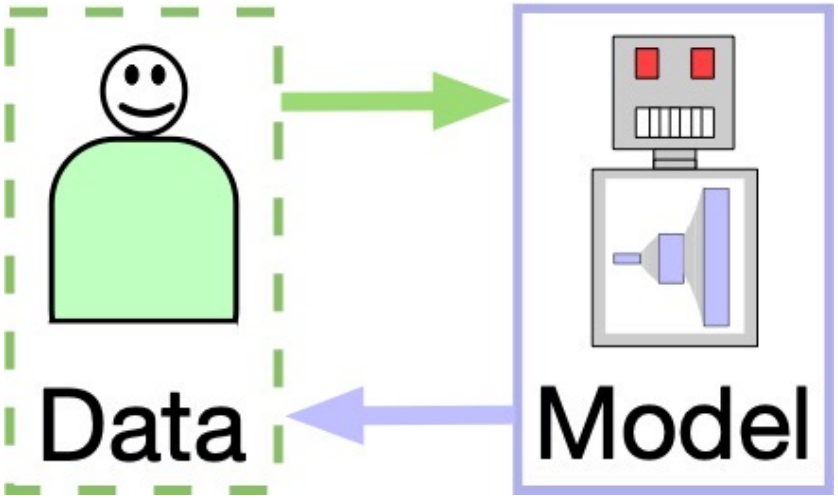
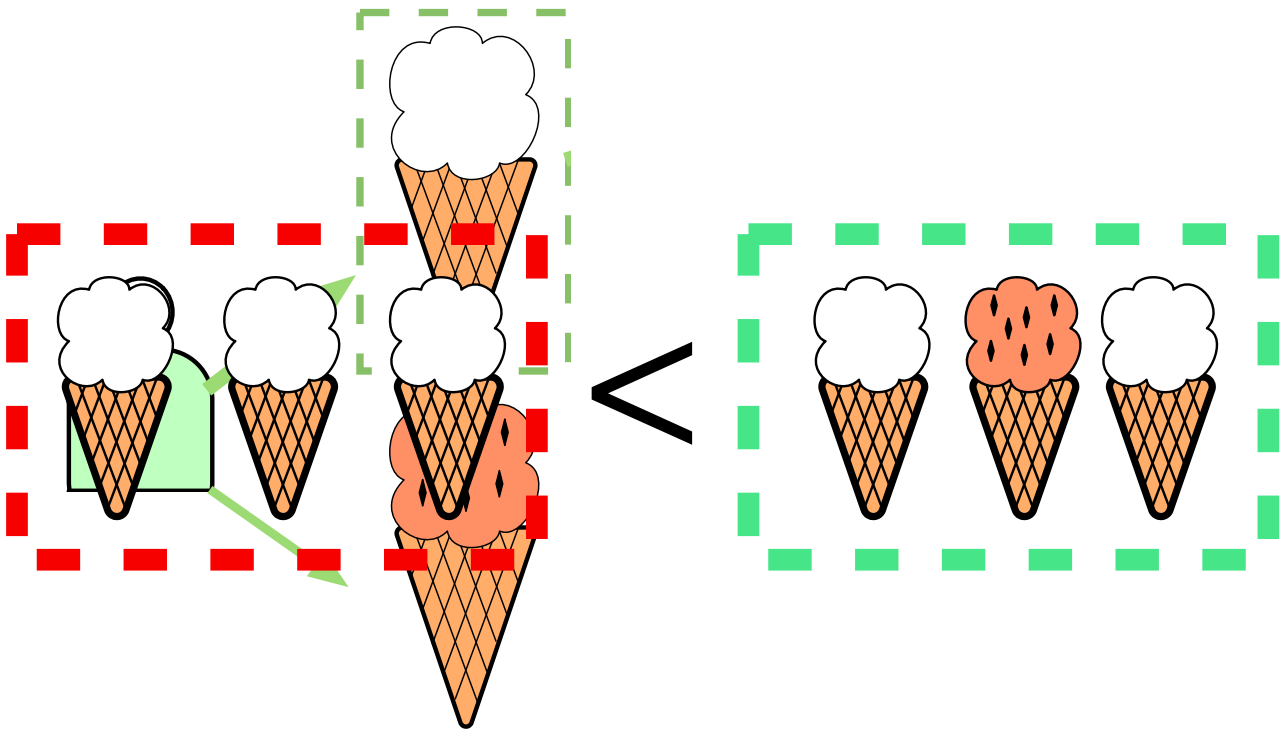
Runaway Feedback Loops in AI Systems

Keith Burghardt
Information Sciences Institute USC
Email: *keithab@isi.edu*
January 2022

What is a Feedback Loop?



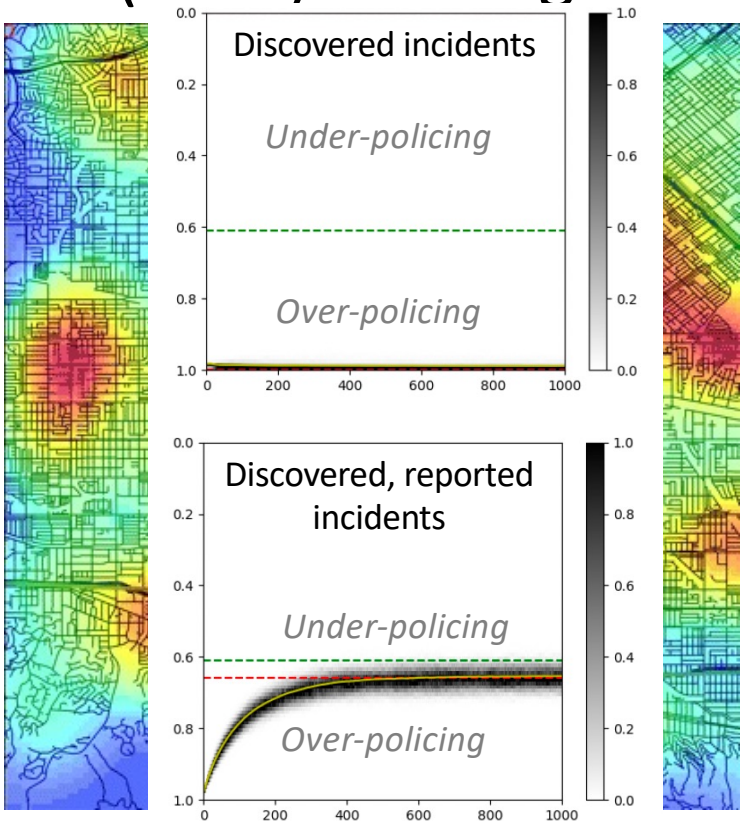
- *Interaction between users (data) and model*
- Models need data from users, but model affects user decisions
- New (potentially biased) decisions fed back into model



- Imagine you love ice cream! If given ice cream, you will eat it
- Slight preference for vanilla ice cream
- Model sees you choose ice cream, only recommends vanilla
- But you might prefer variety, or get sick of same flavor

Why Feedback Loops Can Be Unethical: Runaway Feedback and Predictive (Over)-Policing

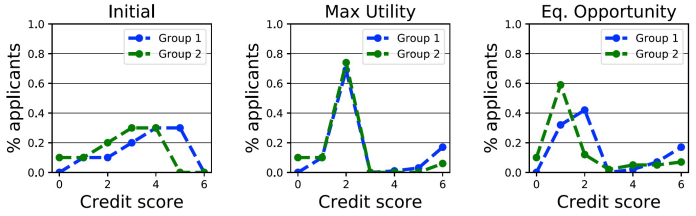
- Historically, there has been over-policing of minority neighborhoods
- PredPol was launched to improve policing (police only where necessary)
- Crime predictions based on past (biased) data
- *What impact does this have?*
  - **Feedback loop:** simulations show **over-policing** of higher crime neighborhoods
  - **Impact:** more crime, minorities unfairly targeted



Ensign, et al. FAT*'18.

AI Feedback Loops

- These feedback loops are common
  - Bank loans
    - use past data to recommend new loans
  - Bail
    - companies like COMPAS use past data to determine recidivism risk
  - Recommendation engines:
    - Recommend videos, social media, items, etc., from past data



- What problems do feedback loops create?

D'Amour et al., FAT* '20

Ethics of Feedback Loops

Fairness:

Focus of previous talks

Unfair predictions create more unfair policies

Filter Bubbles:

People see small window of content, increases polarization

Chaotic Trends:

Positive feedback loops create artificial, unpredictable trends

Takeuchi, *Plastic Love*

Undermining Utility of AI:

Poor predictions

Filter Bubbles

- Goal of recommendation systems is to offer you what you like and avoid what you do not
- This goal implies recommendation systems also create filter bubbles
  - Can force users to see less diverse content*
  - Could polarize users**
- Extremism is an ongoing problem, which recommendation systems could enhance without incentives for diverse content

*Ge et al., SIGUR '20

**Levy, American Economic Review, 2021

https://www.scientificamerican.com/article/calling-truce-political-wars/

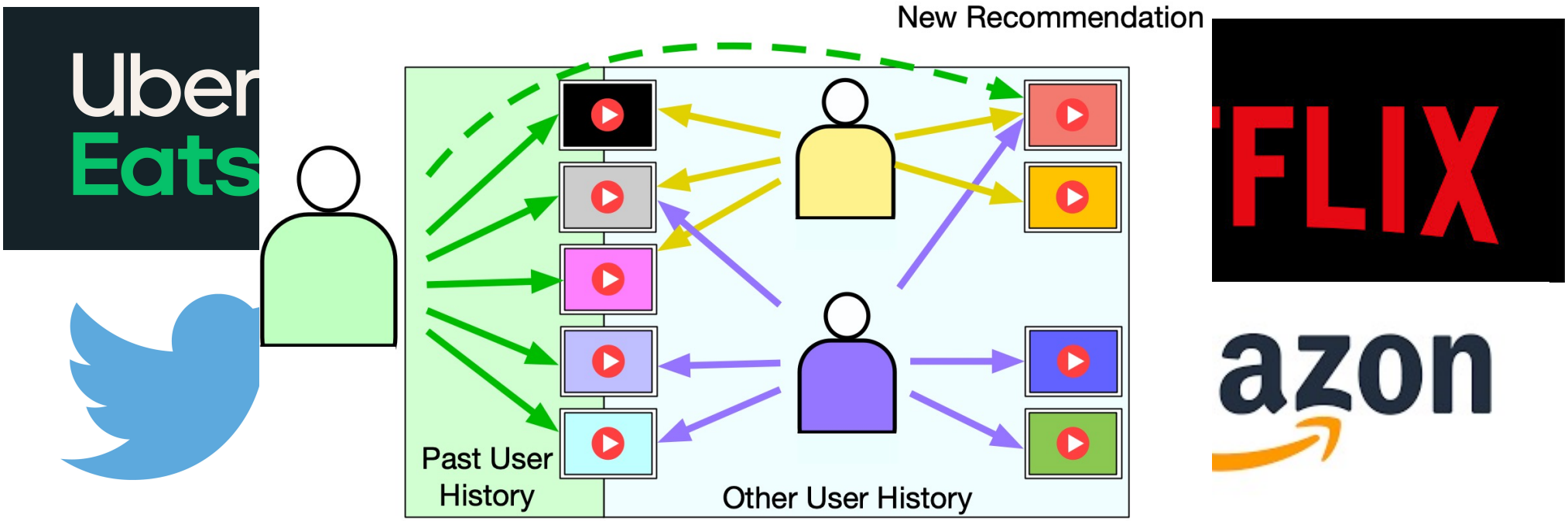
Chaotic Trends

- "City Pop" was a Japanese music genre popular in the 1980s
- 2018: *Plastic Love* widely recommended by YouTube, new popularity of City Pop
- Example of artificial new trends
  - (I say this as a fan of City Pop)
- Recommendation systems offer new way to increase popularity of content (whether good or mediocre)

https://www.openculture.com/2018/10/youtubes-algorithm-turned-obscure-1980s-japanese-song-enormously-popular-hit-discover-mariya-takeuchis-plastic-love.html

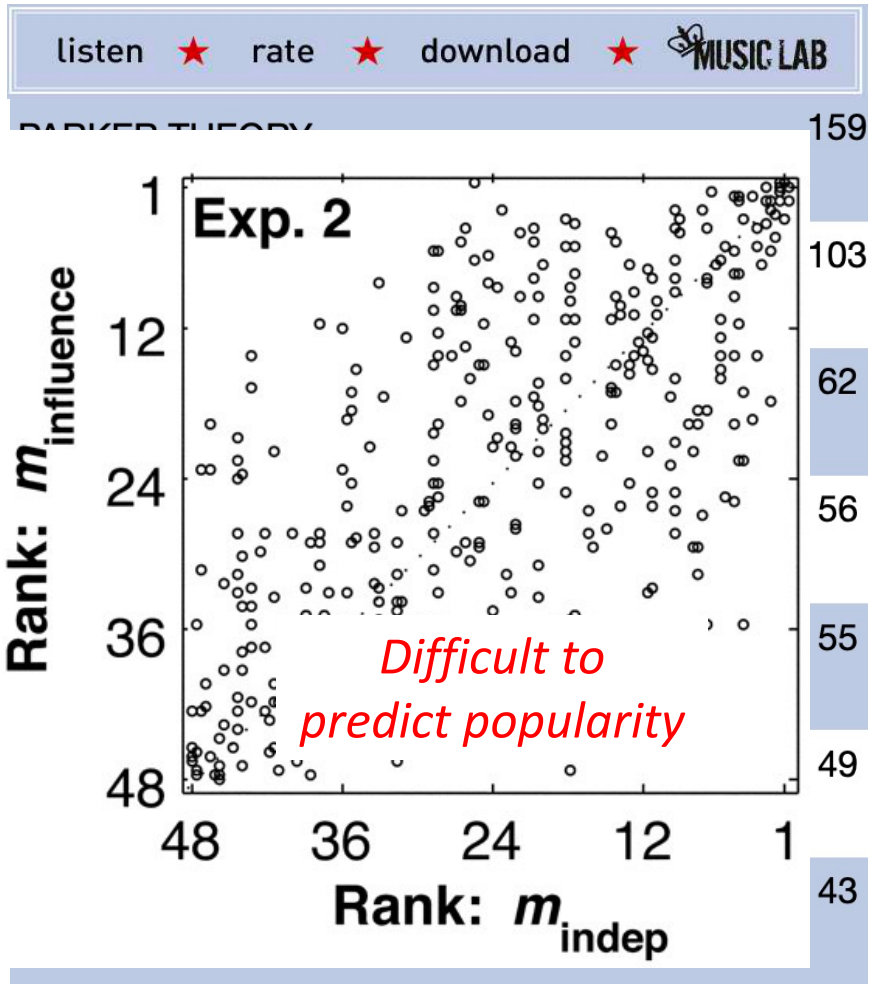
Chaotic Trends & Poor Predictions in Recommendation Systems

- Personalized recommendation is common AI task
- We will explore chaotic trends, predictions in these systems



Chaotic Trends

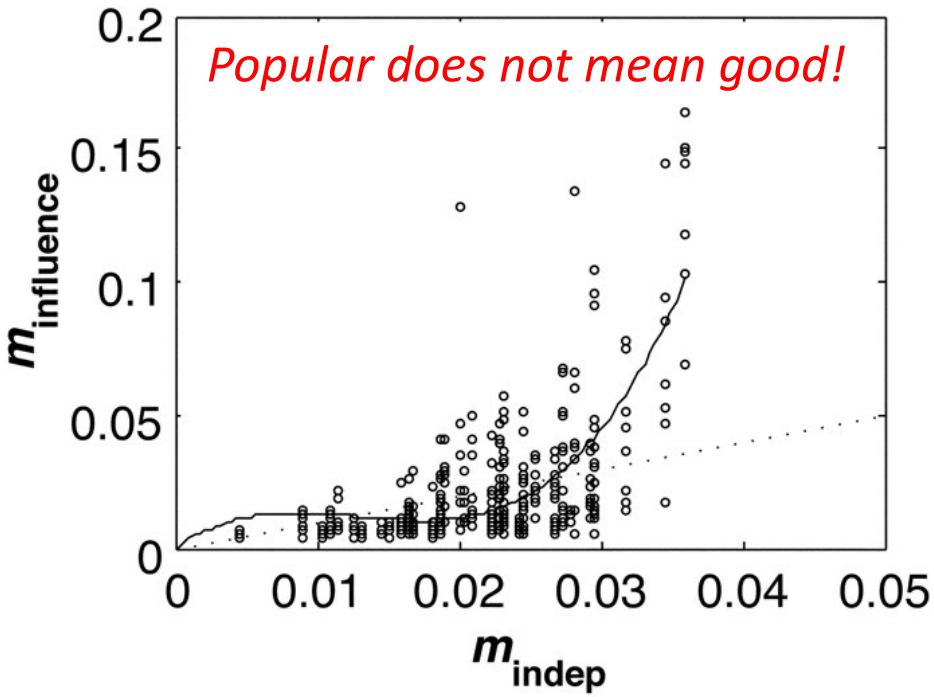
- Simple feedback loops show chaotic trends (*instability*)
- MusicLab study:
  - Rate music sorted by popularity
  - Expect good content to be at top
  - Same song could be popular or unpopular from initial conditions!
- Algorithm: *use past decisions* to rank content
- Result: features besides quality dictate what content people see



Salganik et al., *Science* (2006)

Poor Predictions

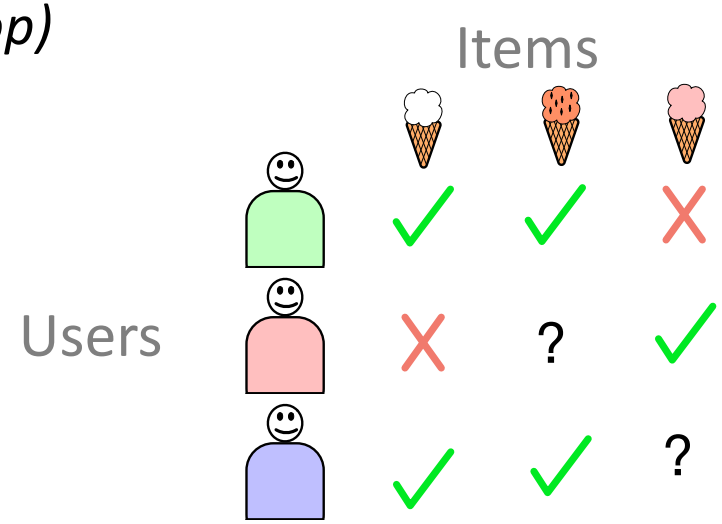
- Plot of popularity versus quality proxy: non-linear, noisy relationship
  - Popular content is poorly correlated with "good" content
- This simple example shows poor prediction can exist
- But how does this extend to "typical" AI recommendations?



Salganik et al., *Science* (2006)

AI Recommendations

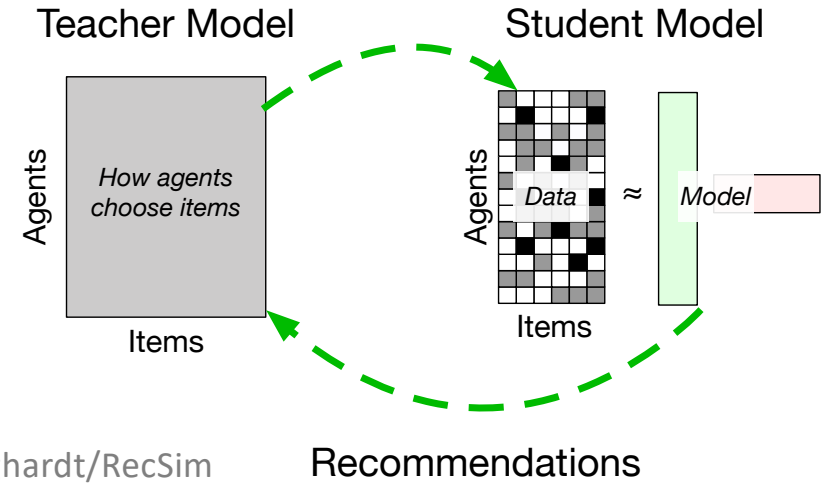
- Recommendations represented as table of users and “items”
- Examples: Users and the videos they watch, ice cream they like
- Model fit to data -> recommend content -> data fed back into model *(feedback loop)*



Models try to guess missing data

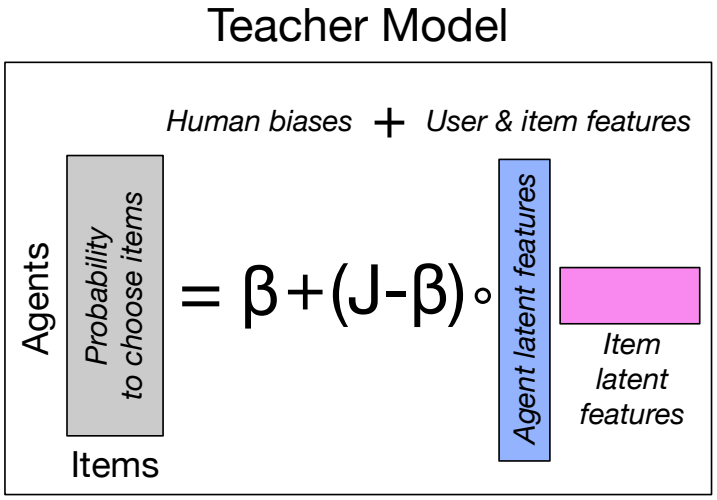
Auditing Recommendation Algorithms

- We can simulate how feedback loop works!
- *Teacher-student framework:*
  - **Teacher model:** Simulate how people will choose videos
  - **Student model:** fit available data (try to match teacher model)



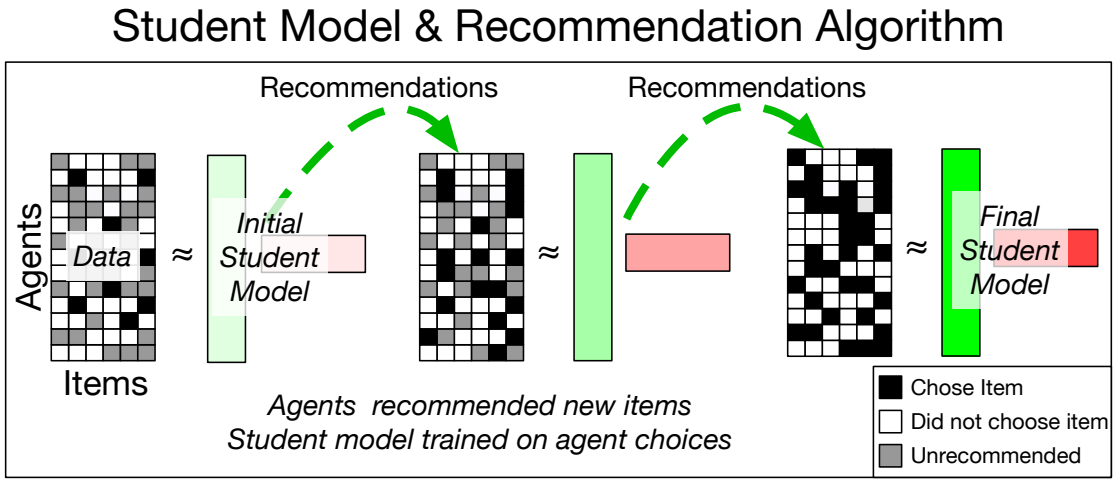
Code: https://github.com/KeithBurghardt/RecSim

Simulation Framework



- Teacher model assumes items are chosen because of intrinsic features (user preferences for action movies, for example)
- Human biases: choosing content at random, or just because it is shown to you

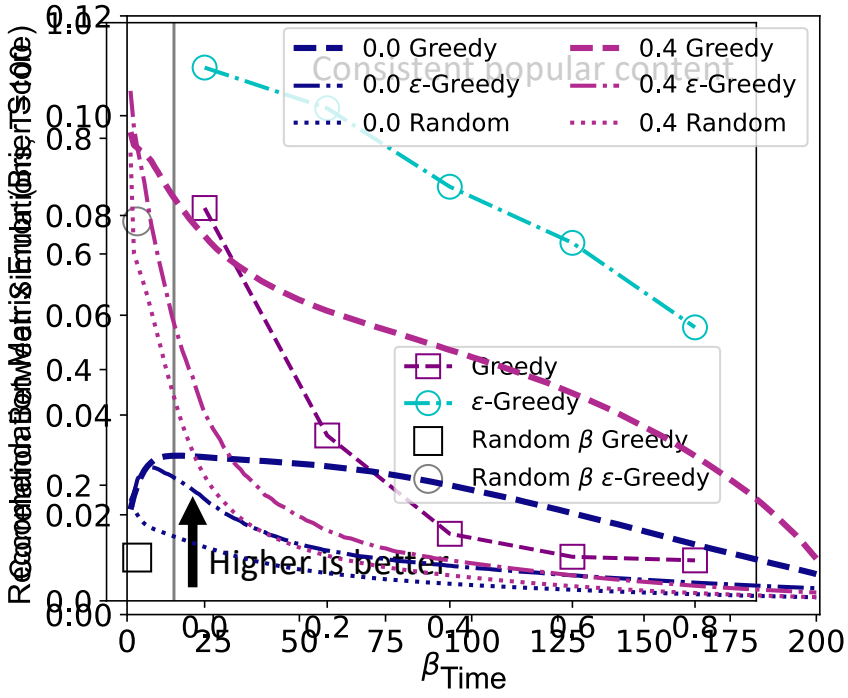
Simulation Framework



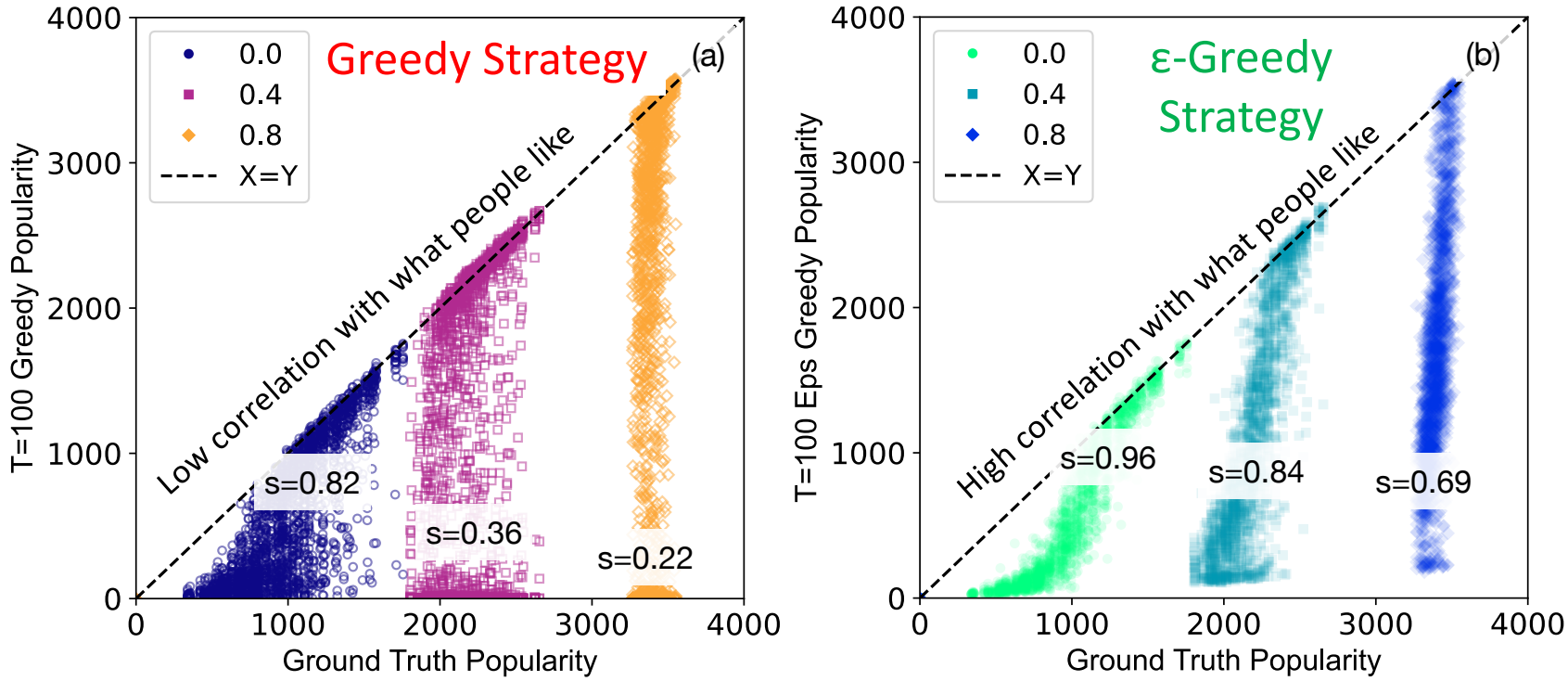
- Train model to data (assumes users prefer some latent item features)
- New recommendations made
- Model retrained on new & old data
- Simplified version of typical systems (only 2 choices, all users make decisions, etc.)

Chaotic Trends and Poor Predictions

- Test two different strategies
  - Greedy: *recommend estimated best content*
  - ε-Greedy: *sometimes offer random/unexpected content*
- Chaotic Trends
  - Greedy method: inconsistent trends
  - ε-Greedy method: more consistency
- Prediction error
  - Greedy method: higher error
  - ε-Greedy method: low error

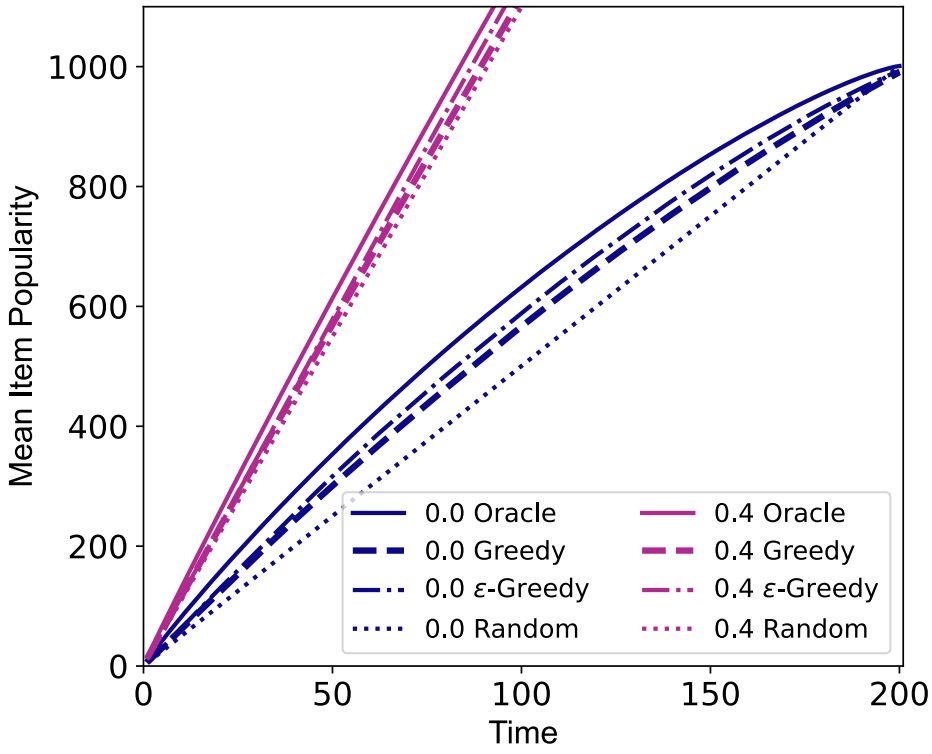


ε-Greedy Strategy Better Ranks Content



Randomness Improves Revenue

- Good predictions = more views
  - Randomly recommending content makes *better recommendations*!
  - ε-Greedy strategy increases views by ~3%
  - Small but significant improvement
- Some sites like Youtube may barely break even*
  - Small change in margins could change profitability



*https://www.investopedia.com/articles/personal-finance/053015/how-youtube-makes-money-videos.asp

Summary of Results

- ε-Greedy strategy has potential to make better recommendations on 3 fronts:
  - *Higher accuracy*
  - *More stable predictions*
  - *Fewer filter bubbles*
- **Future work:**
  - Impact on filter bubbles:
    - More diverse content could push people out of filter bubbles
  - Impact on other systems
    - Does randomness benefit other important decisions?
    - Is it ethical for some decisions to be random ("experiments" with real people)

Summarizing Feedback Loop in AI

- Feedback loop implies **maximizing accuracy is not everything**
- New strategies needed to uncover long-term impact of algorithms
- Problems of feedback loop
  - Fairness
  - Filter bubbles
  - Unpredictable trends
  - Poorer performance
- Nuanced understanding of feedback loop could make better algorithms

Reducing Feedback Loops In Other Systems

- Mitigating this problem is generally hard
- But previous results show auditing and improving algorithms is critical
  - *Status quo:* Improve algorithms with better models and more data
  - *Critique:* It is not so much the model you use but how you use it
    - Simulations used the same model, just difference recommendations
- Potential benefits
  - More loans to successful businesses
  - More accurate credit scores (improve home-buying, increase wealth)
  - Better policing *if done correctly*
  - Reducing polarization
  - Less online extremism (offline I can discuss latest work on this front)

Thank you!