement (Thurstone ’27, Bradley-Terry ’57)
  - Learning: “ranking from pairwise comparisons”

Q: “Which did users click on?”

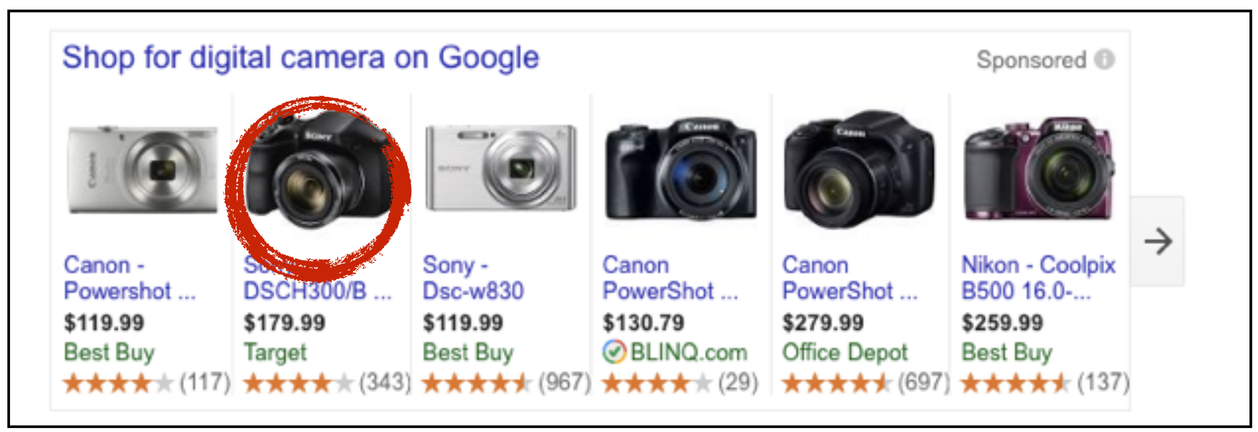


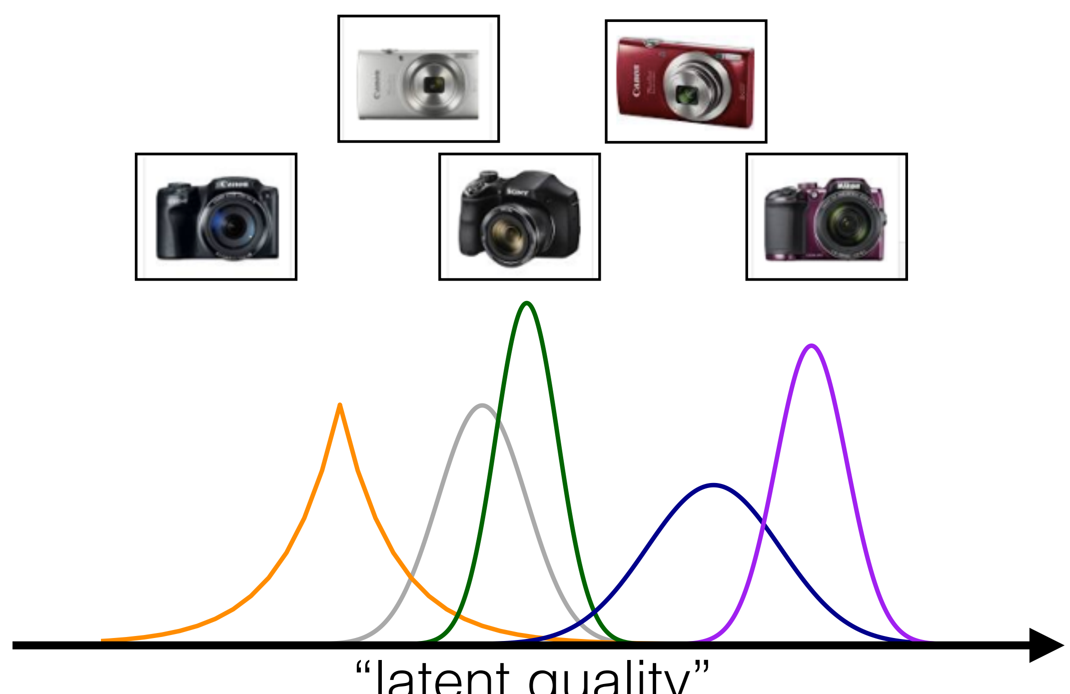
$$\Pr(a \text{ over } b) = \Pr(X_a > X_b)$$

Pairwise to Setwise

- Random Utility Models (Luce '59, McFadden '68, Manski '77)
  - Learning linear models: Multinomial Logistic Regression ("MNL")
  - Regression can also incorporate features of items, users

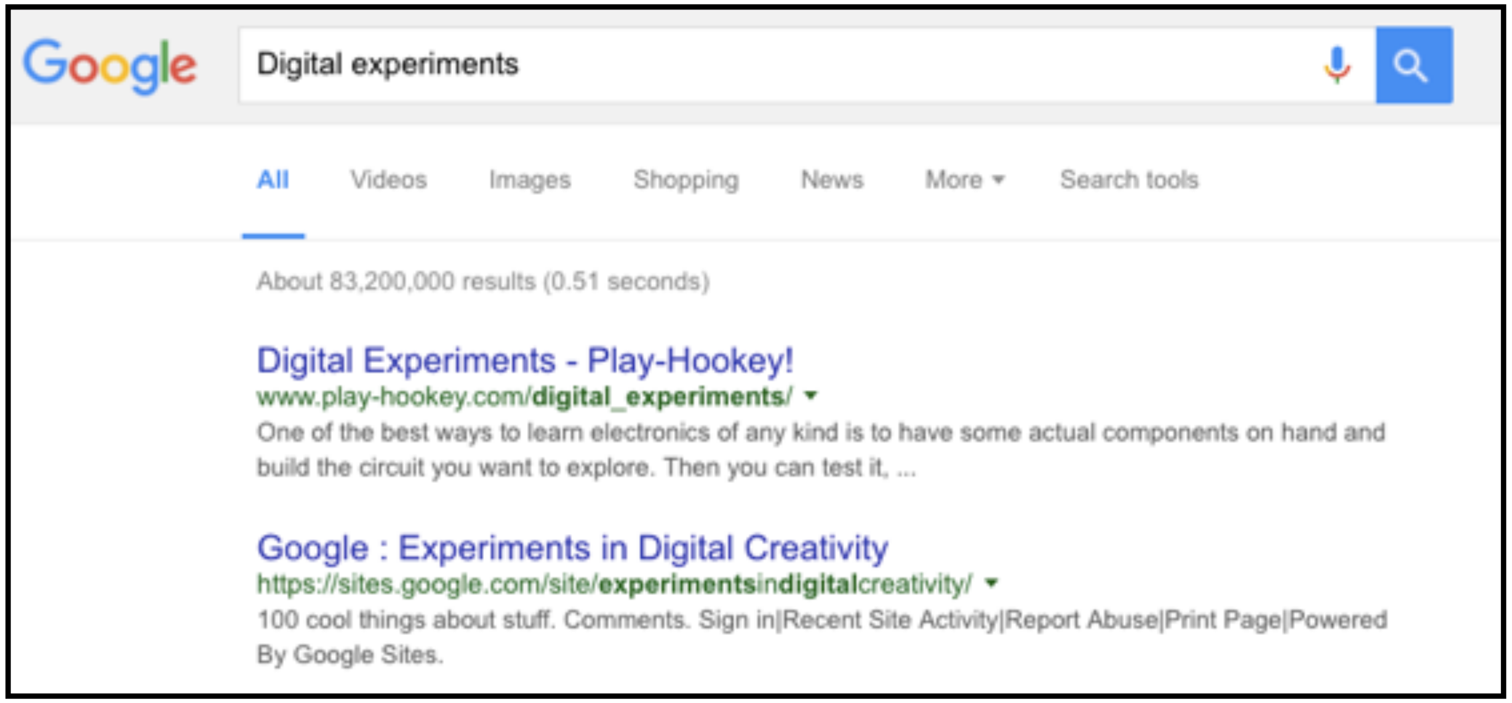
Q: "Which did users click on?"





Learning to Rank

- Explosion of optimization-based approaches to turn click data into optimized rankings (Joachims '02, …)



- Pair-wise, list-wise, point-wise methods.
- Lots of experimentation challenges, e.g. position bias hard to control
- Google since early 2000s: "PageRank is just a feature"

(Bloomberg, 10/2015)

Learning to Choose

- Three assumptions in ranking/RUMs that translate poorly to choices:
  - **Stochastic transitivity**:

$$\left.\begin{aligned}\Pr(a \text{ over } b) > 0.5\\ \Pr(b \text{ over } c) > 0.5\end{aligned}\right\} \Rightarrow \Pr(a \text{ over } c) > 0.5$$

  - **Regularity** between choice sets S, T:

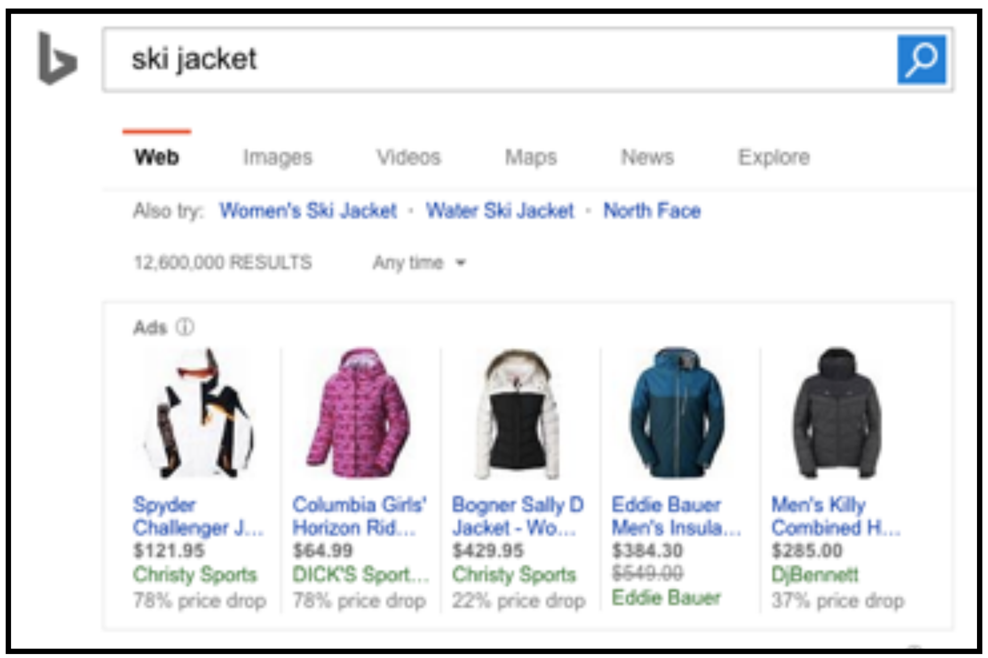
$$S \subseteq T \Rightarrow \Pr(x \text{ from } S) \geq \Pr(x \text{ from } T)$$

  - **Independence of Irrelevant Alternatives** ("IIA"), a.k.a. "choice set effects":

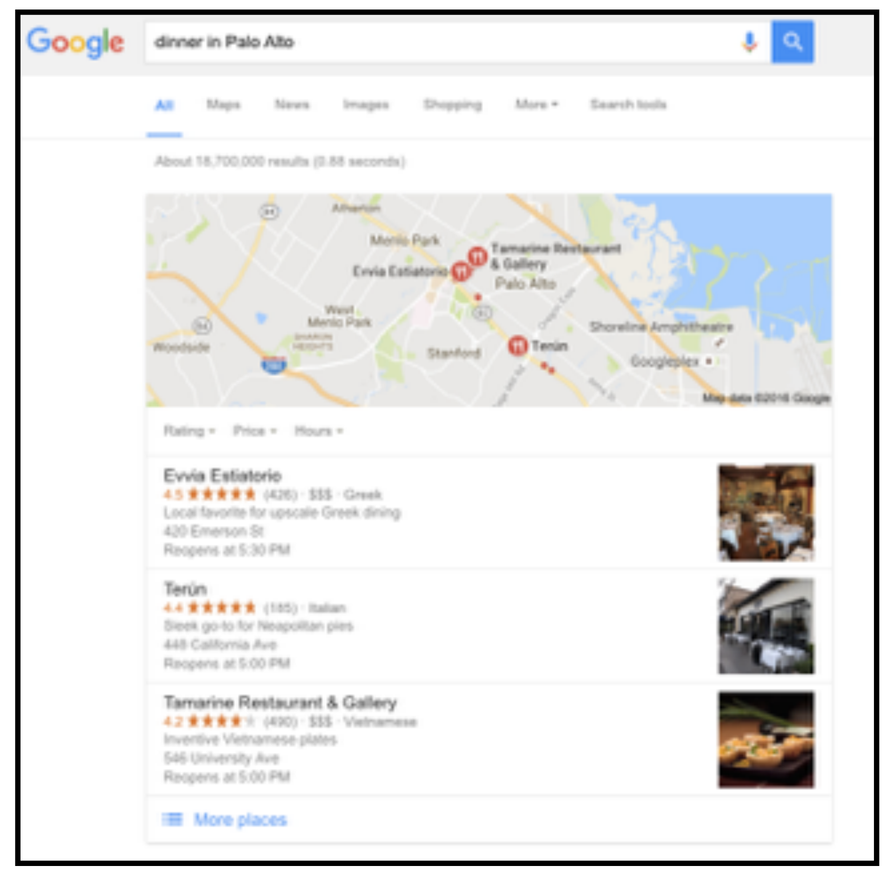
$$\left.\begin{aligned}a, b \in S\\ a, b \in T\end{aligned}\right\} \Rightarrow \frac{\Pr(a \text{ from } S)}{\Pr(b \text{ from } S)} = \frac{\Pr(a \text{ from } T)}{\Pr(b \text{ from } T)}$$

Recent ML work on IIA

- Measurement and models with limited success:
  - Search engine ads (Ieong-Mishra-Sheffet ’12, Yin et al. ‘14)
  - Google web browsing choices (Benson-Kumar-Tomkins ’16)



“ad group quality”

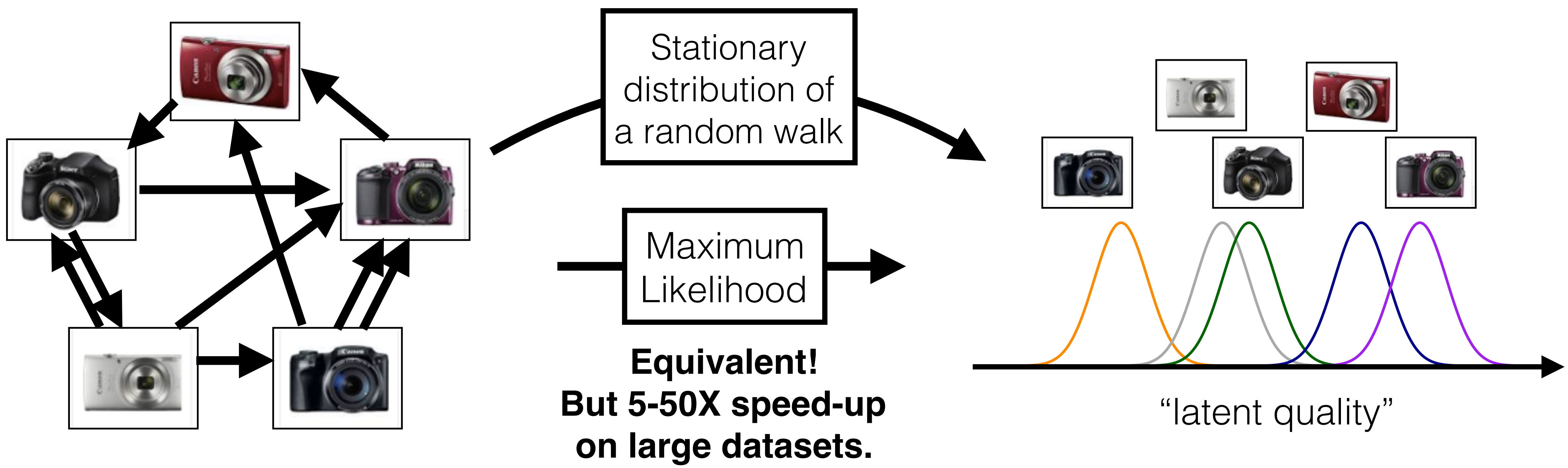


sequential browsing -> choices

- Lots of violations of IIA observed.

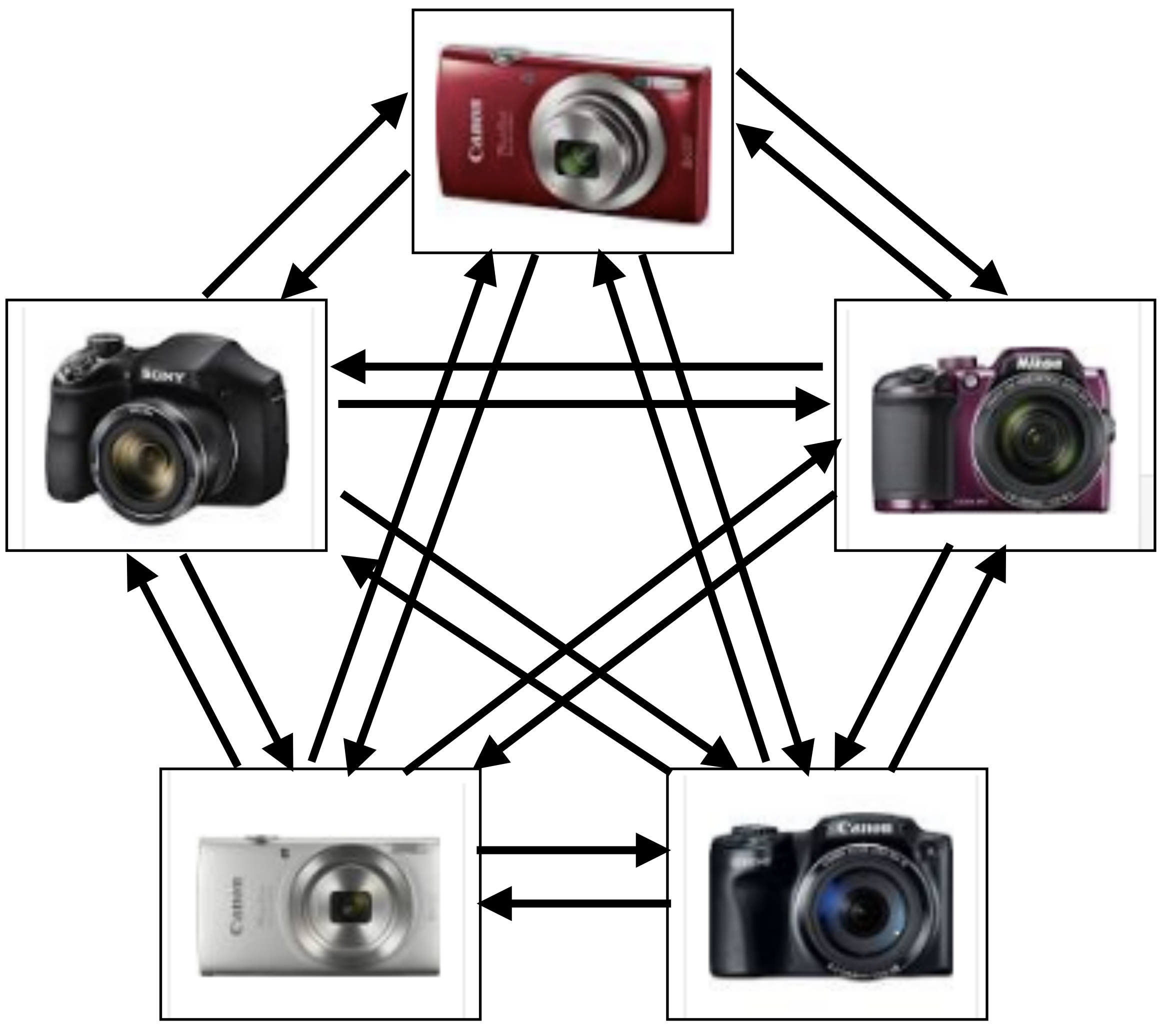
New eyes for an old problem

- Learning BTL/PL/MNL recently connected to Markov chains:
  - “RankCentrality” (Negahban-Oh-Shah ’12)
  - “Luce Spectral Ranking” (Maystre-Grossglauser ’15)



Luce Spectral Ranking

- Maystre & Grossglauser noticed that the stationary conditions of an optimization routine for MNL coincide with the stationary conditions of a particularly parameterized Continuous-Time Markov Chain.



- **If** rates are set to pairwise choice probabilities:

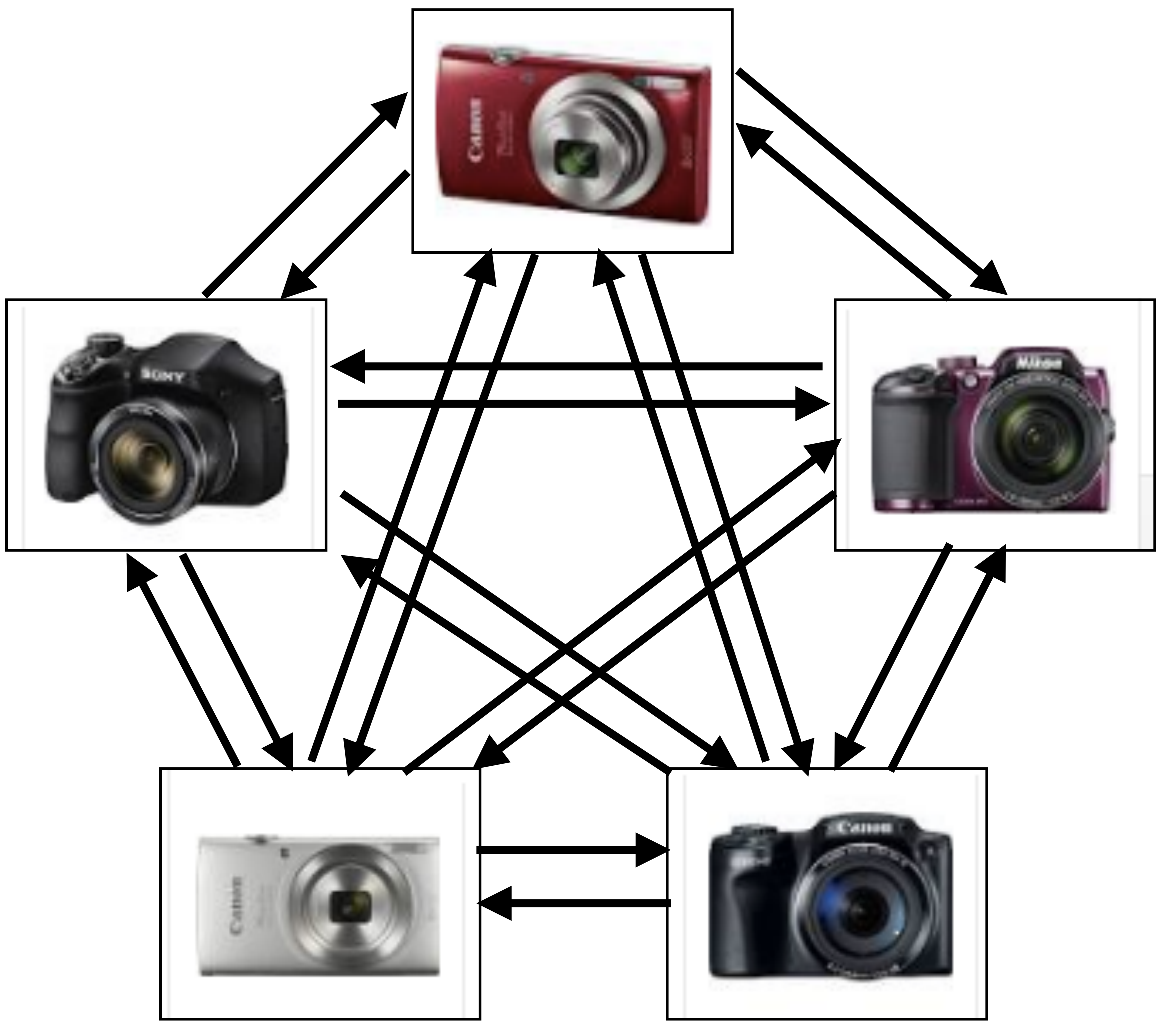
$$q_{ij} = \frac{\gamma_i}{\gamma_i + \gamma_j}, \forall i, j, i \neq j$$

- **Then** normalized “quality” is the stationary distribution:

$$\frac{1}{\sum_{i=1}^{n} \gamma_i} [\gamma_1 \ldots \gamma_n]^T Q = 0$$

Luce Spectral Ranking

- Maystre & Grossglauser noticed that the stationary conditions of an optimization routine for MNL coincide with the stationary conditions of a particularly parameterized Continuous-Time Markov Chain.



- **If** rates are set to pairwise choice probabilities:

$$q_{ij} = \frac{\gamma_i}{\gamma_i + \gamma_j}, \forall i, j, i \neq j$$

- **Then** normalized “quality” is the stationary distribution:

$$\frac{1}{\sum_{i=1}^{n} \gamma_i} [\gamma_1 \dots \gamma_n]^T Q = 0$$

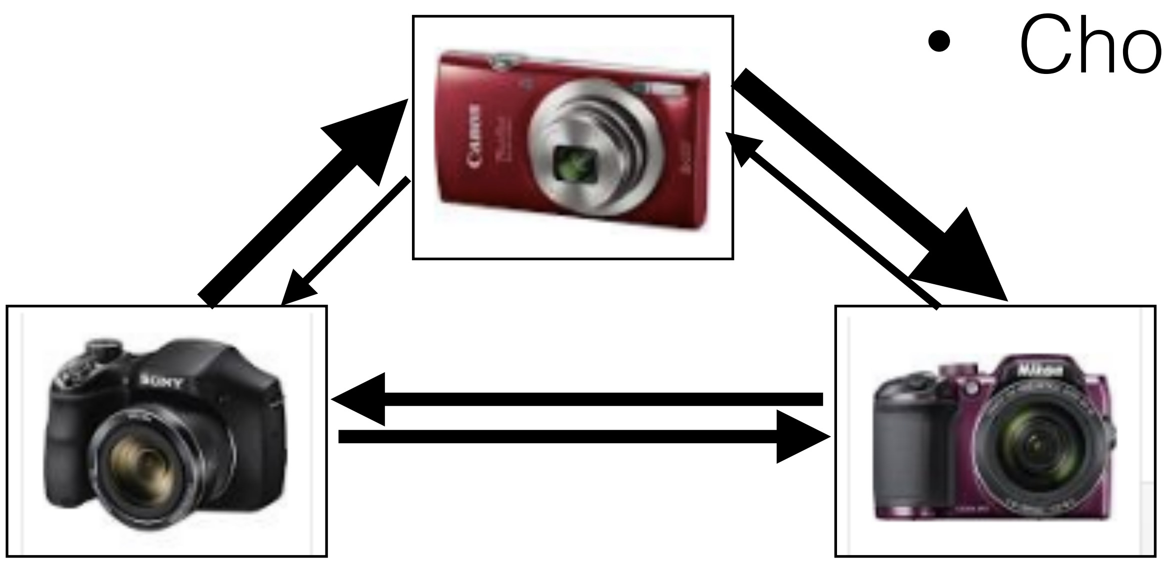
- **Implies a much more general choice model:** why restrict to that parameterization?

Pairwise choice Markov chains

- New model that very naturally models choice set effects
- Model choice probabilities for set S as the stationary distribution of pairwise CTMC on S with rates q_{ij} as parameters.

- Choice from {i,j}:

$$\pi^T \begin{bmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{bmatrix} = 0$$



- Choice from {i,j,k}:

$$\pi^T \begin{bmatrix} -\sum_{i\neq 1} q_{1i} & q_{12} & q_{13} \\ q_{21} & -\sum_{i\neq 2} q_{2i} & q_{23} \\ q_{31} & q_{32} & -\sum_{i\neq 3} q_{3i} \end{bmatrix} = 0$$

- Same parameters interleaved across different set sizes.

Key properties of PCMC model

- **No** assumptions of **transitivity**, **IIA**, or **regularity**
  - No regularity means not even a RUM!
  - Even “Elimination by Aspects” (Tversky ’72) is a RUM.

Key properties of PCMC model

- **No** assumptions of **transitivity**, **IIA**, or **regularity**
  - No regularity means not even a RUM!
  - Even “Elimination by Aspects” (Tversky ’72) is a RUM.
- PCMC **does** satisfy axiom of **uniform expansion** (Yellott ’77)
  - “probabilities are unchanged by making copies of the set”
  - UE in an independent RUM implies Luce’s Axiom (and thus MNL)
  - PCMC satisfies UE without being a RUM

vs.

Key properties of PCMC model

- **No** assumptions of **transitivity**, **IIA**, or **regularity**
  - No regularity means not even a RUM!
  - Even “Elimination by Aspects” (Tversky ’72) is a RUM.
- PCMC **does** satisfy axiom of **uniform expansion** (Yellott ’77)
  - “probabilities are unchanged by making copies of the set”
  - UE in an independent RUM implies Luce’s Axiom (and thus MNL)
  - PCMC satisfies UE without being a RUM

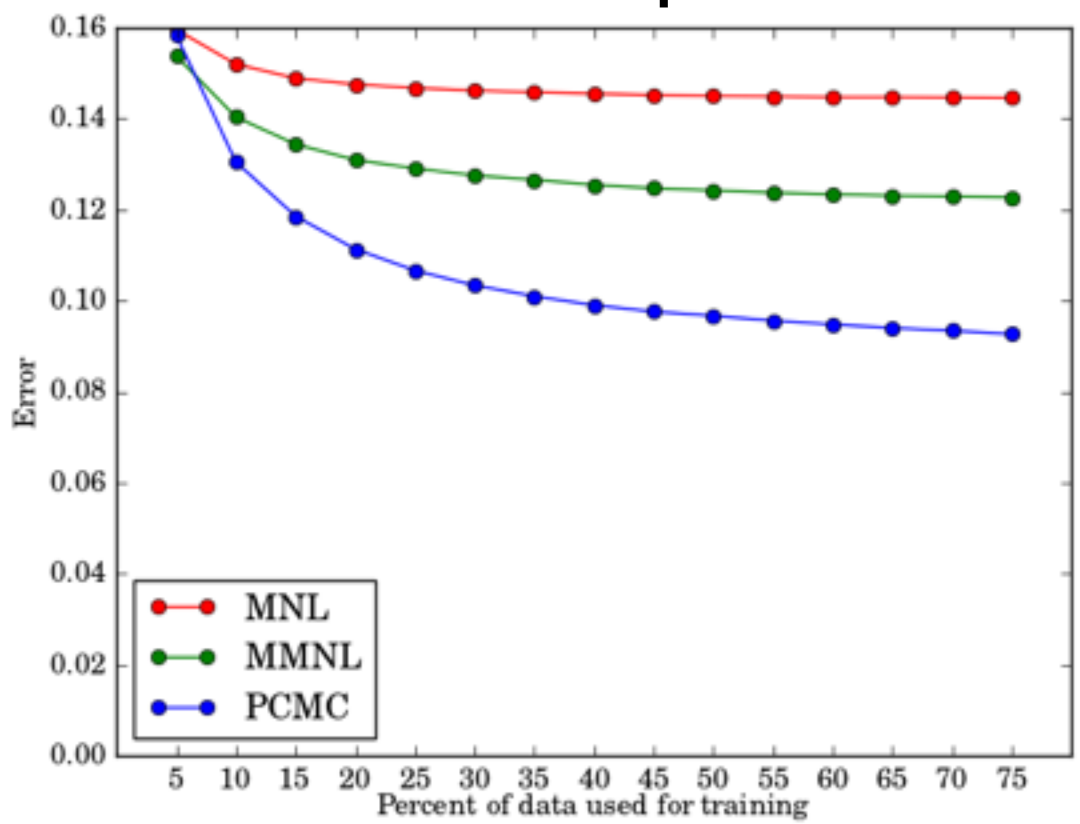
vs.

- We also generalize UE to a stronger property we call **contractibility** (addresses a thought experiment by Debreu)

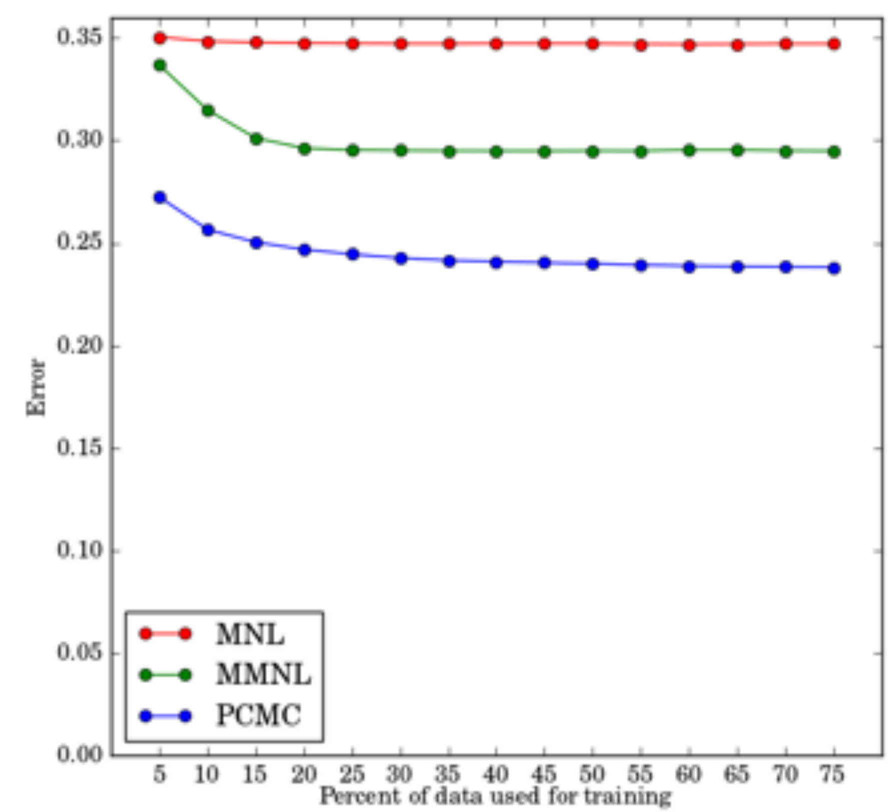
PCMC Predictions

- Dataset: transportation choices around SF for commuting and shopping.
- People had 2-8 options to choose from
- Many apparent violations of IIA

work trips



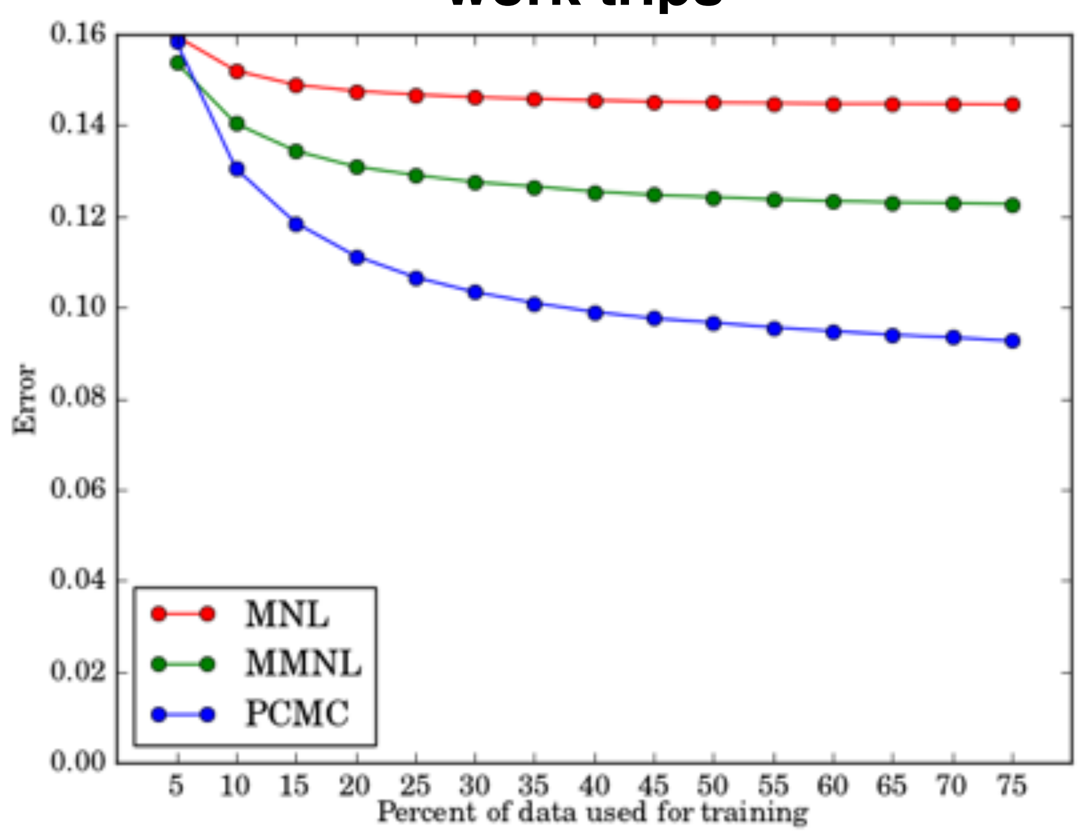
shopping trips



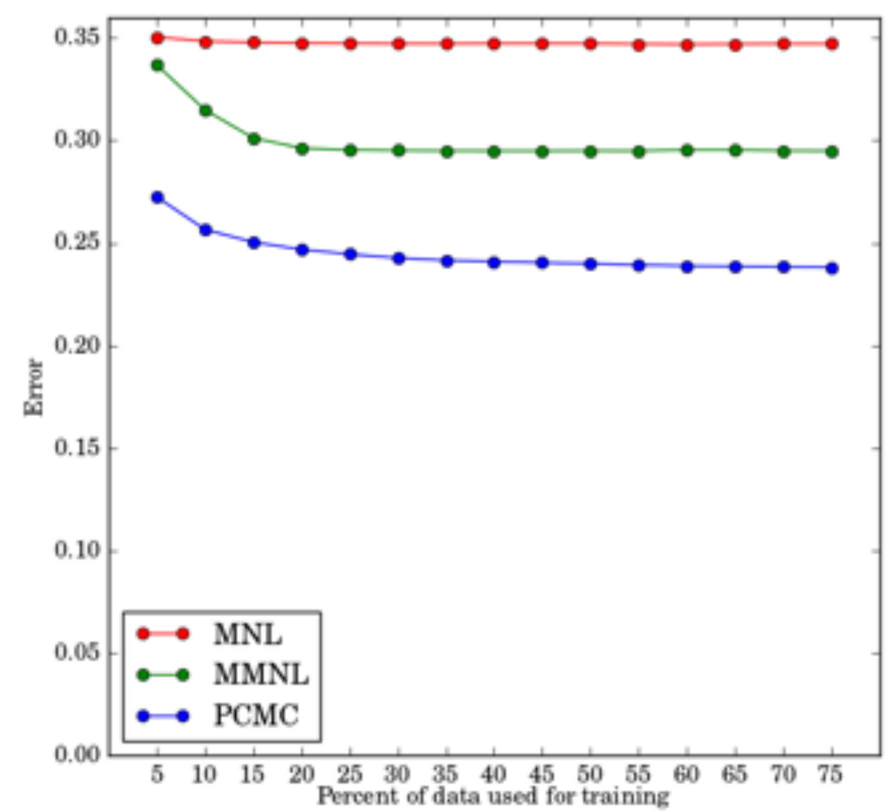
PCMC Predictions

- Dataset: transportation choices around SF for commuting and shopping.
- People had 2-8 options to choose from
- Many apparent violations of IIA

work trips



shopping trips

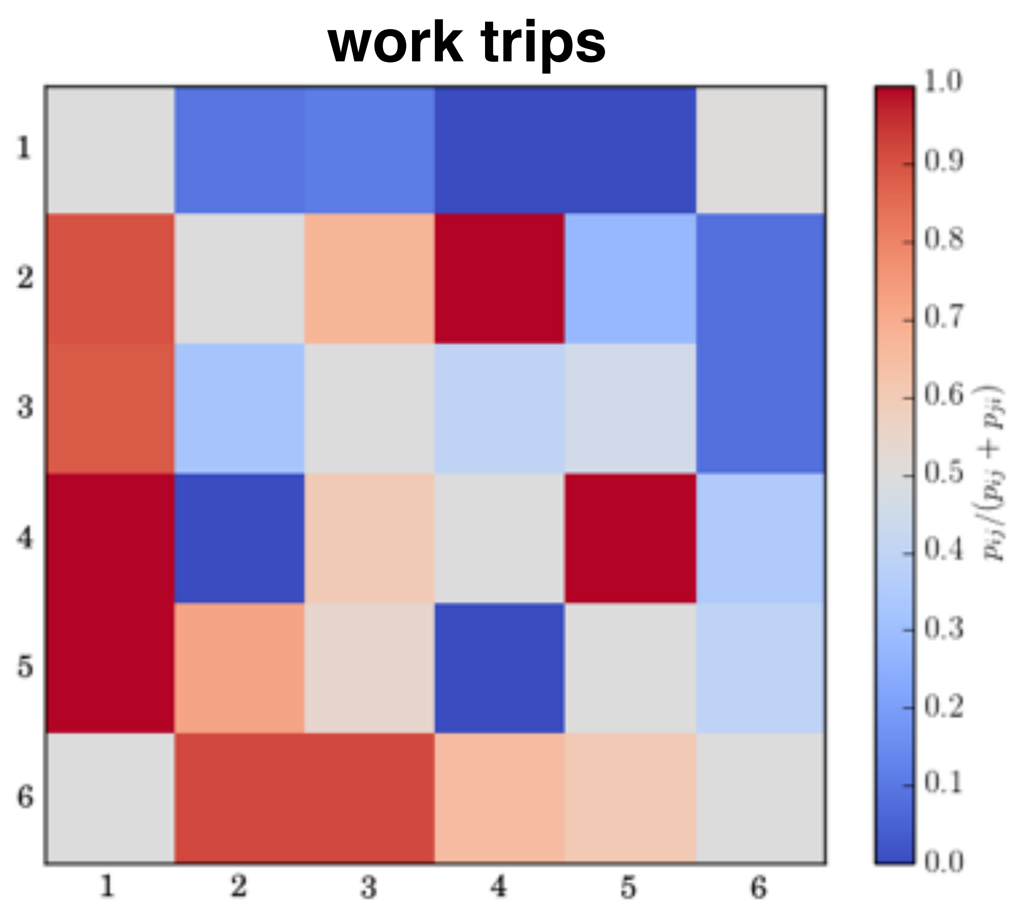


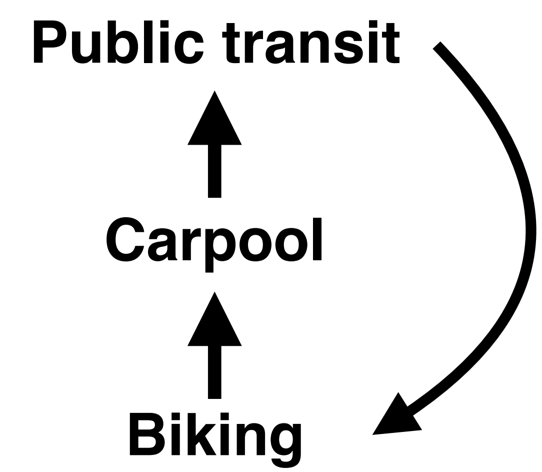
- In data with violations of IIA, PCMC does **20-30%** better at prediction out of sample. Without violations, PCMC falls back to MNL.

PCMC Pairwise Probabilities

- Inferred pairwise probabilities are highly non-transitive:

1. Driving alone
2. Carpool (1)
3. Walking
4. Public transit
5. Biking
6. Carpool (2+)

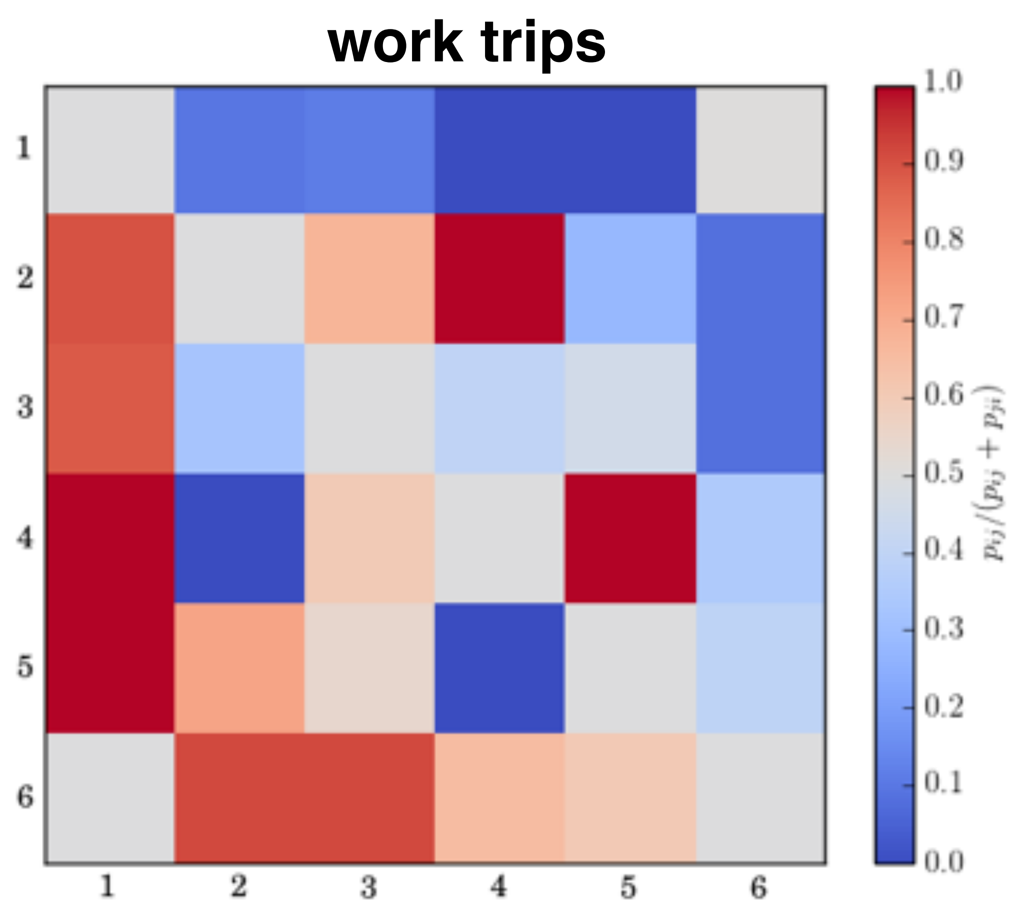


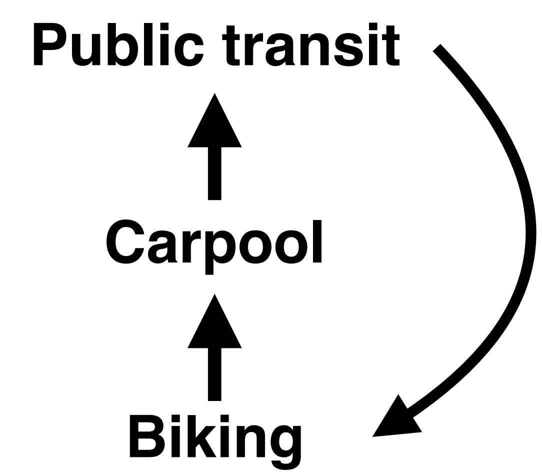


PCMC Pairwise Probabilities

- Inferred pairwise probabilities are highly non-transitive:

1. Driving alone
2. Carpool (1)
3. Walking
4. Public transit
5. Biking
6. Carpool (2+)





- Low-dimensional parameterization of pairwise probabilities:
  - Very recent "Blade-Chest" model (Shen-Joachims '16a, '16b) can embed/represent matrix Q with O(n) parameters without loss of performance.

Machine Learning Choices

- Applications:
  - **Testing:** When do choice set effects exist, when not?
  - **Learning:** what S to query to learn model with regret bounds?
  - **Design:** Given x, what set S maximizes probability of x?
  - **UX:** do predicted choice set effects persist when explained?
- Open modelling directions:
  - Incorporate covariates
  - “Choosing to Rank”
- Big questions:
  - Divergent goals of “Artificial Human Intelligence” vs. AI?
  - Ethics of libertarian paternalism in designed online systems?