Tell me Why? Tell me More! Explaining Predictions, Iterated Learning Bias, and Counter-Polarization in Big Data Discovery Models

CCS@Lexington, October 16, 2017

Olfa Nasraoui

This work is a Collaboration with:

Behnoush Abdollahi, Mahsa Badami, Sami Khenissi, Wenlong Sun, Gopi Nutakki, Pegah Sagheb: @UofL

& Patrick Shafto: @Rutgers-Newark

Knowledge Discovery & Web Mining Lab

Computer Engineering & Computer Science Dept.

University of Louisville

http://webmining.spd.louisville.edu/

olfa.nasraoui@louisville.edu

Acknowledgements:

National Science Foundation:

NSF INSPIRE (IIS)- Grant #1549981

NSF IIS - Data Intensive Computing Grant # 0916489

Kentucky Science & Engineering Foundation: KSEF-3113-RDE-017

Outline

- What can go Wrong in Machine Learning?
  - Unfair Machine Learning
  - Iterated Bias & Polarization
  - Black Box models
- Tell me <u>more</u>: Counter-Polarization
- Tell me <u>why</u>: Explanation Generation

What Can Go Wrong in Machine Learning?

"Twitter and Facebook can't predict the election, but they did predict what you're going to have for lunch: a tuna salad sandwich. You're having the wrong sandwich."

What Can Go Wrong in Machine Learning?

- We are relying on Machine Learning (**ML**) algorithms to support decisions:
  - **Recommender Systems:**
    - They **guide** humans in discovering **only a few** choices from among **a vast space** of options
    - **Choose among options:** Reading the News, Watching movies, Reading books, Discovering friends, Dating, Marriage, etc
  - **Supervised Learning:**
    - **Predict class label** for given instance
      - Example of label: whether to approve a loan, etc
    - Credit Scoring, Criminal investigation, Justice, Healthcare, Education, Insurance risk modeling, etc

What Can Go Wrong in Machine Learning?

Real life data can include **biases** that can affect the predictions

- May result in **unfair** ML models
  - discriminative,
  - unreasonable,
  - biased...

  - **worse** when models are opaque/black box!

What Can Go Wrong in Machine Learning?

- Increasing (unchecked) Human-ML algorithm interaction...
  - Think about **Recommender Systems**
    - They **guide** humans in discovering **only a few** choices from among **a vast space** of options
    - Why are they needed?
      - Information Overload ⇒ need **Relevance Filters**!
    - But ...
      - could result in **hiding** important information from humans
      - could exacerbate **polarization** around divisive issues
      - could **fail to explain why they recommend** a particular choice (**Black Box** models: e.g, Matrix Factorization, Deep Learning)

What Can Go Wrong in Machine Learning?

Increasing unchecked Human-ML algorithm interaction...

Need for:

- Understanding Impact of interaction
- Limiting or reversing biases

  ⇒ Tell Me **More**!

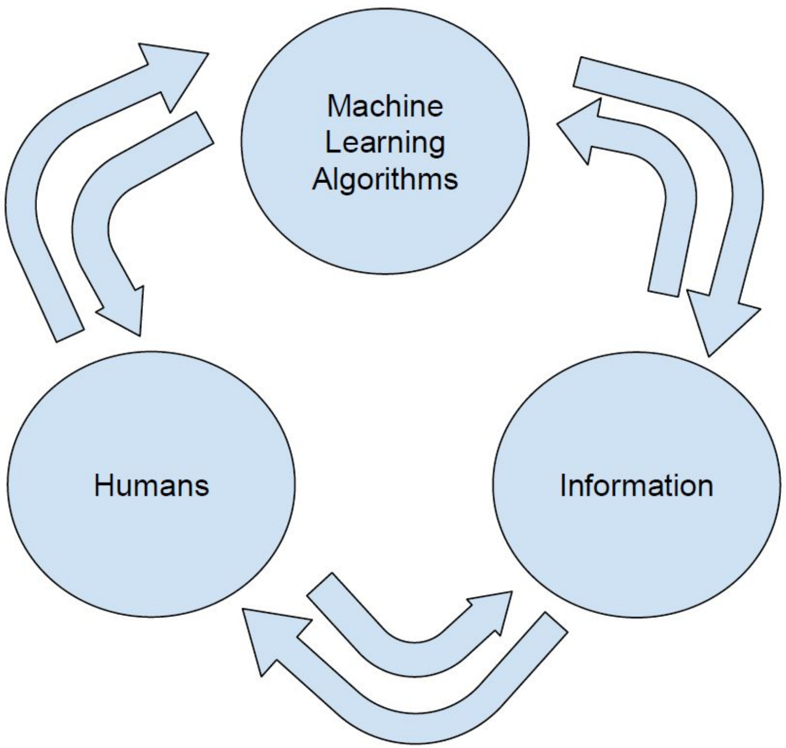
- Adding Transparency / Explanations
  - to scrutinize biased or incorrect predictions
  - ⇒ more trust in ML models!

  ⇒ Tell Me **Why**?

Outline

- What can go Wrong in Machine Learning?
  - Unfair Machine Learning
  - Iterated Bias & Polarization
  - Black Box models
- **Tell me more: Counter-Polarization**
- Tell me why: Explanation Generation

Iterated Bias



Machine Learning: Now & Then...

- In the **past**, Machine learning algorithms relied on **reliable** labels from experts to build predictive models.
  - **Expert** users, **limited** data, **reliable** labels

- **Today**, algorithms receive data from the **general population**
  - Labeling, annotations, etc.
  - **Everybody** is a user, **Big Data**, **subjective** labels

- **Labeled** Data (User **Relevance labels**)

  ⇒ Machine Learning **Models**

  ⇒ **Filtering of information visible to the user**

  ⇒ **Next Labeled Data**

  ⇒ **Next ML Model**

  … etc

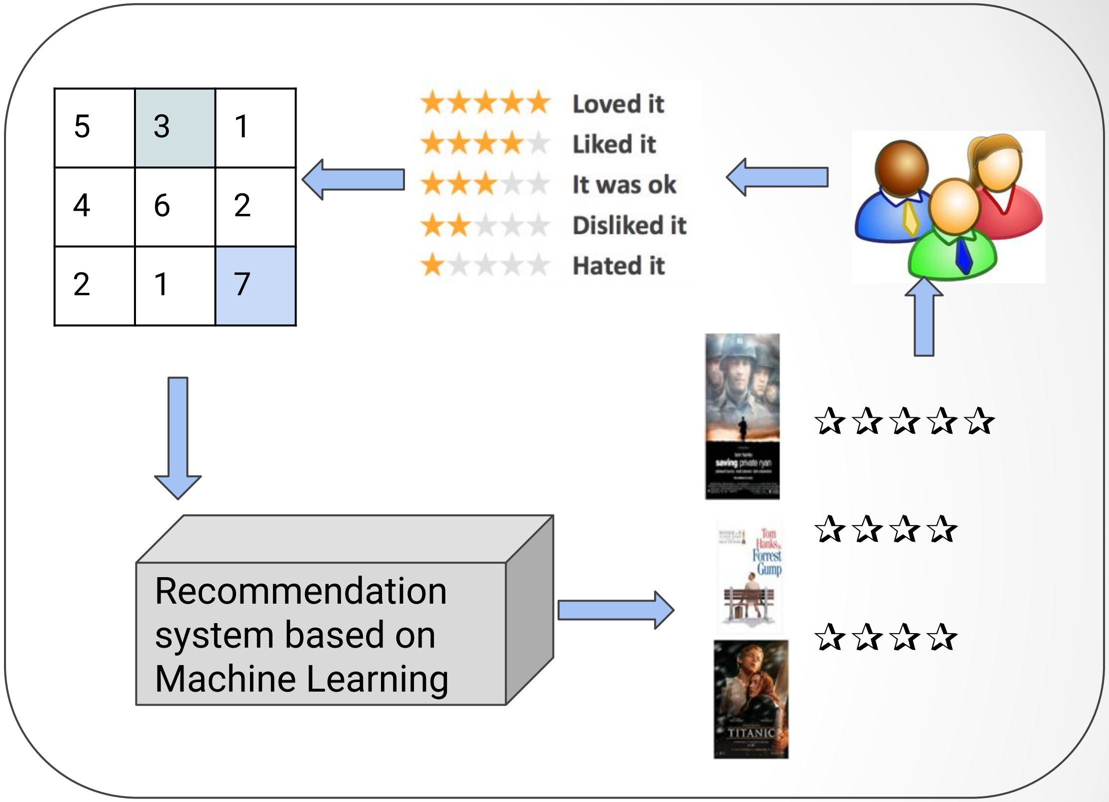
  ⇒ **Bias!** **Iterated Learning Bias!**

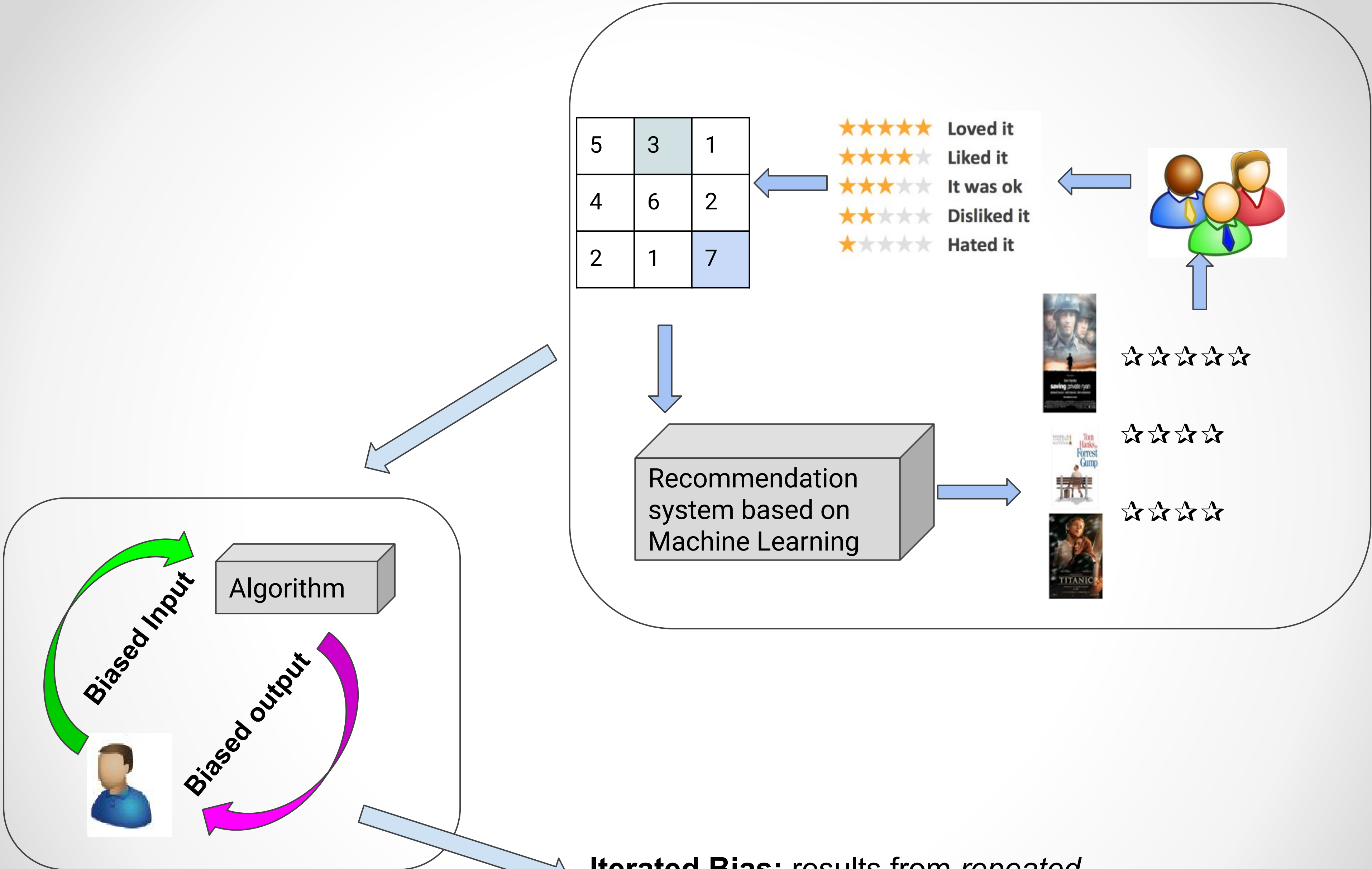
Recommender Systems

Collaborative Filtering

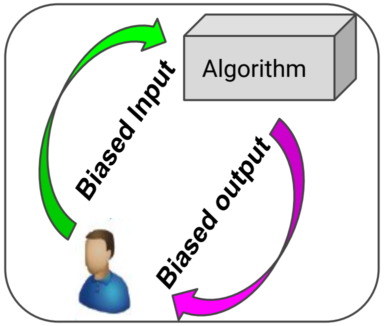
Uses previous ratings of the user to predict future preferences

Recommender Systems ⇒ Iterated Bias





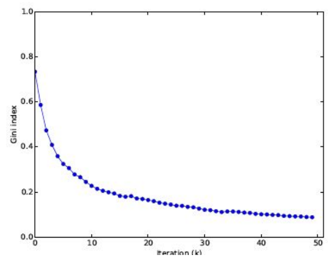
Iterated Bias: results from *repeated* *<u>interaction</u> between user & algorithm*



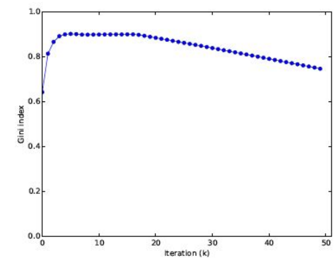
Impact of Iterated Bias on Predicted Ratings

- Collaborative Filtering Simulation: Item-based, U=100, N=200
- **Gini Index of the rating distributions vs iterations** between **rater** and **algorithm**

Open loop



Closed loop



- **<u>Feedback loop</u> / interaction between rater and recommender**
  - ⇒ Increases the divergence between ratings (Likes / Dislikes)
  - ⇒ We are witnessing **<u>the birth of polarization</u>**

Note: **Existing public benchmark data sets are useless for studying this problem!**

(1) they do not record every interaction

(2) they do not have the absolute user preference on each item!

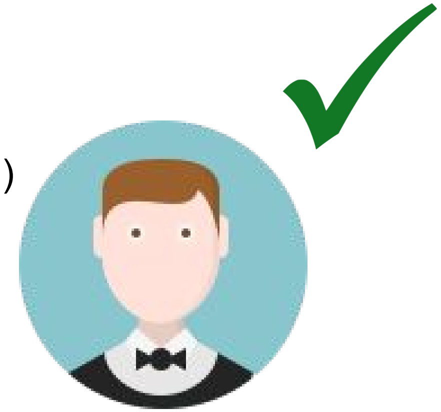
⇒ **Need Benchmark human choice and rating <u>cognitive</u> <u>models!</u>**
(Shafto & Nasraoui, 'Human-Recommender System' RecSys 2016)

Polarization
&
Counter-Polarization
in Recommender Systems

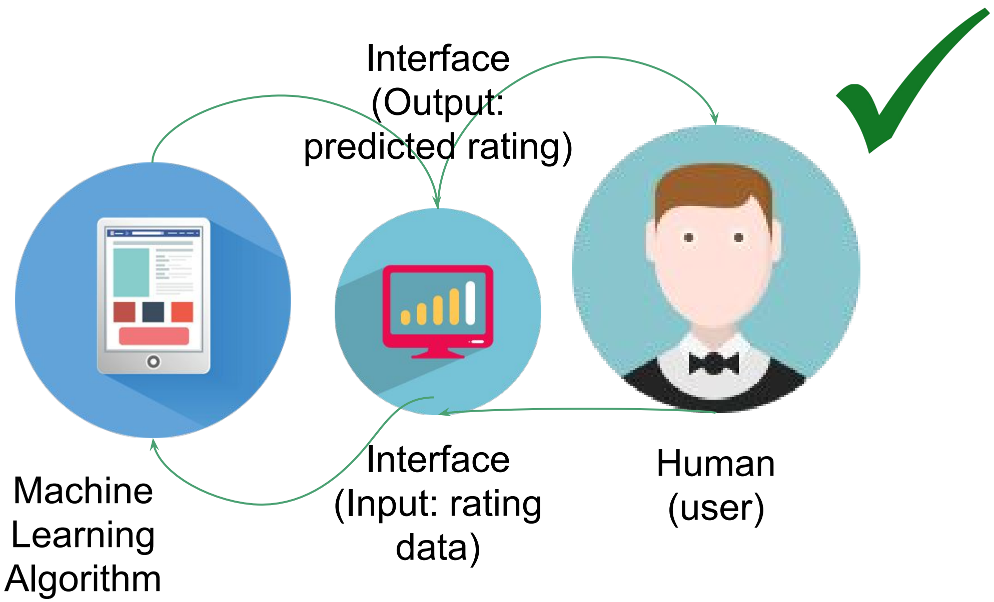
Machine Learning Algorithm

Interface (Output: predicted rating)

Interface (Input: rating data)



Human (user)

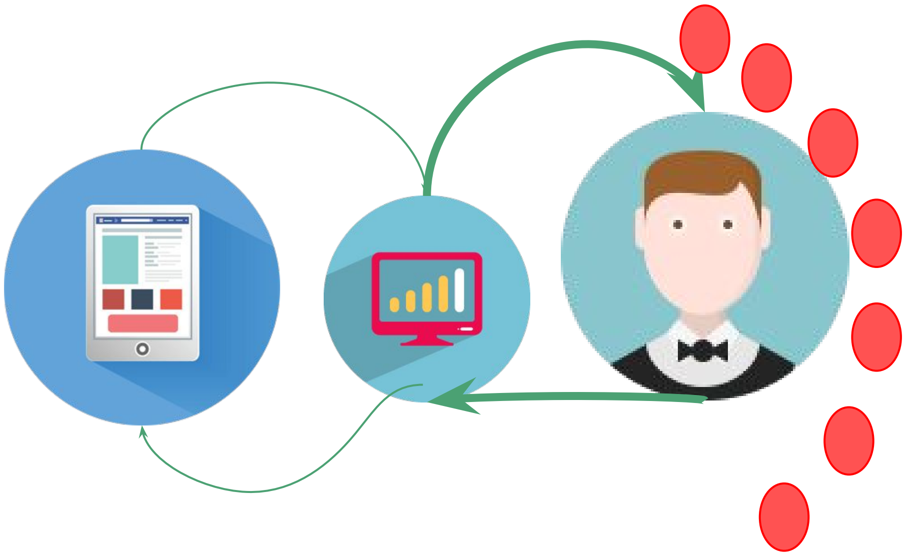


Positive Feedback Loop

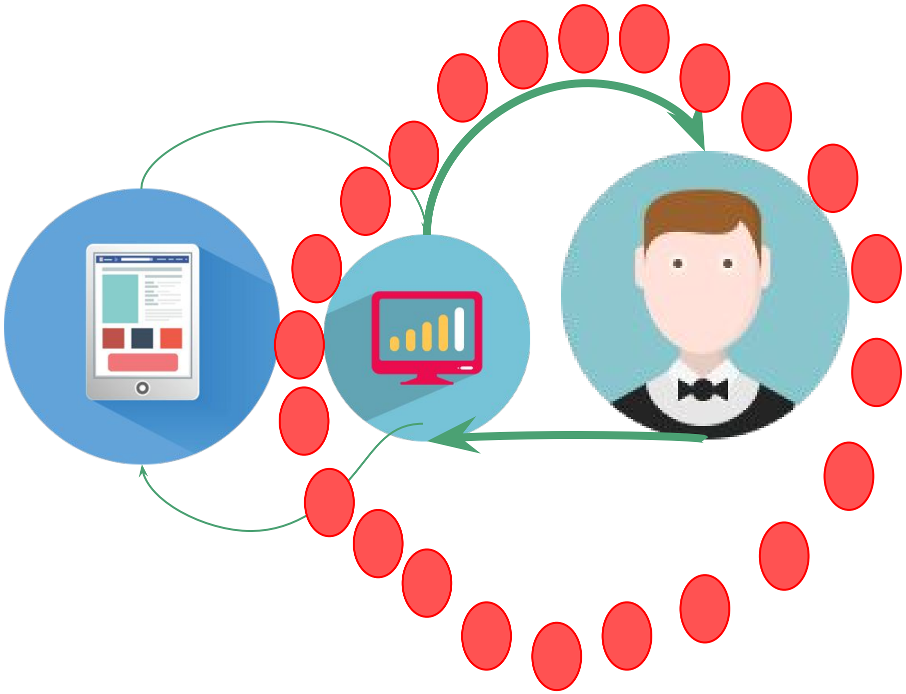
Limited Historical Data

Temporal Discounting

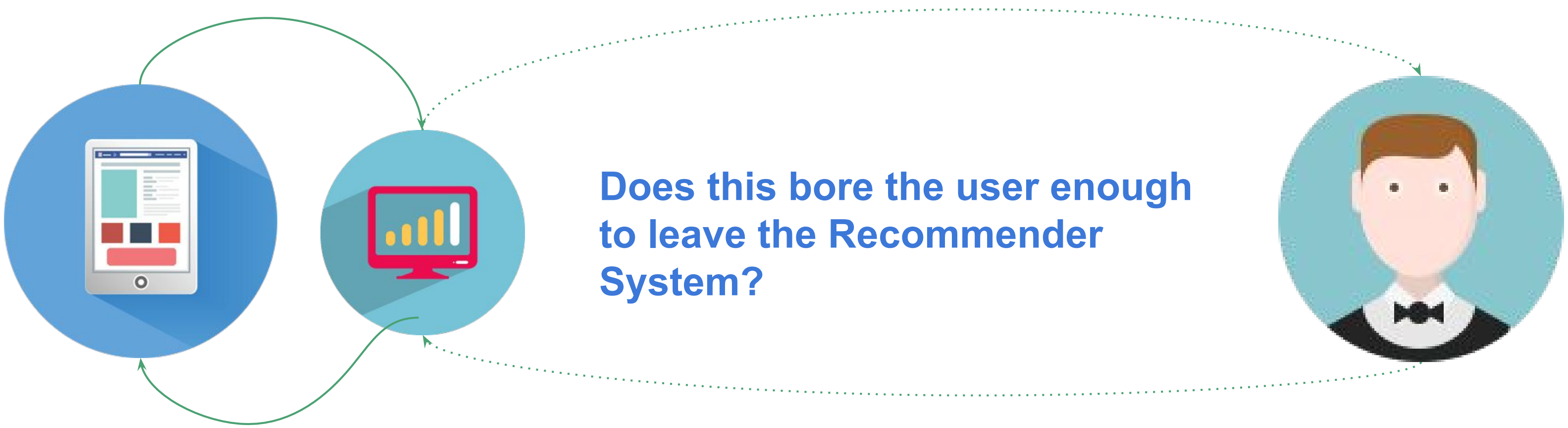
Positive Feedback Loop



Positive Feedback Loop



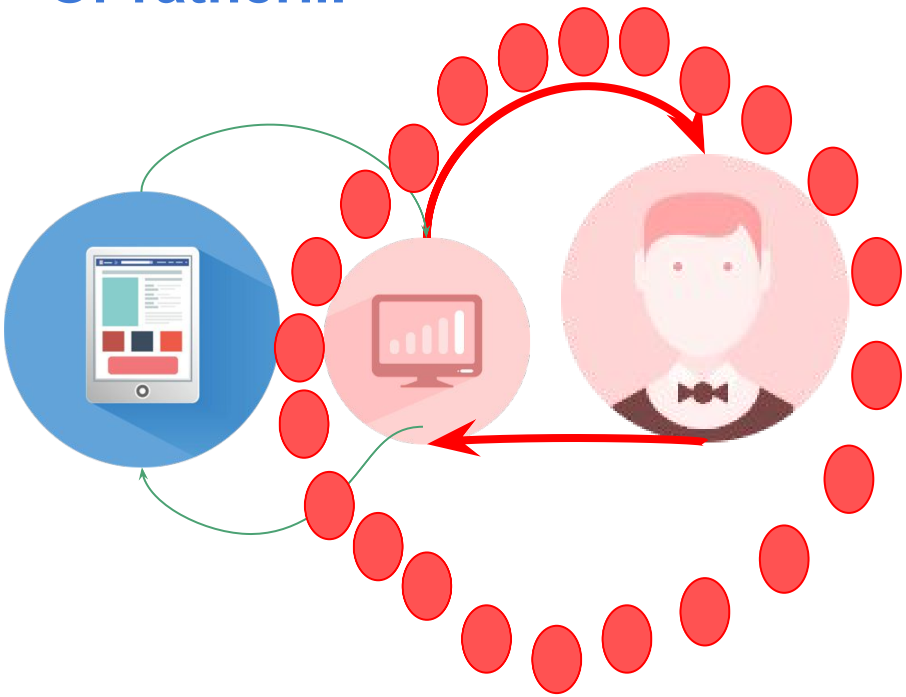
Filter Bubble



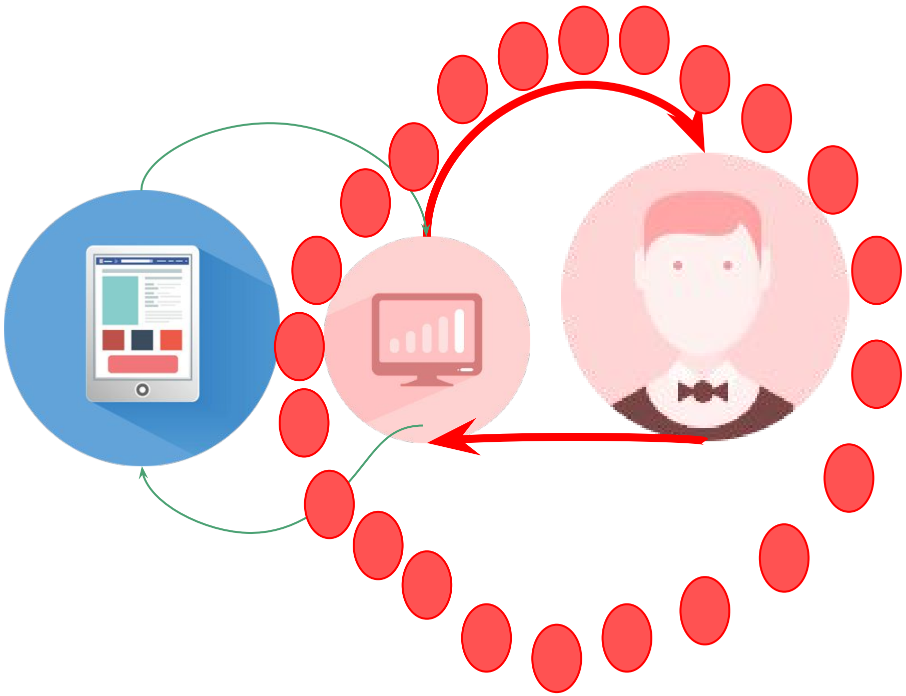
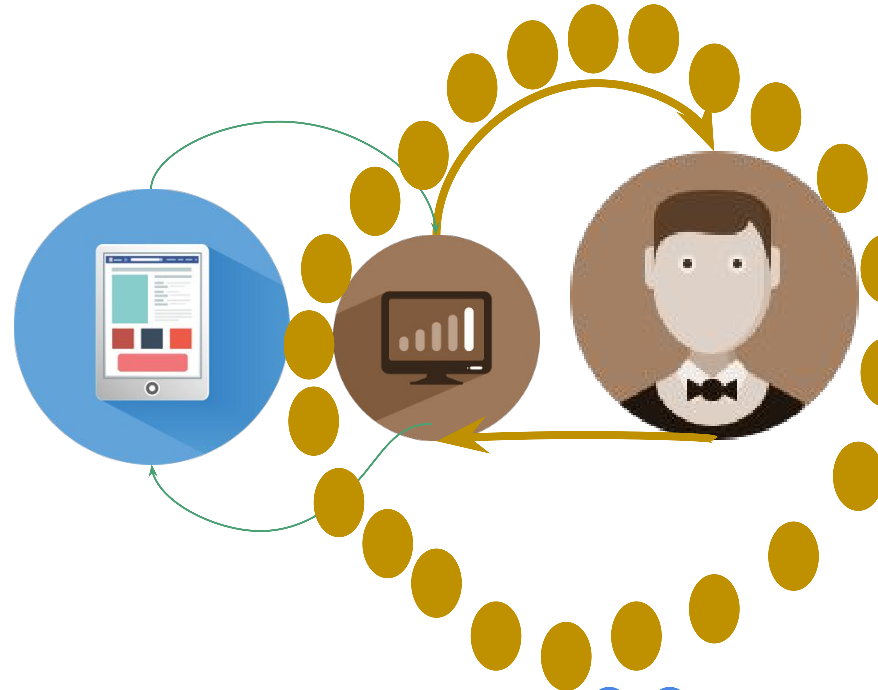
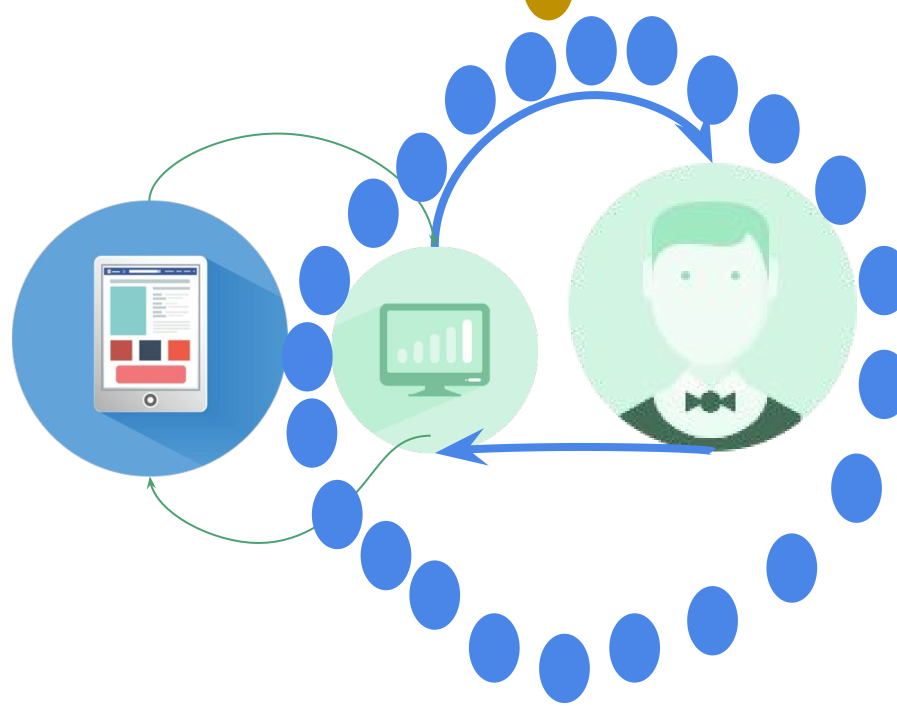
Does this bore the user enough to leave the Recommender System?

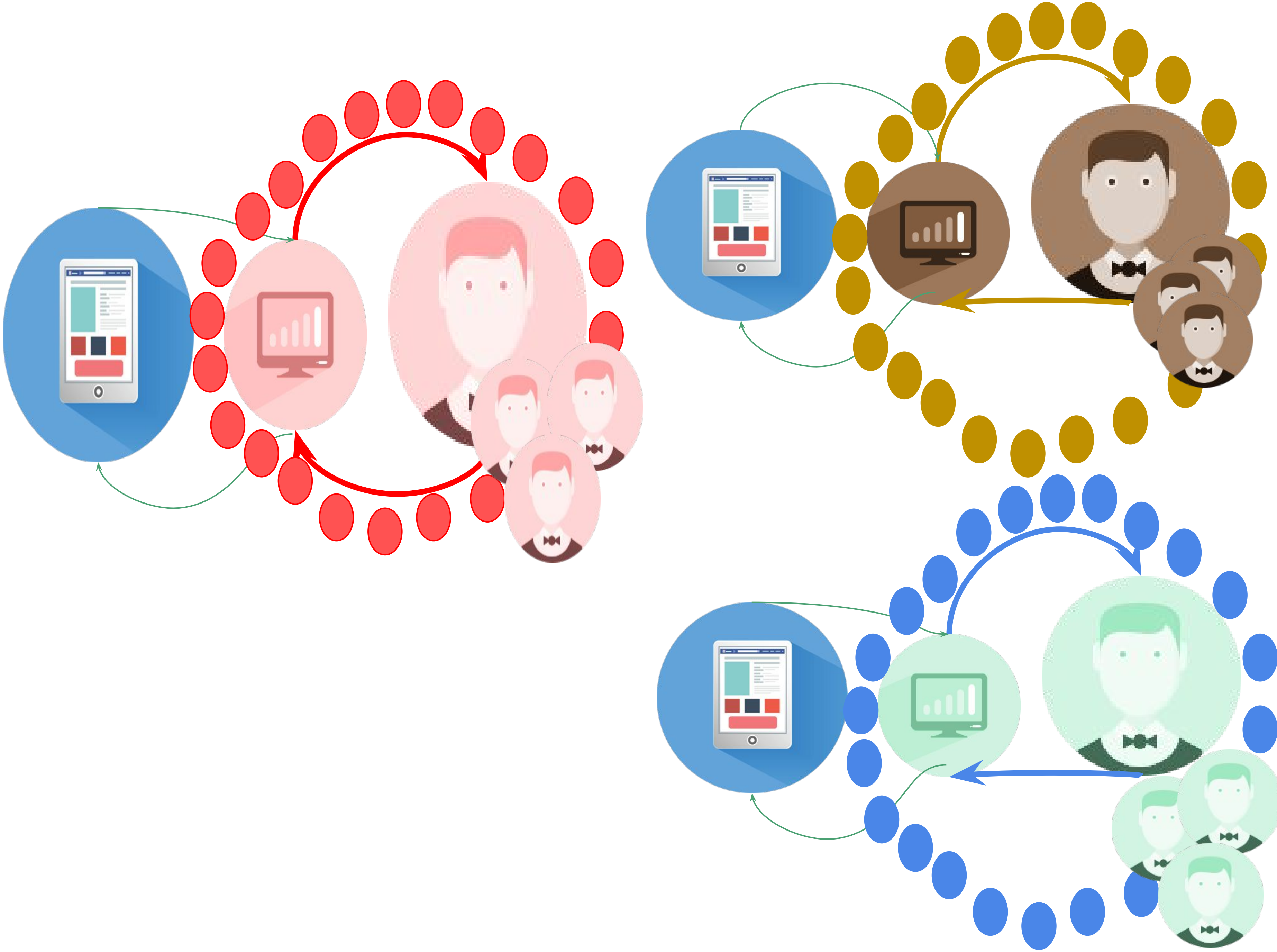
Filter Bubble

Or rather...

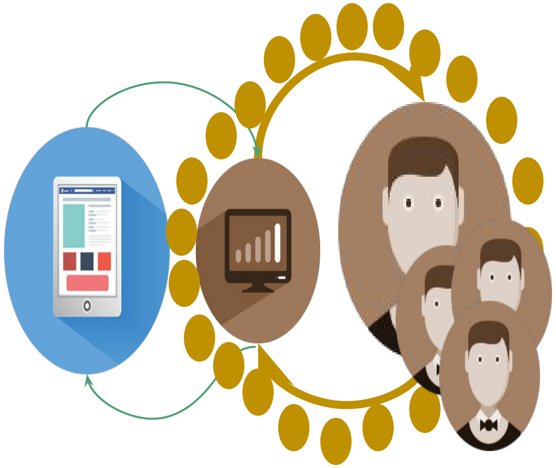


Self-fulfilling Identity

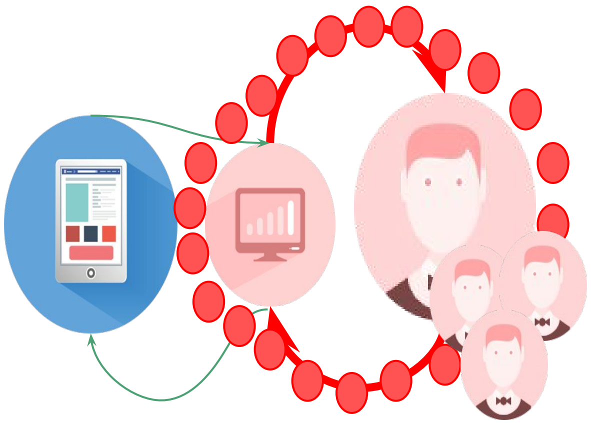






Consequences



Over Specialization



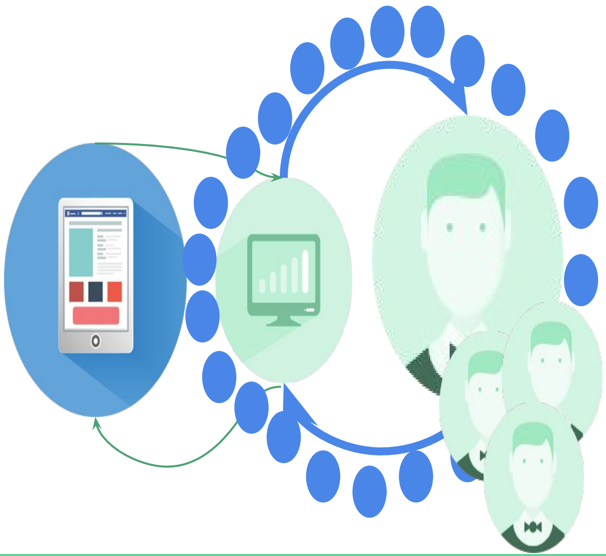
User Unsatisfaction

Polarization

Misperceiving Facts

Deconstructing non-prevailing views, opinions and behaviors

Extreme Attitudes



Low Sales Rates

It gets worse in a *Polarized* environment!

Merriam-Webster

Definition of POLARIZATION

1 : the action of polarizing or state of being or becoming polarized: such as

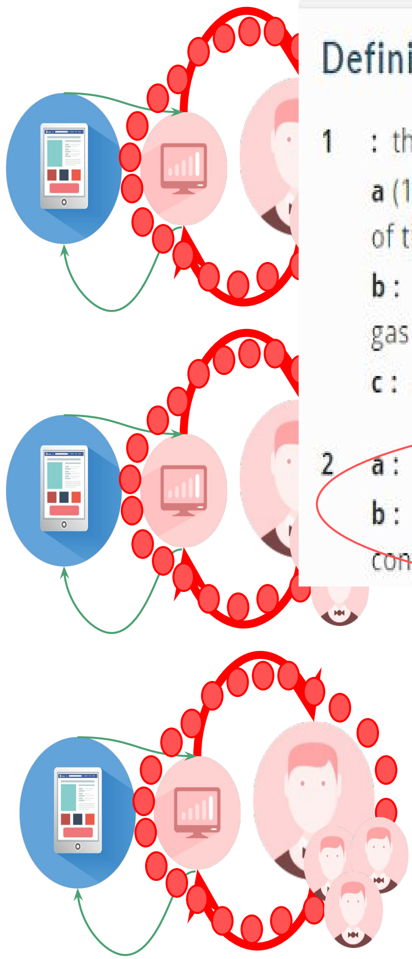
a (1) : the action or process of affecting radiation and especially light so that the vibrations of the wave assume a definite form (2) : the state of radiation affected by this process

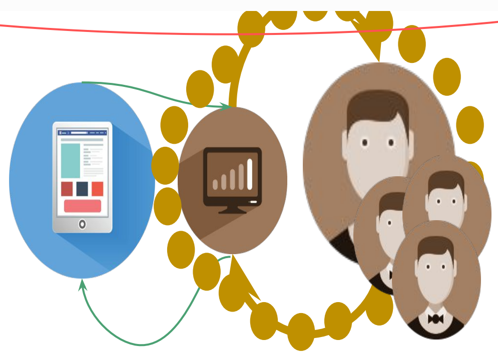
b : an increase in the resistance of an electrolytic cell often caused by the deposition of gas on one or both electrodes

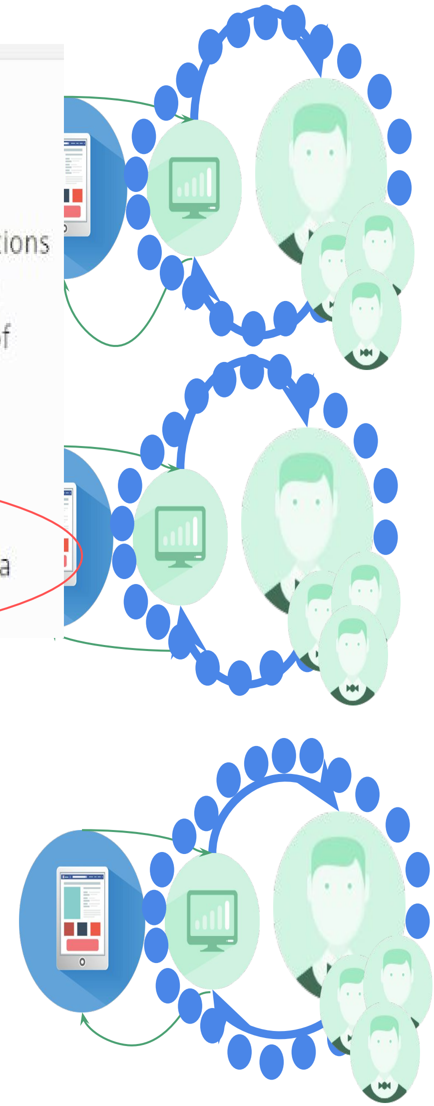
c : MAGNETIZATION

2 a : division into two opposites

b : concentration about opposing extremes of groups or interests formerly ranged on a continuum







Polarization

Our survey ⇒

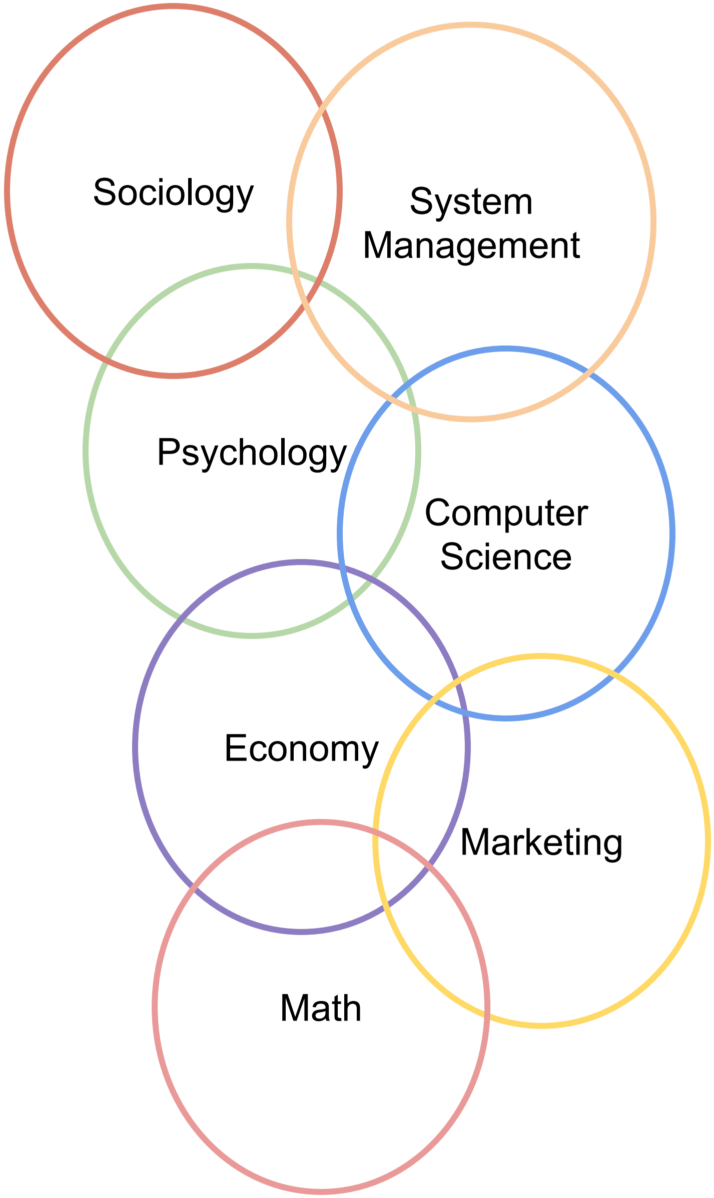
The field of polarization is rather not unified in

- how polarization is defined?

  and

- what is done after recognizing it?

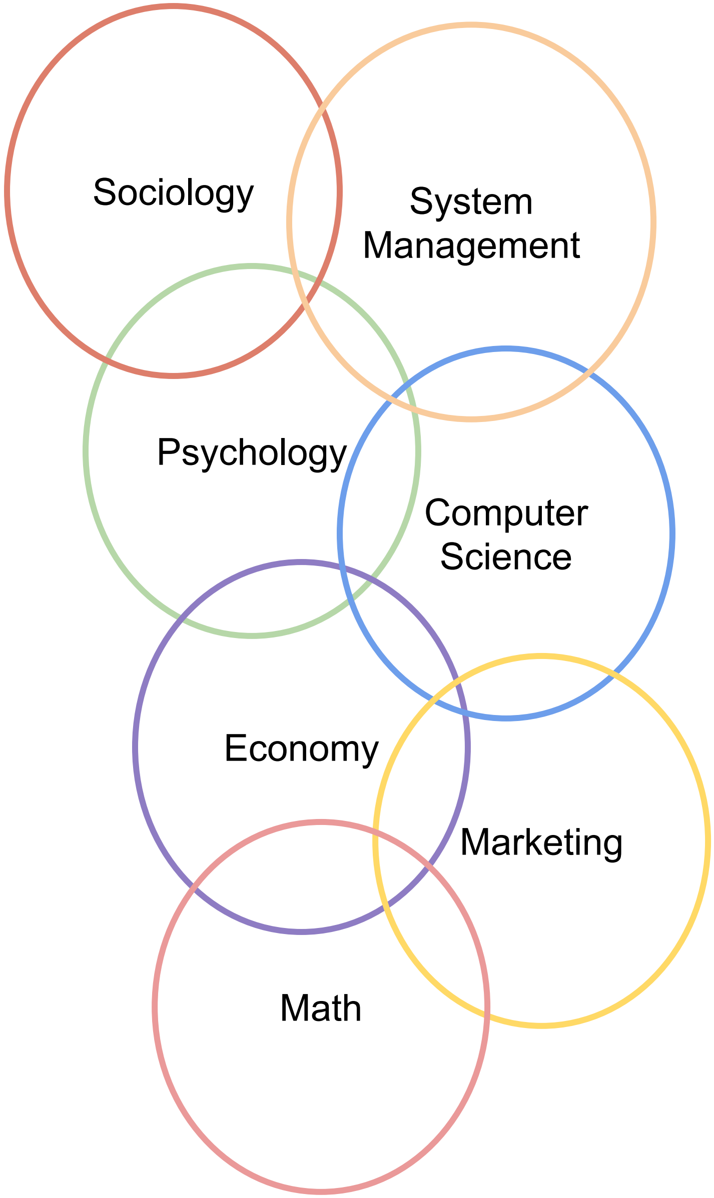
  almost nothing...



Basic Polarization Taxonomy

1. **Social Polarization:** how people **congregate** with one another,
2. **Written Polarization:** how people **write** about topics,
3. **Rated and Recommended Polarization:** how people **behave**, **consume** and express their **preferences,**

   How they <u>interact</u> with <u>algorithms</u>.

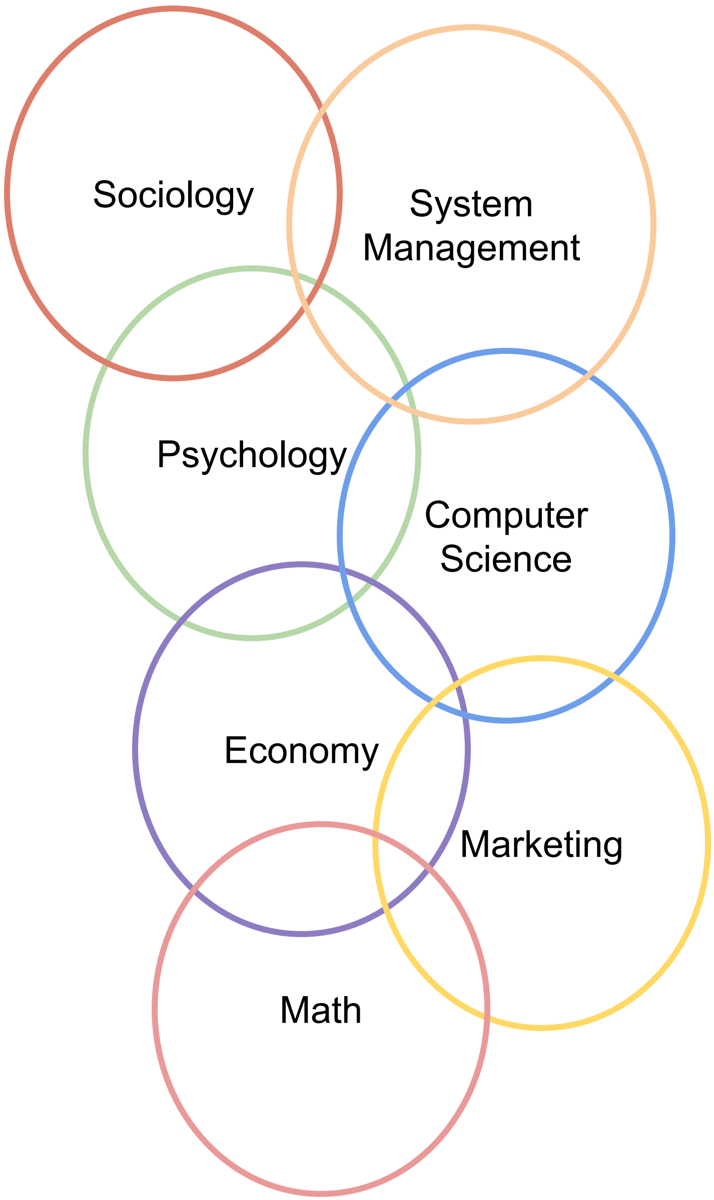


Basic Polarization Taxonomy

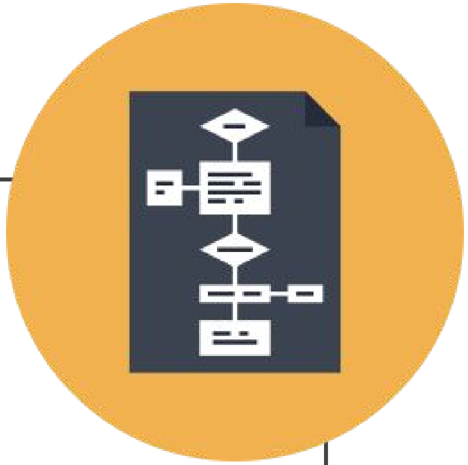
1. **Social Polarization:** how people congregate with one another,
2. **Written Polarization:** how people write about topics,
3. **<u>Rated</u> and <u>Recommended</u> Polarization:** how people behave, consume and express their preferences:

   How they <u>interact</u> with <u>algorithms</u>

What can we do about it?

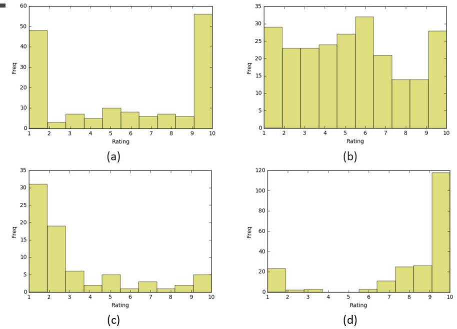


Polarization Detection Classifier - PDT

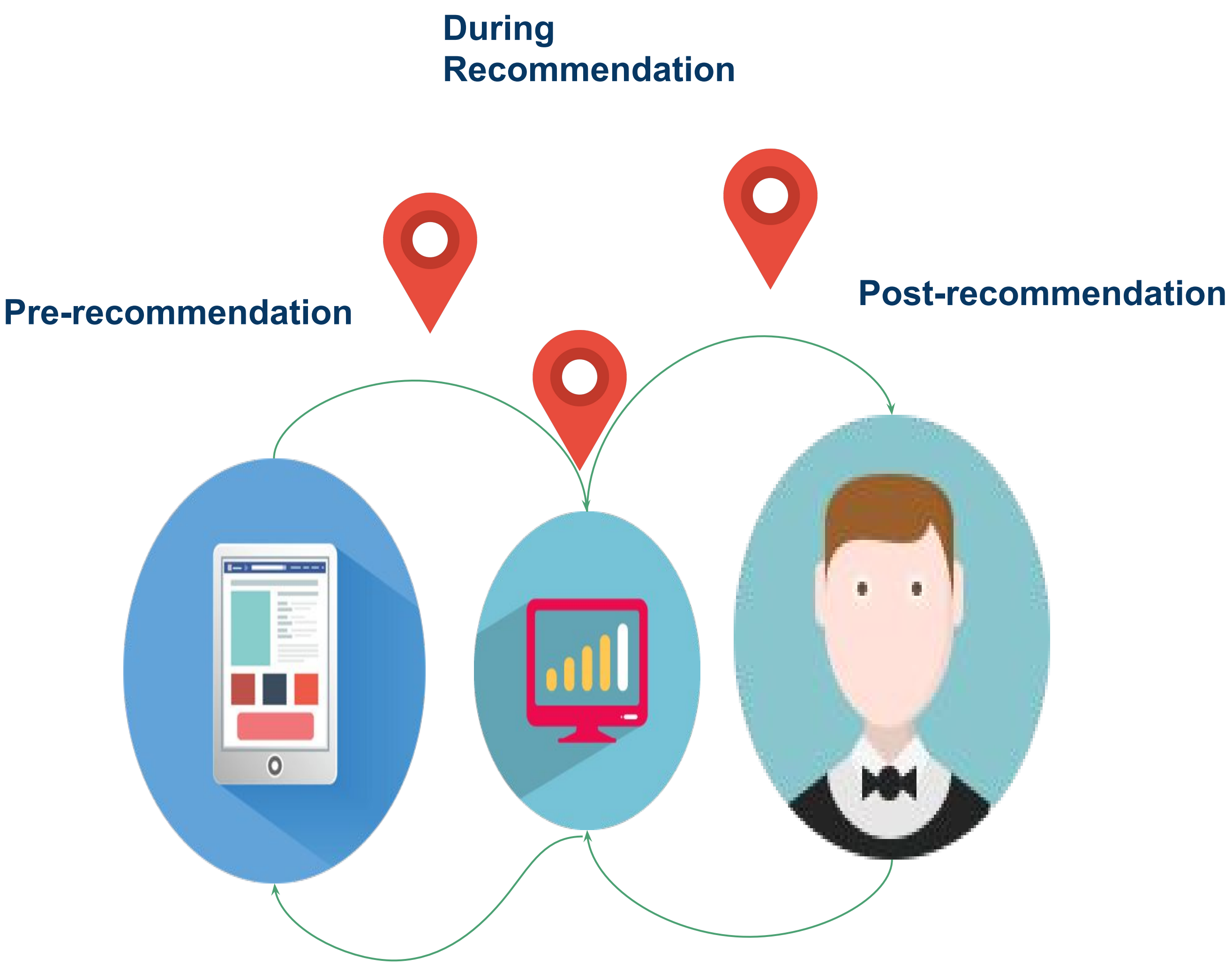


Data Science Pipeline:

- **Data-driven problem formulation**
- Feature engineering
- Modeling
    - Training a classifier using rating data
    - **Polarization Score** = predicted probability of belonging to the **polarized** class
- Evaluation
- Interpretation



Recommender System Counter Polarization Methods: RS-CP



Pre-recommendation Countering Polarization - PrCP

Why do we need it?

- Changing the Recommender System algorithm may not be always feasible
  - Black box
  - or too complex to modify ...

What do we do?

- **Transform the source data** to mitigate extreme ratings that make an item polarized.
- Take into account the **user's relative preferences**,
  - yet **reduce extreme recommendation** that can be generated from a standard recommender system algorithm.

Pre-recommendation -based Countering Polarization - PrCP

Mapping Function:

$$f : (U, I, R) \rightarrow (U, I, R') \text{ with probability of } p$$

User Discovery Factor

Polarization Score

User Preference Threshold

$$r'_{ij} = r_{ij} - \lambda_i \times (\bar{r} + g_i) \times \Phi_j^{\lambda_i + r_{ij}} \quad \text{if } r_{ij} \text{ is } \geq \delta$$

$$r'_{ij} = r_{ij} + \lambda_i \times (\bar{r} - g_i) \times \Phi_j^{\lambda_i + r_{ij}} \quad \text{if } r_{ij} \text{ is } < \delta$$

Item gap ratio

Initial rating

Average ratings of the user

Polarization-aware Recommender Interactive System - PaRIS

Goal:

Design a recommendation system which not only recommends **relevant items**

but also may include **opposite views**

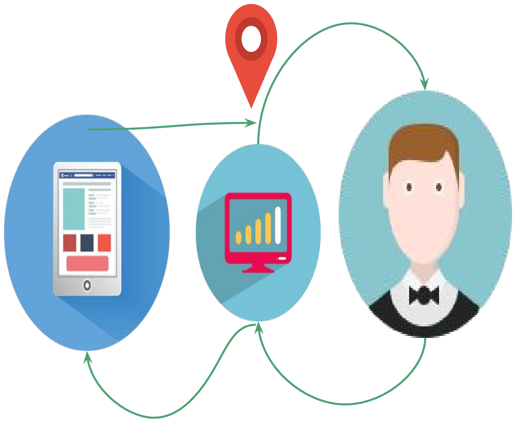
<u>in case the user is **interested** to **discover new items**</u>

Polarization-aware Recommender Interactive System - (PaRIS)

Goal: Design a recommendation system which not only recommends **relevant items** but also includes **opposite views** in case the user is **interested** to **discover new items**.

Our Baseline: Non-negative Matrix Factorization (NMF)-based recommender systems:

- Good scalability
- High predictive accuracy
- Flexibility for modeling various real-life situations
- Easy incorporation of additional information



NMF: Matrix Factorization (Koren et al - 2009)

Input: Rating matrix

	item v	
user u	r_{uv}	

Rating from user u to item v

Idea: Learn p and q to predict all values of the rating matrix

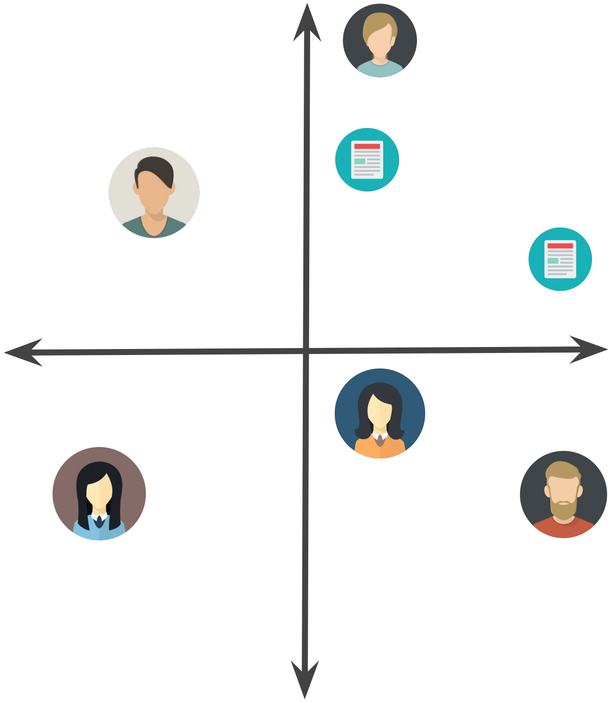
- *p* and *q* are the representation of the user *u* and item *v* in a latent space.

$$r_{uv} = q_v^T * p_u$$

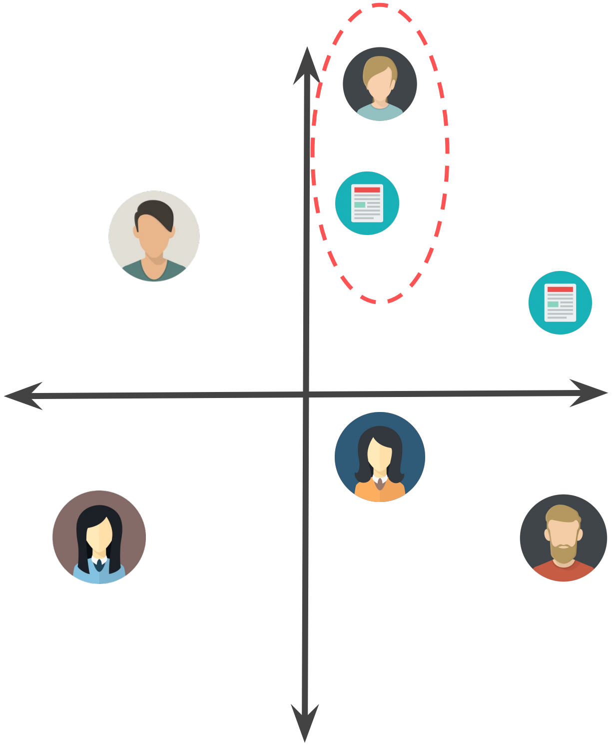
Learning process:

$$\min_{P,Q} = \sum_{(u,v)\in R} (r_{uv} - q_v^T p_u)^2 + \lambda(||(q_v^2|| + ||(p_u^2||)$$

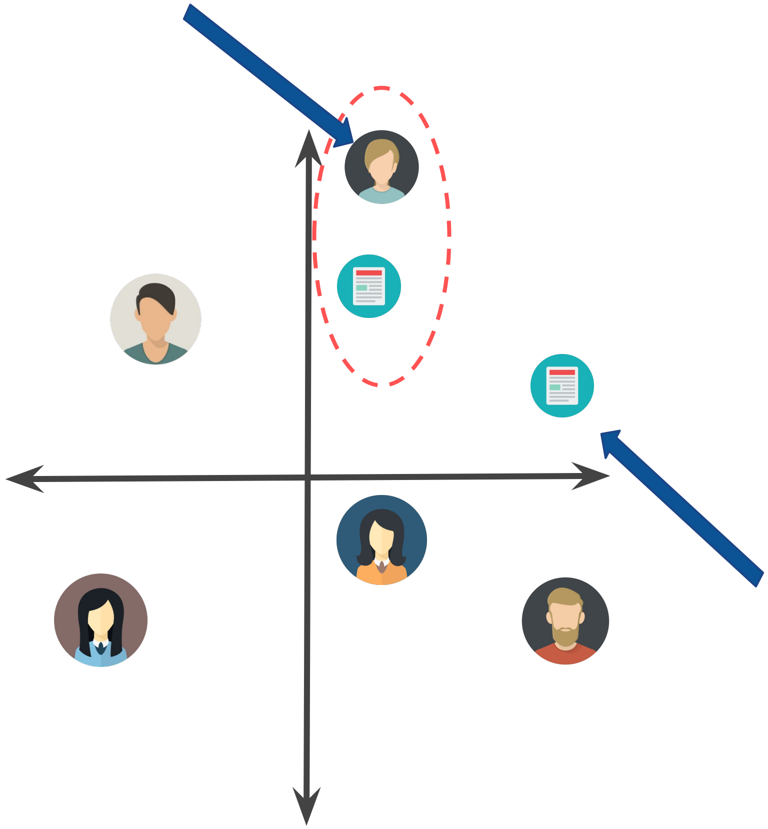
PaRIS - Intuition



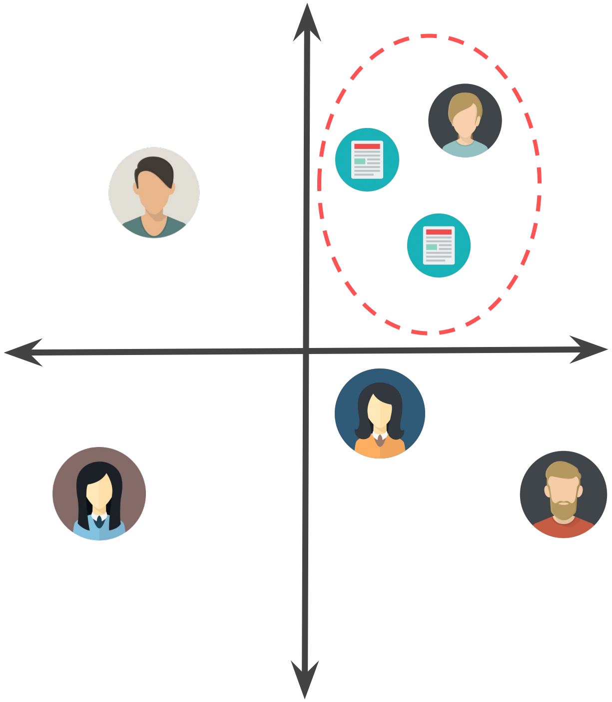
PaRIS - Intuition



PaRIS - Intuition



PaRIS - Intuition



Polarization-aware Recommender Interactive System - PaRIS

$$\min\,(1-\lambda_i)\times||r_{ij}-p_iq_j||^2+\lambda_i\times||r'_{ij}-p_iq_j||^2$$

$$r'_{ij}=r_{ij}-(\bar{r}+g_i)\times\Phi_j^{\lambda_i+r_{ij}}\quad \text{if } r_{ij} \text{ is } \geq\delta$$

$$r'_{ij}=r_{ij}+(\bar{r}-g_i)\times\Phi_j^{\lambda_i+r_{ij}}\quad \text{if } r_{ij} \text{ is } <\delta$$

User Discovery Factor

User Preference Threshold

Initial rating

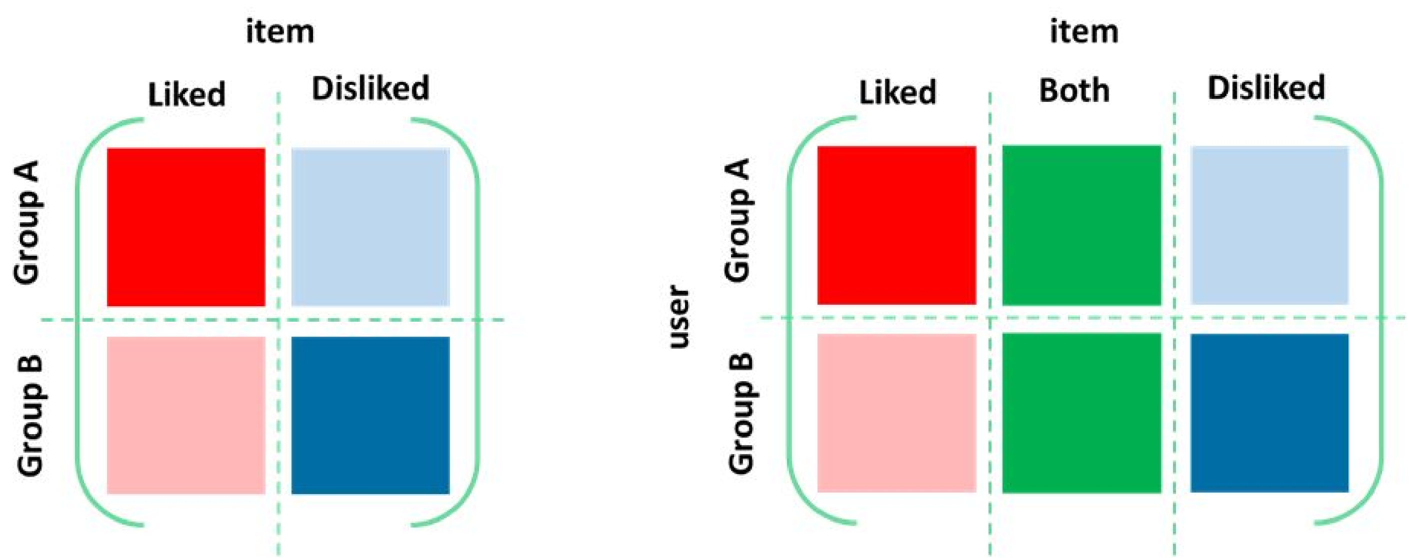
Average rating

Polarization Score

Item gap ratio

Experiments

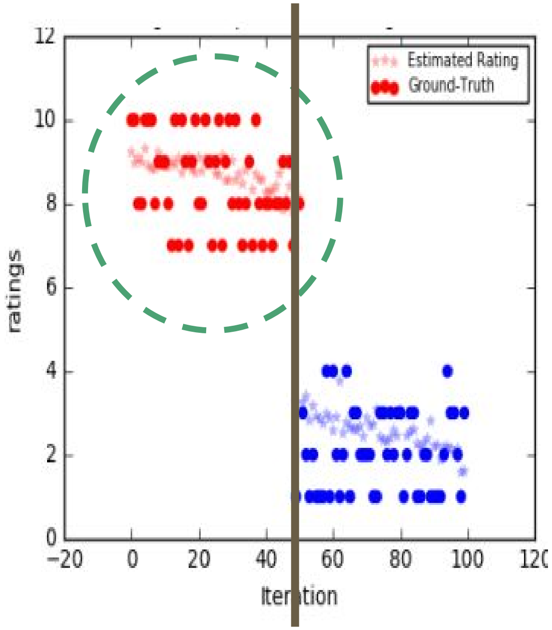
Definition 3 : *Let the number of users, $|U| = n$ and number of items, $|I| = m$. A recommender system algorithm takes environment G as input along with a user $u \in U$, and outputs a set of items $i_1, ..., i_{k_t} \in I$.*

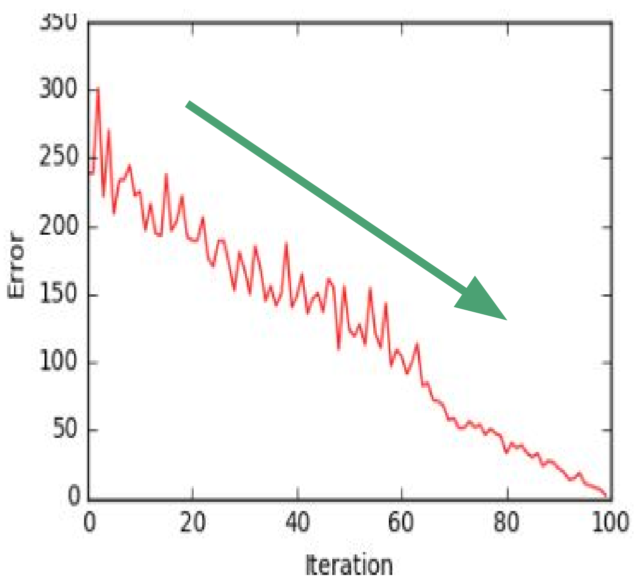


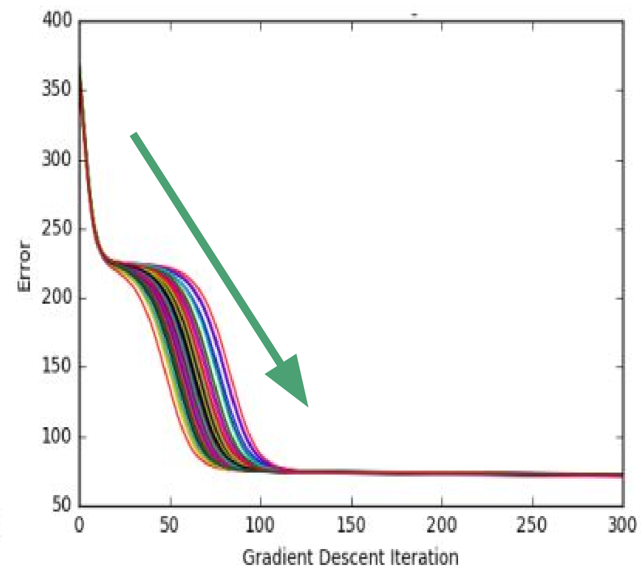
Fully Polarized Environment

Partially Polarized Environment

NMF: Fully Polarized Environment



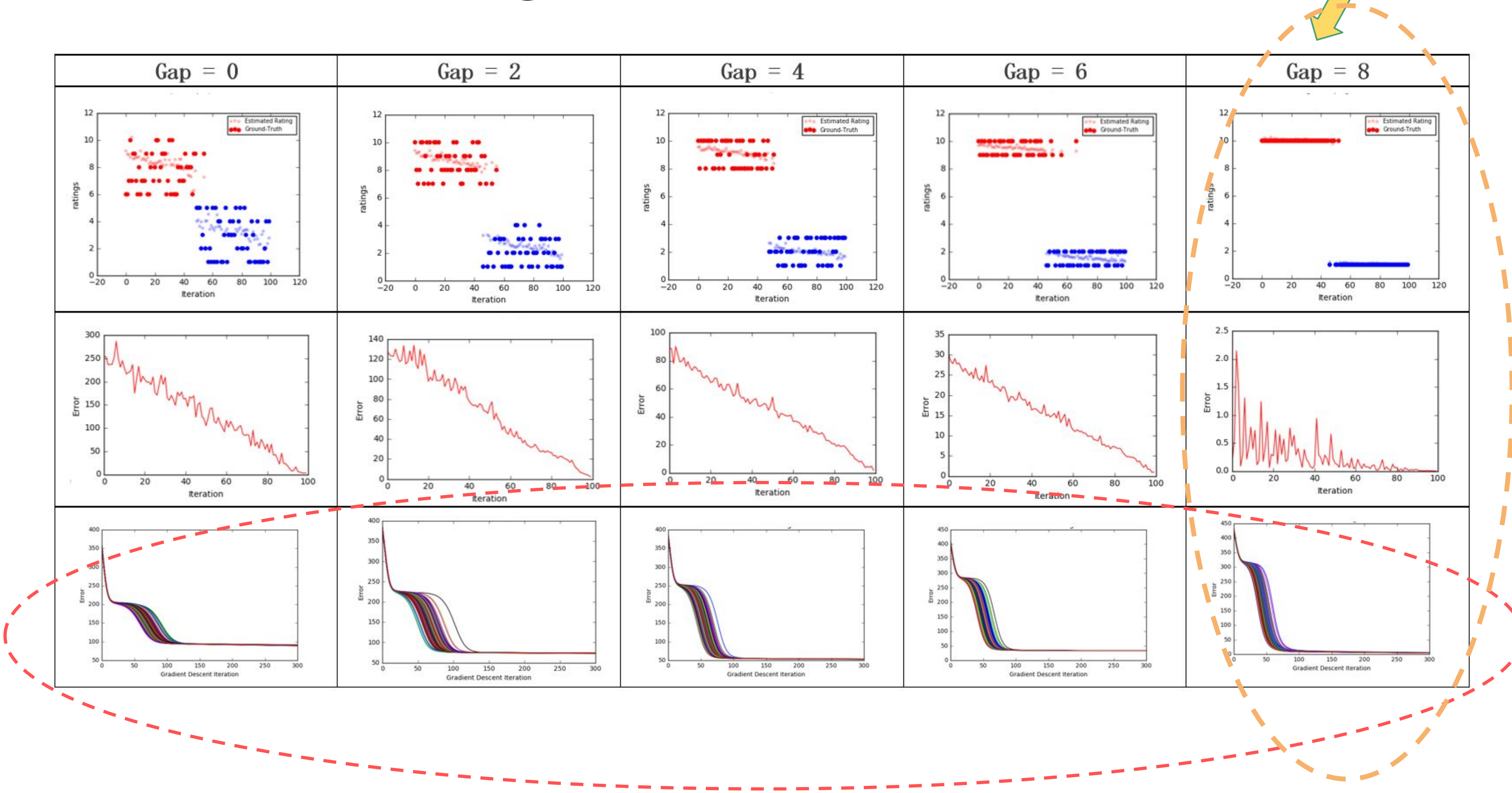




- It is **<u>easy</u>** and **<u>fast</u>** to learn discriminating models in a polarized environment!
  - The result: Keep each user in the safety of their preferred viewpoint

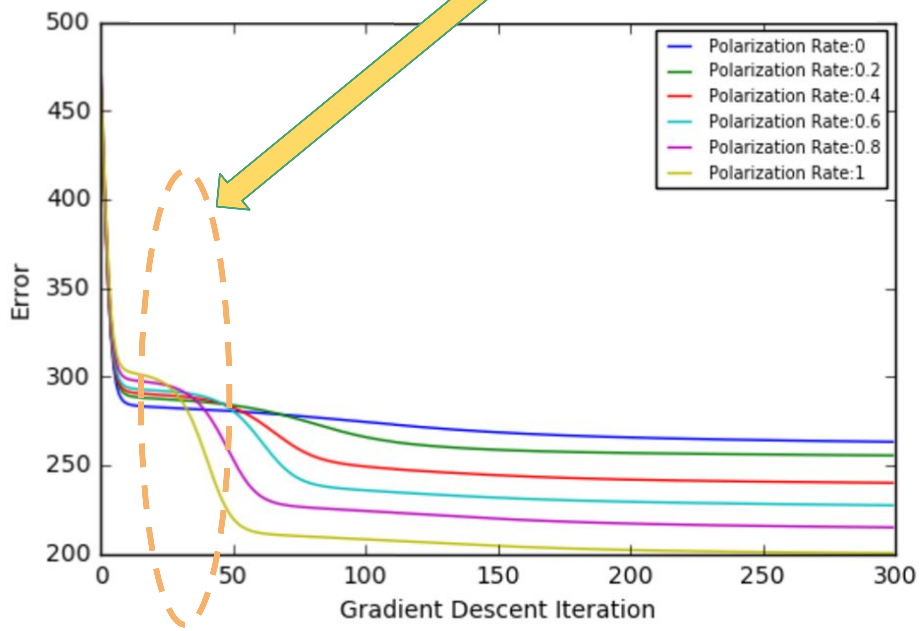
Effect of **Increasing Polarization** on NMF

Extreme Polarization!!



Effect of Polarization on NMF

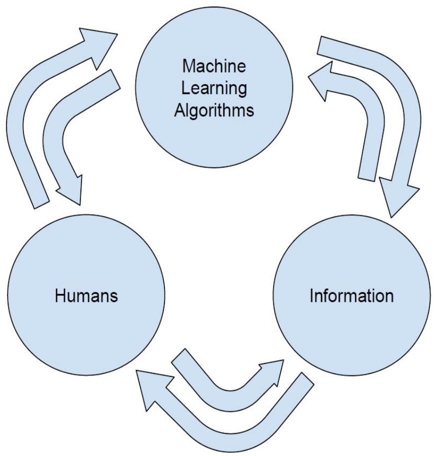
Can <u>monitor</u> convergence trend to <u>detect</u> emergence of polarization!!



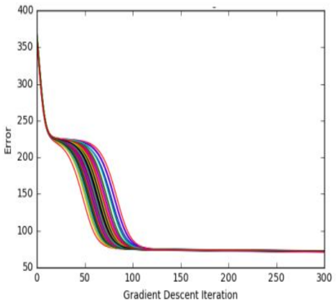
Counter Polarization Methods:
Recommend **More** Items from Opposite View

		Opposite View Ratio		Mean Square	
		$OVHR_u$	$OVHR_{tk}$	MSE_{Train}	MSE_{Test}
		mean, std	mean, std	mean, std	mean, std
Classic NMF		$0.0\% \pm 0.00$	$0.0\% \pm 0.00$	22.02 ± 5.27	138.96 ± 12.55
PrCP	$\lambda_i = \mathbf{0.2}$	$5.4\% \pm 0.073$	12.32 ± 0.31	123.92 ± 36.76	813.01 ± 36.76
	$\lambda_i = \mathbf{0.5}$	$6.0\% \pm 0.08$	$18.1\% \pm 0.21$	124.46 ± 37.29	299.82 ± 76.01
	$\lambda_i = \mathbf{0.7}$	$61.0\% \pm 0.17$	$31.0\% \pm 0.167$	209.73 ± 59.53	967.103 ± 145.92
	$\lambda_i = \mathbf{1.0}$	$67.0\% \pm 0.24$	$68.0\% \pm 0.24$	361.77 ± 102.74	1883.50 ± 237.83
PaRS	$\lambda_i = \mathbf{0.2}$	$5.4\% \pm 0.73$	$4.9\% \pm 0.021$	123.92 ± 36.76	813.01 ± 36.76
	$\lambda_i = \mathbf{0.5}$	$6.2\% \pm 0.075$	$5.2\% \pm 0.042$	122.56 ± 39.081	804.01 ± 75.88
	$\lambda_i = \mathbf{0.7}$	$7.0\% \pm 0.075$	$5.4\% \pm 0.033$	120.97 ± 35.19	803.65 ± 64.65
	$\lambda_i = \mathbf{1.0}$	$6.8\% \pm 0.064$	$5.8\% \pm 0.03$	119.76 ± 34.93	801.86 ± 65.07

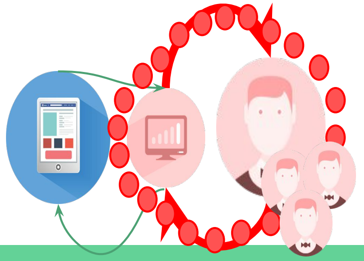
Conclusion



- ★ **Iterated Learning Bias**: theory and simulations
- ★ **Counter-polarization**
  - ○ Empower the users who are increasingly entrapped in algorithmic filters
  - ○ Allows **humans to regain control** of algorithm-induced filter bubble traps,
  - ○ Impact on **information filtering / recommender systems**
    - ■ News, social media, e-commerce, e-learning, etc



- ★ We uncovered **patterns** that are characteristic of environments where **polarization** emerges
  - ○ Can monitor objective function optimization trend
  - ○ ⇒ **detect** and **quantify** the evolution of **polarization**
- ★ ⇒ **allow users to <u>break free from their algorithmic chains</u>**!



Outline

- What can go Wrong in Machine Learning?
  - Unfair Machine Learning
  - Iterated Bias & Polarization
  - Black Box models
- Tell me <u>more</u>: Counter-Polarization
- **Tell me <u>why</u>: Explanation Generation**

Why is Explainability So Important?

Transparency is crucial to scrutinize:

- incorrect predictions
- biased predictions

More trustworthy ML models!

Black Box vs. White Box

- Black Box (opaque) predictors such as Deep learning and matrix factorization are accurate,
  - *…. but lack interpretability and ability to give explanations*

- White Box models such as rules and decision trees are interpretable (explainable)
  - *… but lack accuracy*

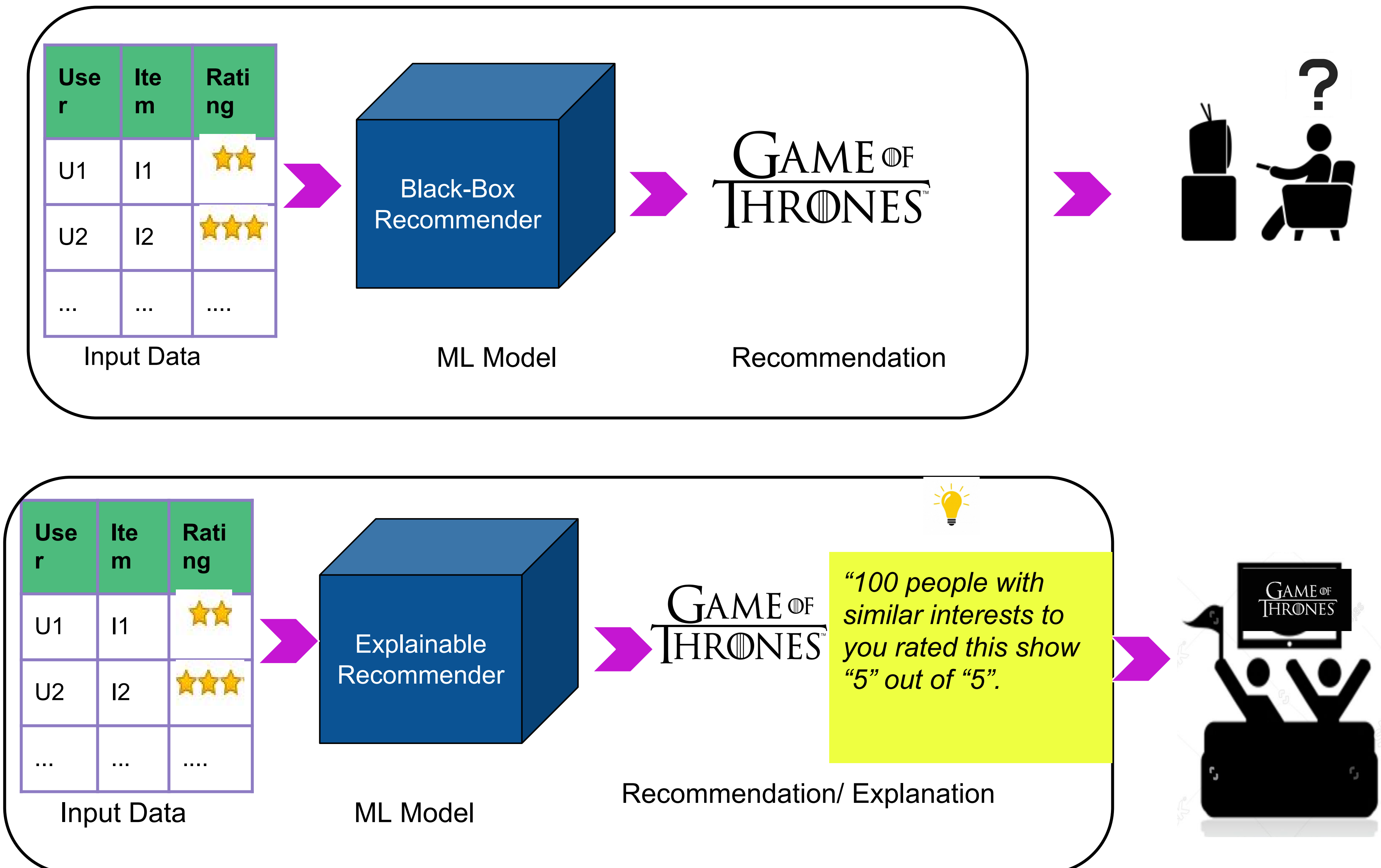
- Explanations provide a rationale behind predictions
  ➔ help the user gauge the validity of a prediction
  ➔ may reveal prediction errors and reasons behind errors
  ➔ increase trust between human and machine

Our Focus: Explanations in Recommender Systems

Recommender Systems

Collaborative Filtering

Uses previous ratings of the user to predict future preferences



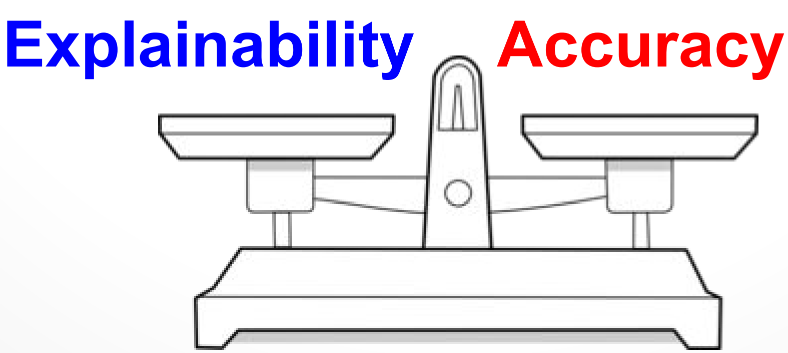
User
Item
Rating
U1
I1
U2
I2
...
...
....
Input Data
Black-Box Recommender
ML Model
GAME OF THRONES
Recommendation
?
User
Item
Rating
U1
I1
U2
I2
...
...
....
Input Data
Explainable Recommender
ML Model
GAME OF THRONES
"100 people with similar interests to you rated this show "5" out of "5".
Recommendation/ Explanation
GAME OF THRONES

Tradeoff between Accuracy and Explainability

- Using Explanations, we can increase the transparency of the model.

- However there may be a downside:
  - Explainable models should also remain accurate!

Goal : a moderate tradeoff between accuracy and explainability



MF: Matrix Factorization (Koren et al - 2009)

Input Data: Rating matrix

	item v	
user u	r_{uv}	

Rating from user u to item v

Idea: Learn *p* and *q* to predict all missing values of the rating matrix

p and *q* = representation of user *u* and item *v* in a latent space.

$$r_{uv} = q_v^T * p_u$$

Learning process:

$$\min_{P,Q} = \sum_{(u,v)\in R} (r_{uv} - q_v^T p_u)^2 + \lambda(\|(q_v^2\| + \|(p_u^2\|)$$

Main Problem: Matrix Factorization is a **Black Box Model**

EMF: Explainable Matrix Factorization (Abdollahi & Nasraoui, 2016)

Idea: Provide neighborhood style **Explanations** along with recommendations
and **learn** a model that is **explainable**

Recommendation:

Justification:

80% of users who share similar interests with you liked this movie

New objective function:

$$J = \sum_{(u,v)\in R} (r_{uv} - q_v^T p_u)^2 + \frac{\beta}{2}(\|p_u^2\| + \|q_v^2\|) + \boxed{\frac{\lambda}{2}(p_u - q_v)^2 W_{uv}}$$

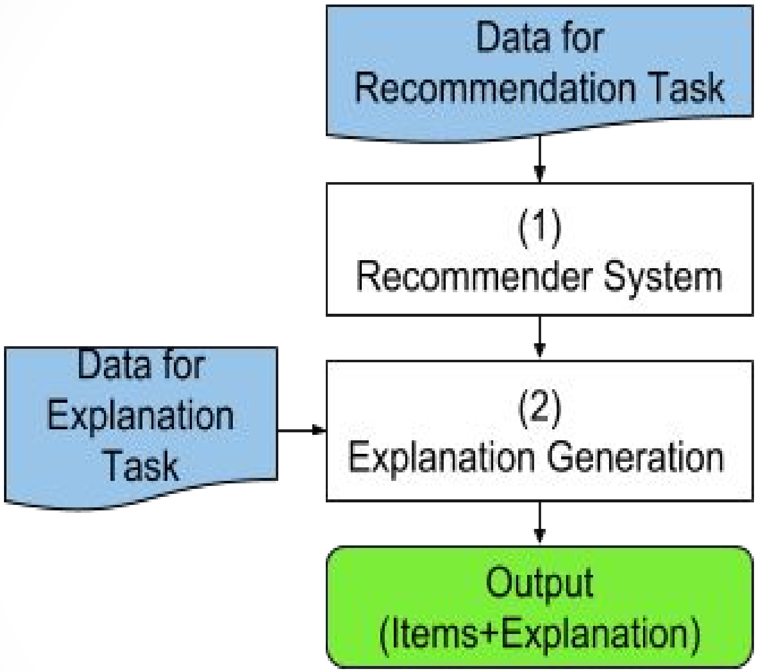
Explainability term to favor users and items with similar p and q

W_{uv} = Explainability score calculated for user u and item v.

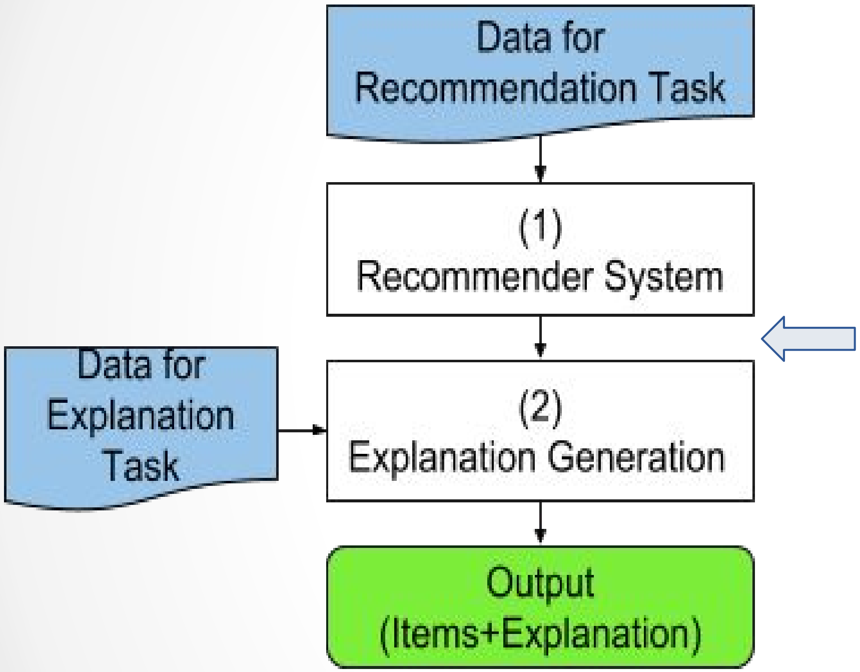
$$W_{uv} \begin{cases} \frac{|N'(u)|}{|N'_k(u)|} \; if \; \frac{|N'(u)|}{|N'_k(u)|} > \theta; \\ 0 \quad\quad Otherwise; \end{cases}$$

- N': total number of neighbors of user u who rated item v
- N'_k: total number of neighbors of user u

Classical Framework



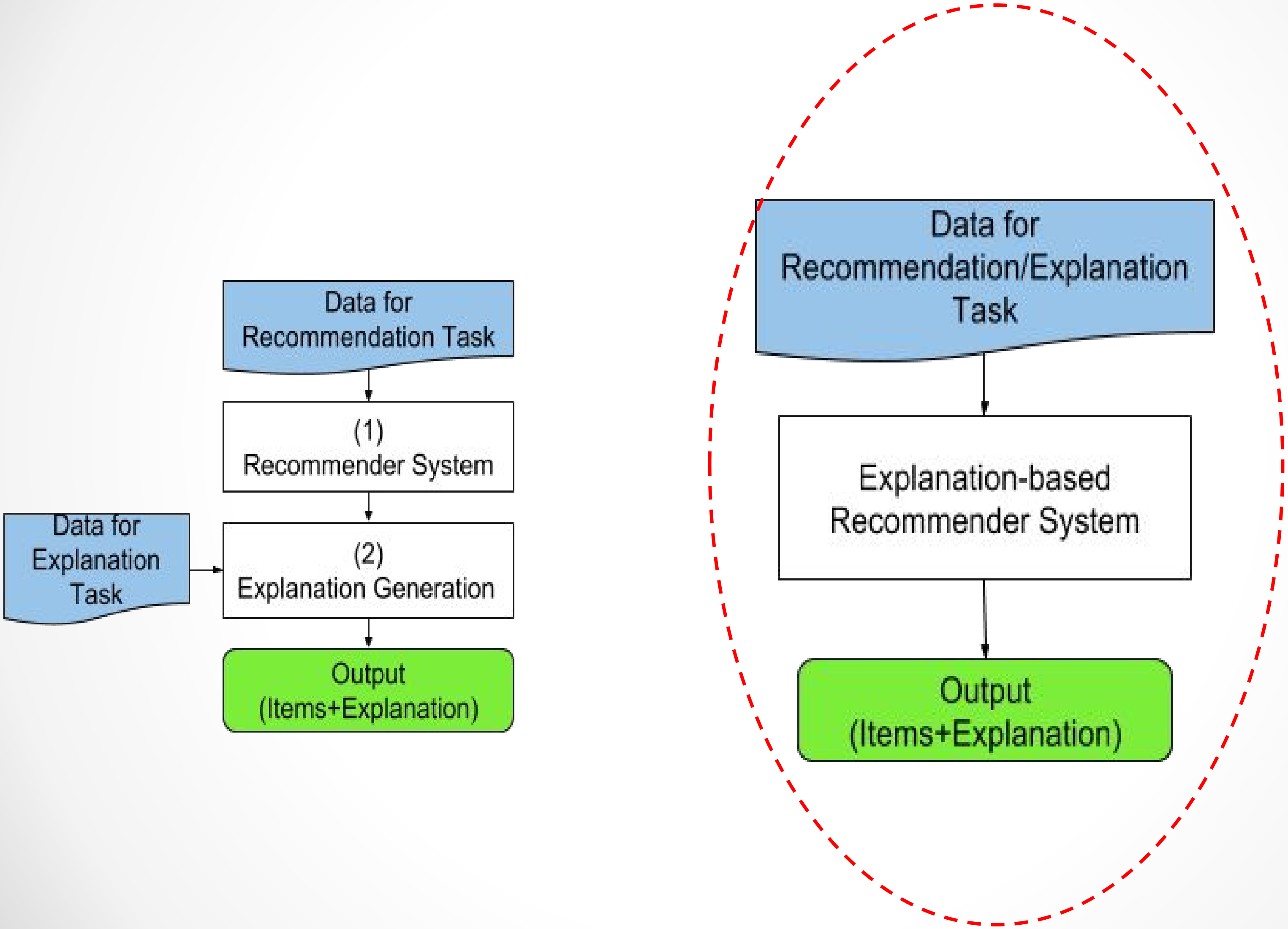
Classical Framework



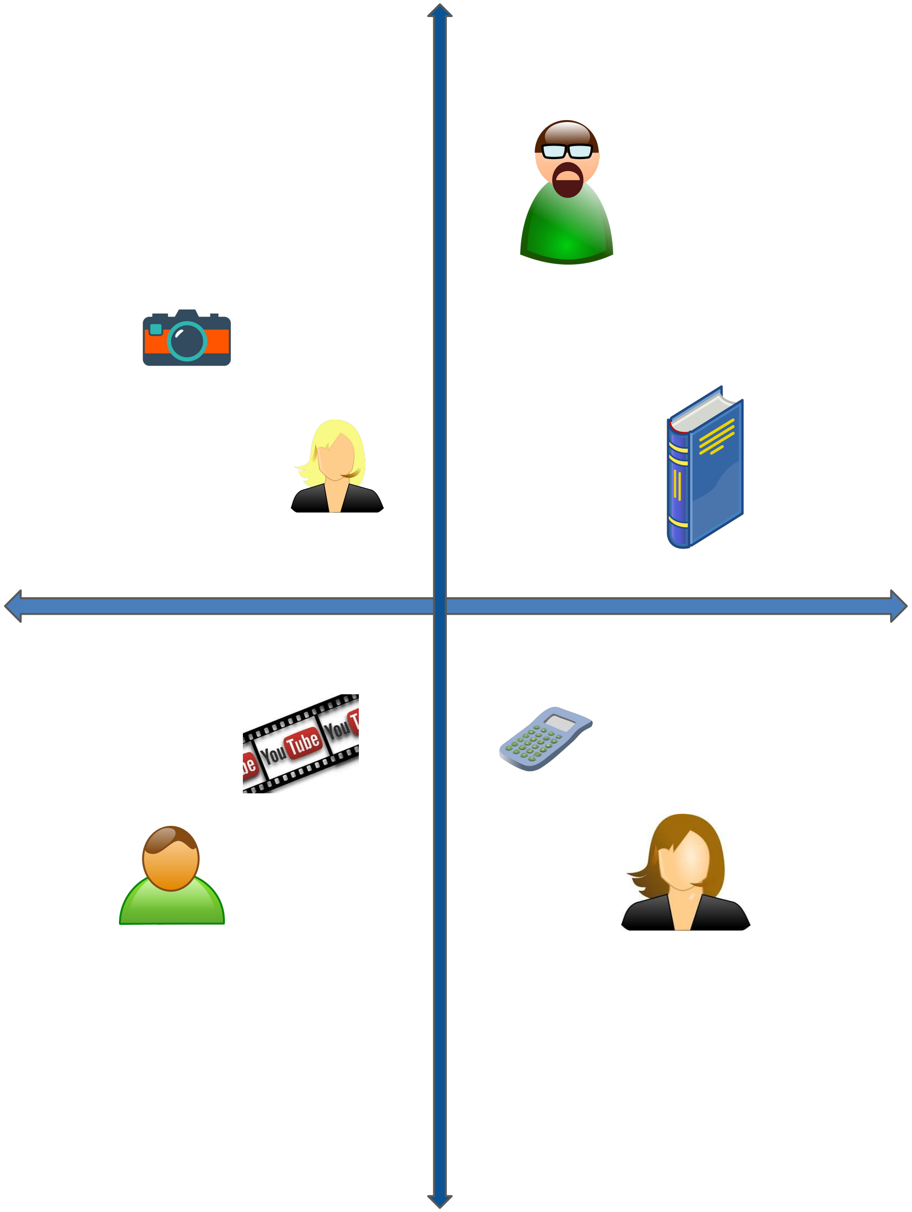
- possible mismatch between (1) and (2)

- generally need to generate explanations at recommendation time (not efficient)

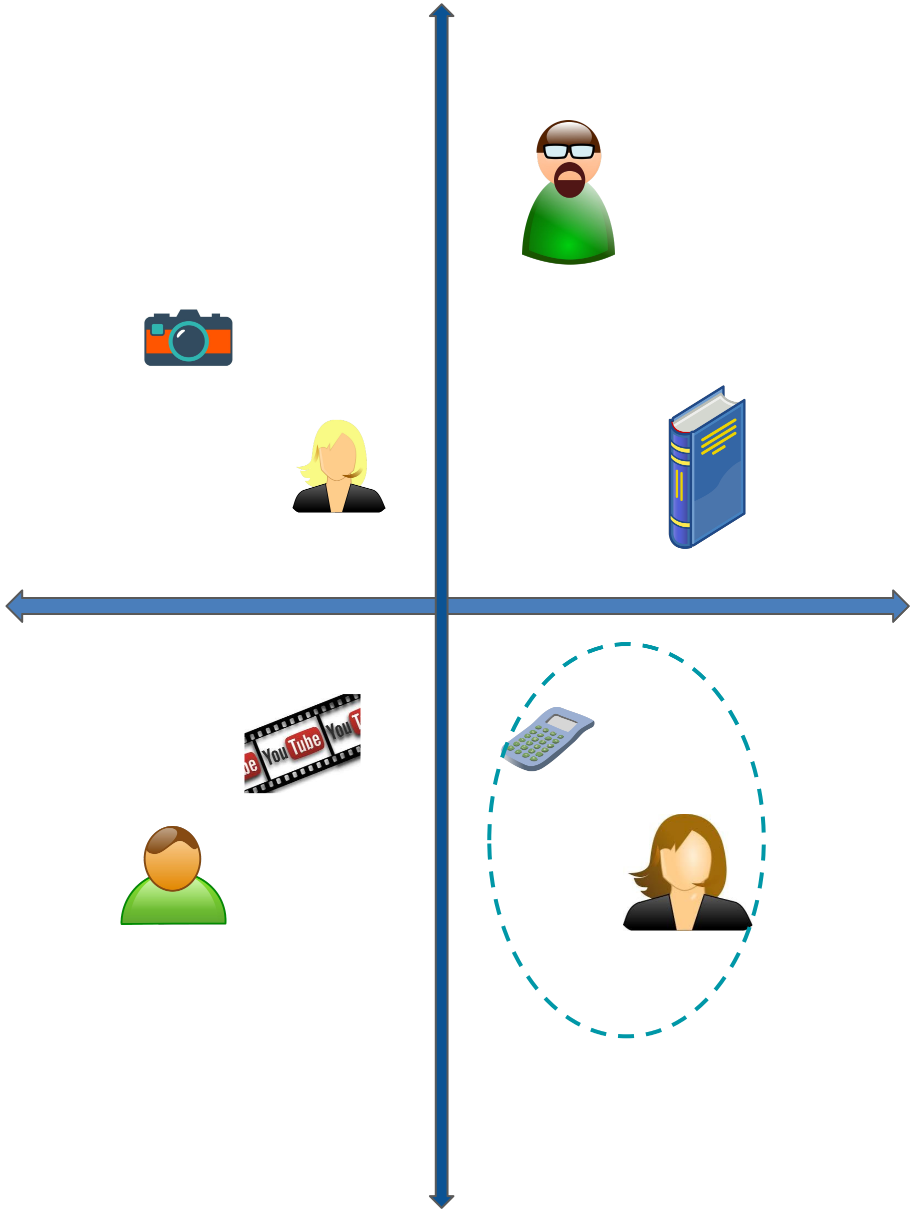
Classical Framework vs Proposed Framework



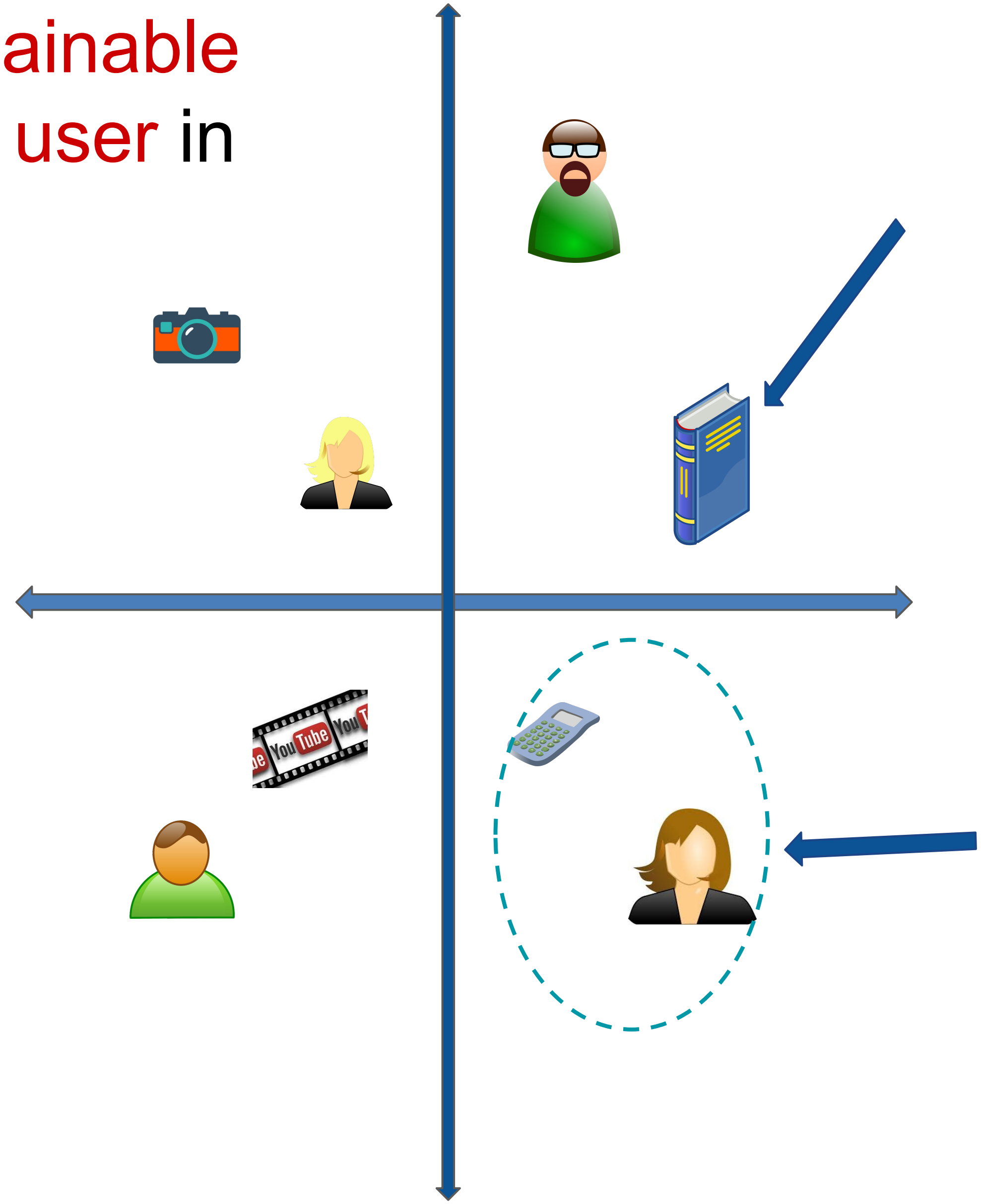
Intuition



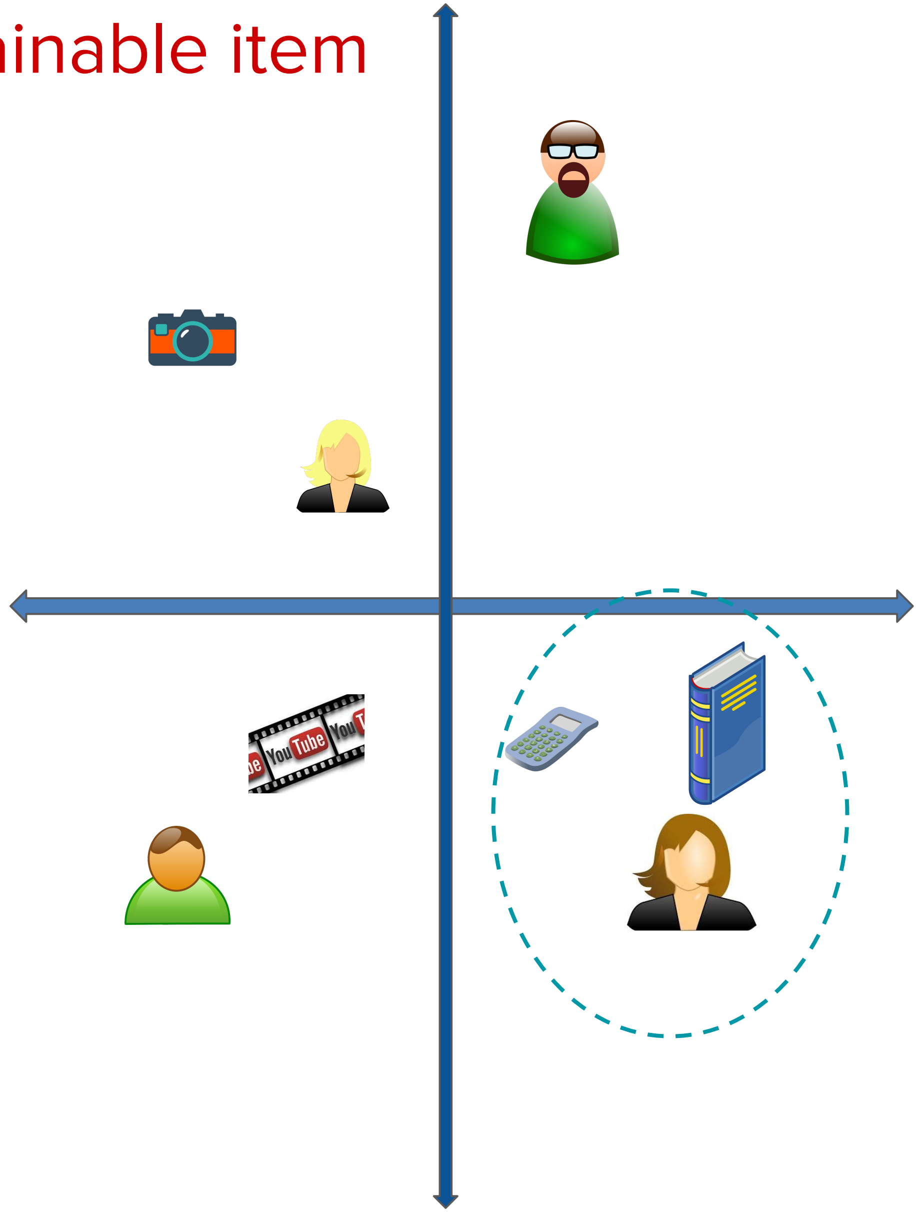
Intuition



Intuition: Bring explainable items **closer** to the user in latent space

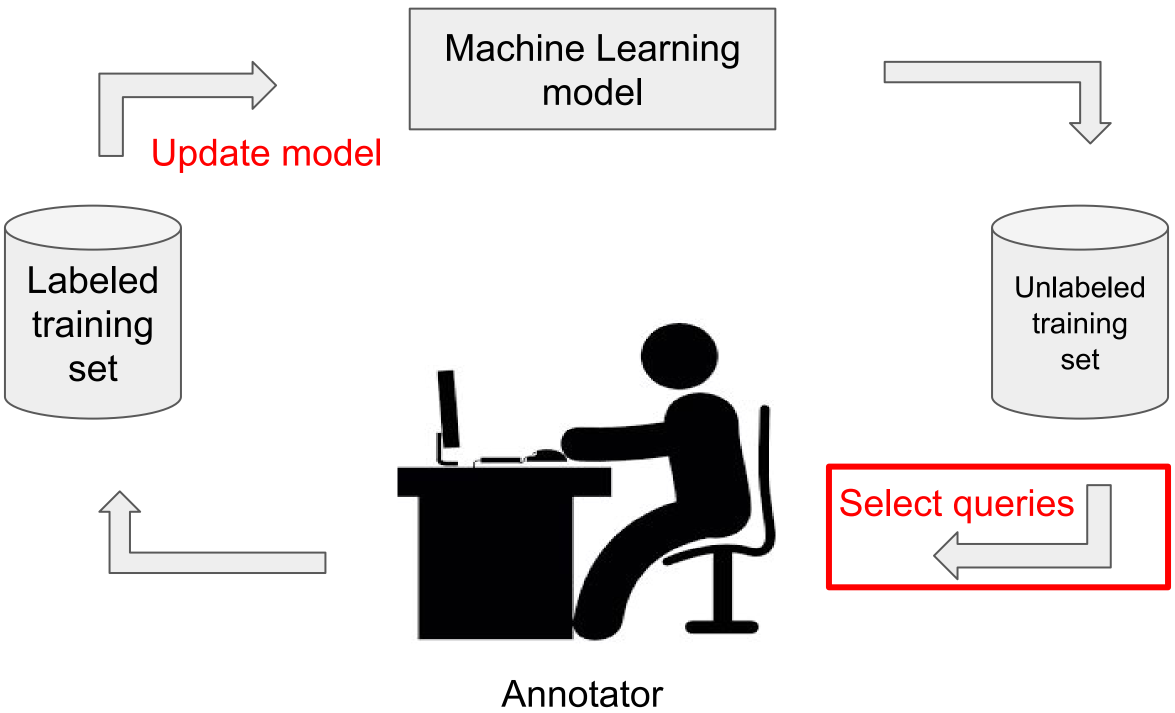


Intuition: Now explainable item is **more likely to be recommended**

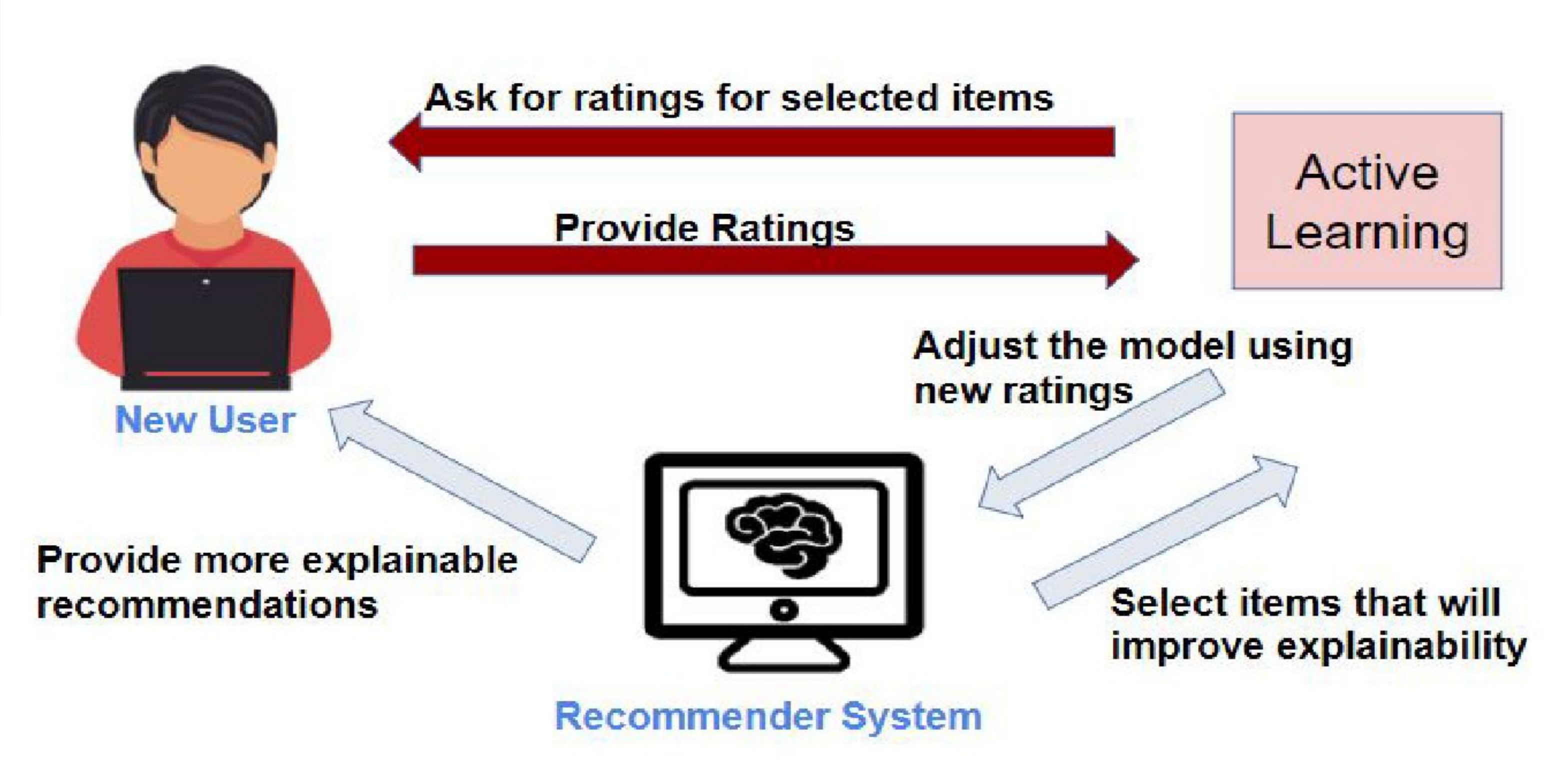


Active Learning

What If we make the algorithm <u>choose</u> the most useful training data?



ExAL: Explainable Active Learning

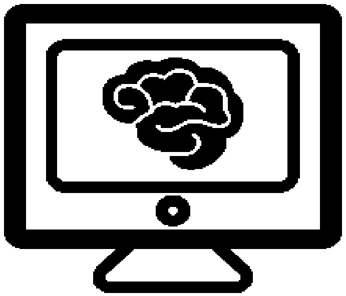


1. Select items from an unlabeled pool of items using an **Active Learning selection strategy**
2. Obtain the true ratings of the selected item from the new user
3. Adjust the parameters of the model using the new ratings
4. Repeat the process until meeting a stopping criterion

Explainable Active learning Strategy Algorithm (ExAL)

New User

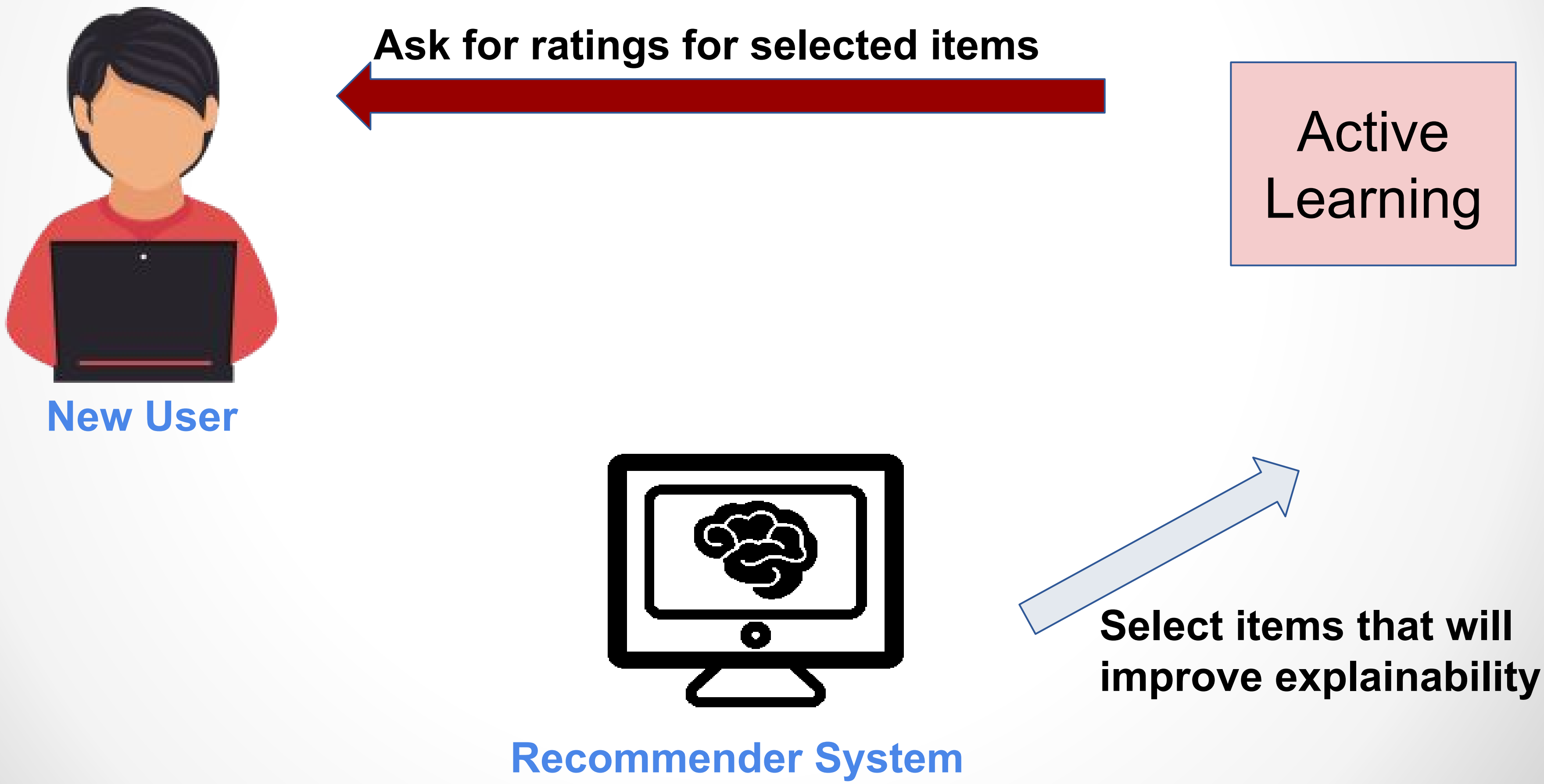
Active Learning



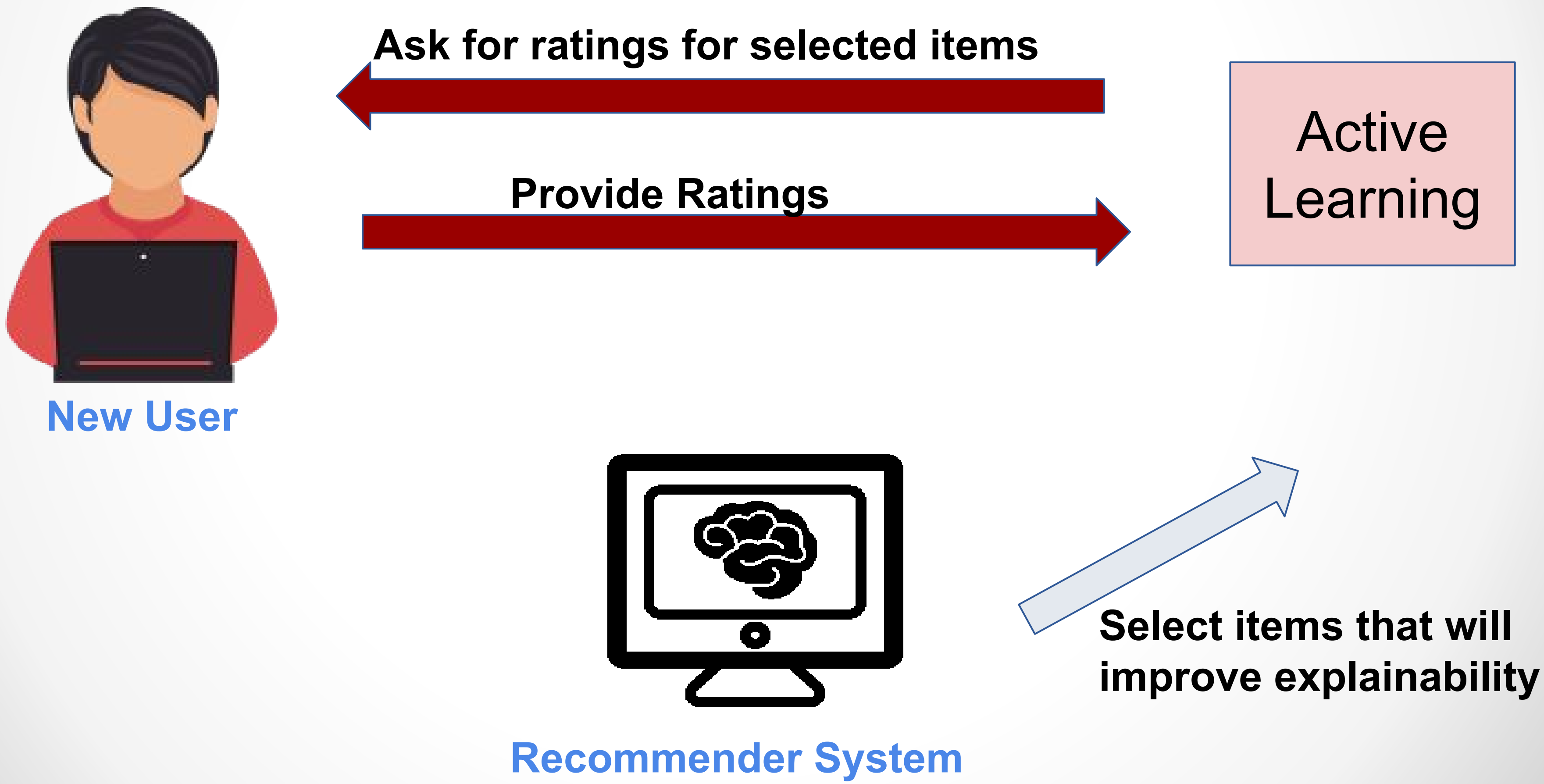
Select items that will improve explainability

Recommender System

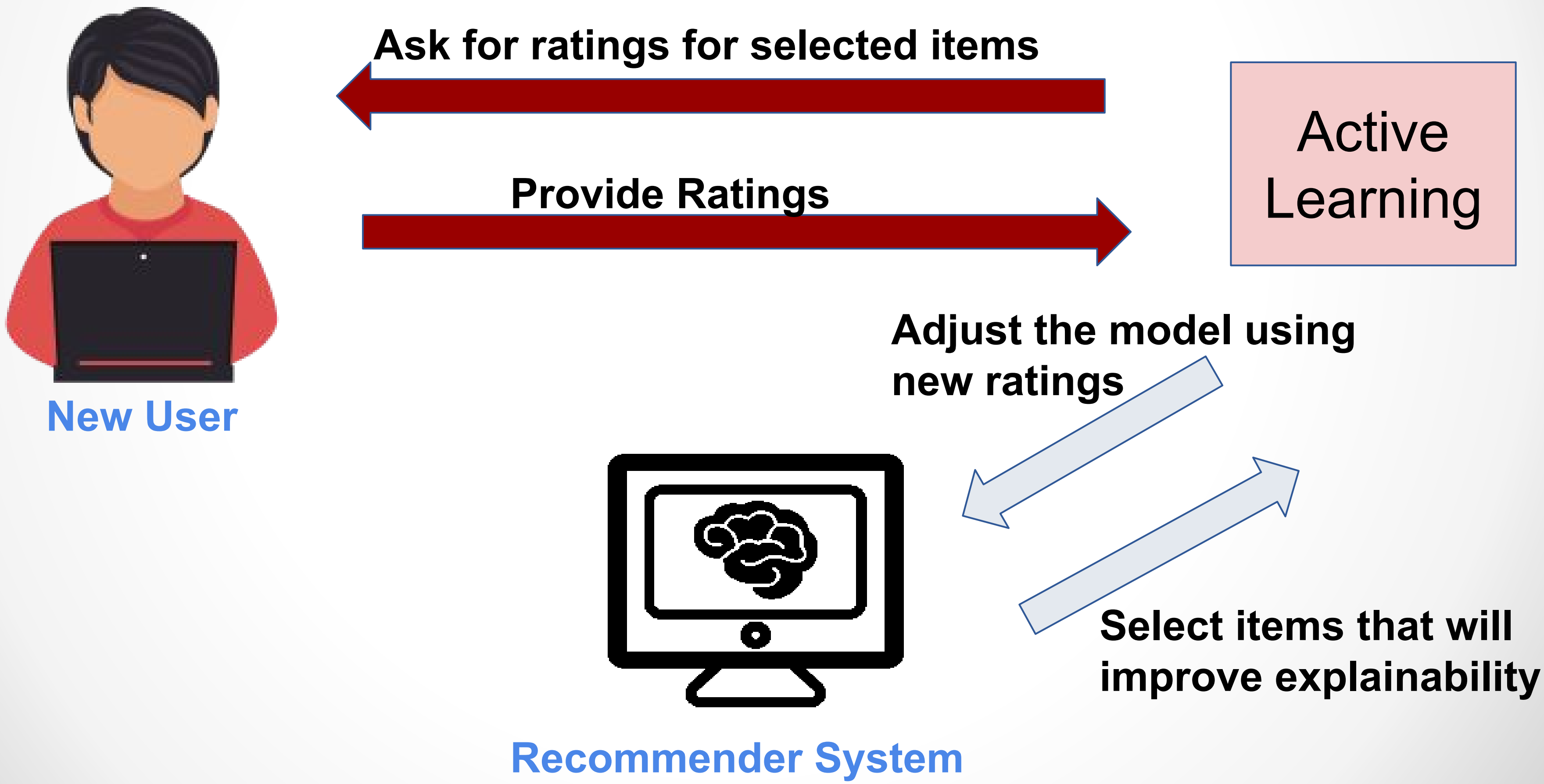
Explainable Active learning Strategy Algorithm (ExAL)



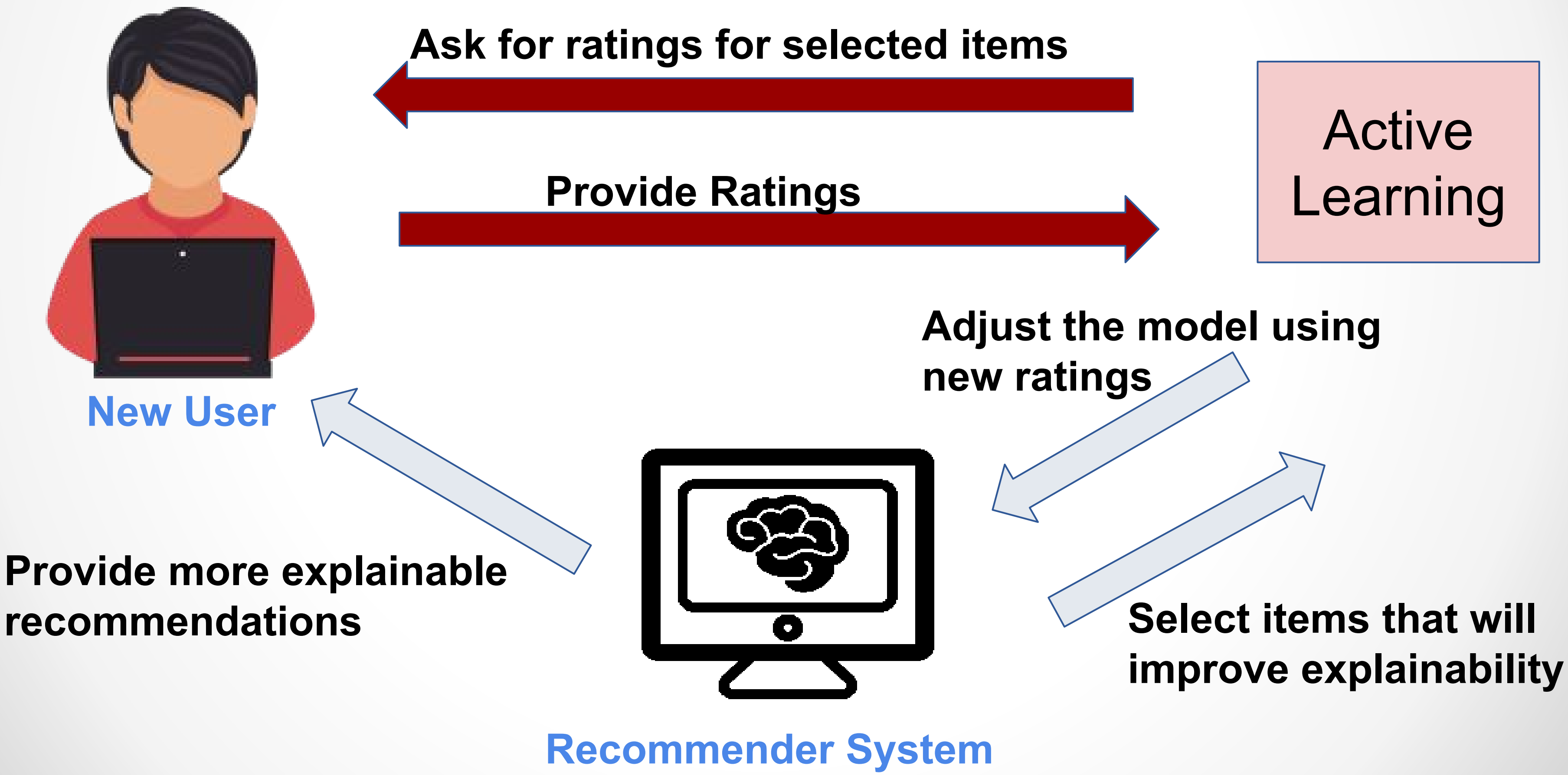
Explainable Active learning Strategy Algorithm (ExAL)



Explainable Active learning Strategy Algorithm (ExAL)



Explainable Active learning Strategy Algorithm (ExAL)



Explainable Active learning Strategy Algorithm (ExAL)

Active Learning to improve explainability in MF

Problem :

How are we going to select the best items to be queried to the user ?

↓

Selection Criterion

Explainable Active learning Strategy Algorithm (ExAL)

Active Learning to improve explainability in MF

Proposition : A selection criterion for EMF to minimize testing error and increase explainability for user u :

i* such that :

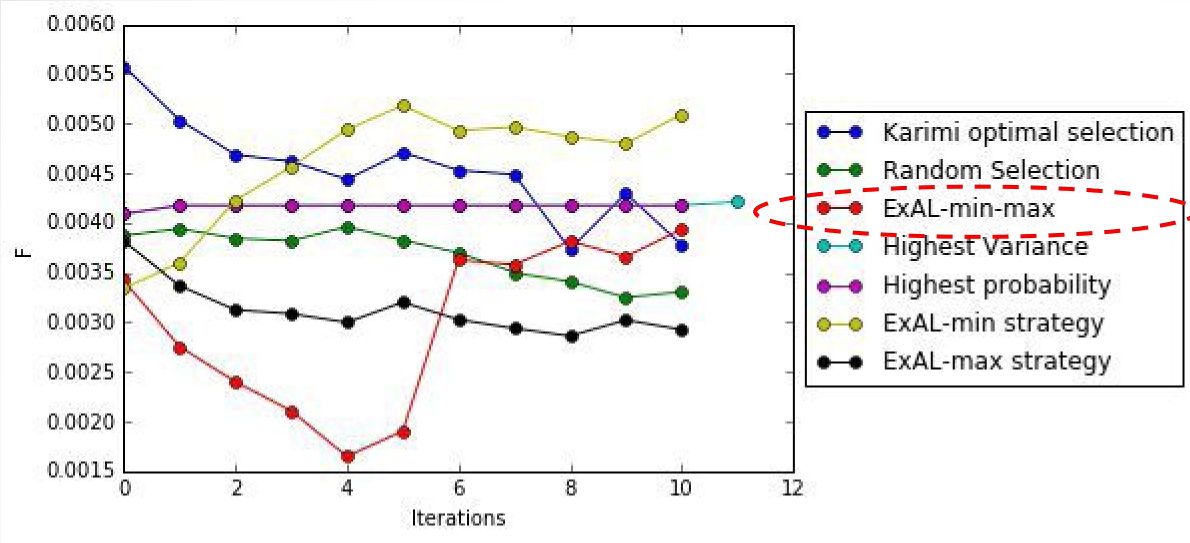
$$i^*_u \simeq \underset{i\in I^u_{pool}}{\operatorname{argmin}} \sum_{j\in I^u_{test}} \left| 1 - r_{uj} + \boxed{2\alpha((r_{ui} - \bar{R}_i)\sum_{f=1}^{k} q_{if}q_{jf}} + \boxed{\lambda W_{ui}(r_{uj} - \sum_{f=1}^{k} q_{if}q_{jf}))} \right|$$

Index of the item that will be queried from the user

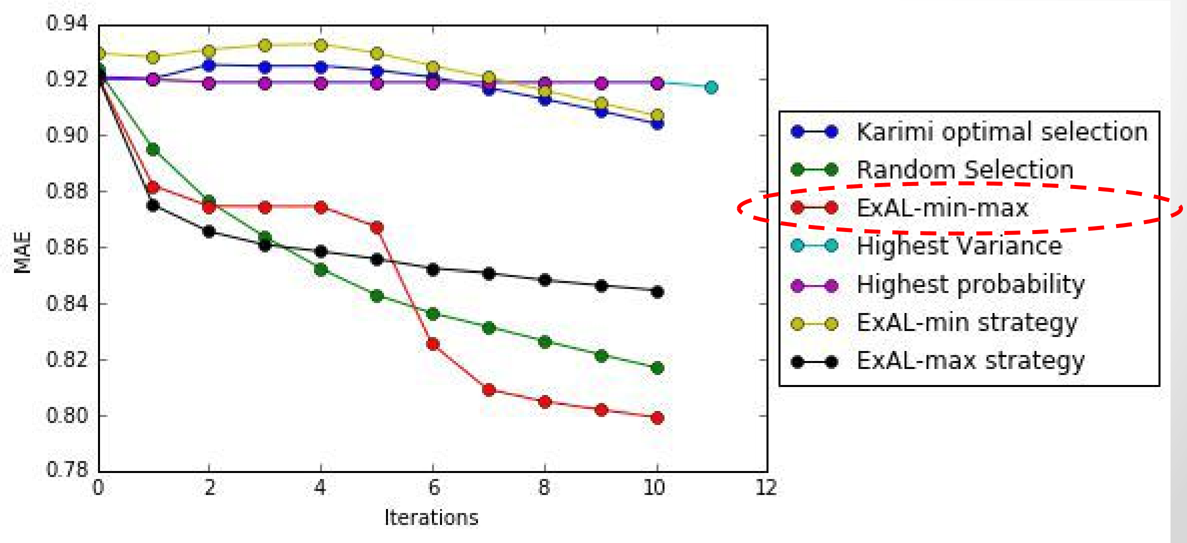
Expected change in the accuracy of the testing error

Explainability term that takes into consideration explainability as a selection criterion

Explainability F-score



Predictive Error (MAE)



Summary of Explainable Recommender Systems

- **EMF**: Explainable Matrix Factorization
  - Explainable Latent Factor Model
- **ERBM**: Explainable Restricted Boltzman Machines for Recommender Systems
  - **Explainable Deep Learning Approach** for Collaborative Filtering
- Both EMF and ERBM:
  - improve explainability
  - without significant loss in accuracy
- **ExAL**: An **Active learning** approach to Explainable Recommendations
  - improves explainability **and** accuracy

References

- Kirby, Simon, Tom Griffiths, and Kenny Smith. (2014) "Iterated learning and the evolution of language." *Current opinion in neurobiology* 28: 108-114.
- Nasraoui, O. & Shafto, P. (2016). Human-algorithm interaction biases in the Big Data cycle: A Markov Chain Iterated Learning framework. arXiv preprint arXiv:1608.07895.
- Shafto, P. & Nasraoui, O. (2016). Human-recommender systems: From benchmark data to benchmark cognitive models. Proceedings of the 10th ACM Conference on Recommender Systems (RecSys 2016), 127-130

- Abdollahi, Behnoush, and Nasraoui, Olfa. (2016). Explainable Matrix Factorization for Collaborative Filtering."In Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee.
- Abdollahi, Behnoush, and Nasraoui, Olfa. (2016). Explainable Restricted Boltzmann Machines for Collaborative Filtering." ICML Workshop in Human Interpretability.
- D. D. Lee and H. S. Seung, (2001). "Algorithms for non-negative matrix factorization," Advances in neural information processing systems, pp. 556–562.
- Abdollahi, Behnoush, and Nasraoui, Olfa. (2017). Explainable "Using Explainability for Constrained Matrix Factorization", Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017).
- Karimi, R., Freudenthaler, C., Nanopoulos, A., and Schmidt-Thieme, L. (2011b). Towards optimal active learning for matrix factorization in recommender systems. In Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence.
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8).

Thank You!

Tell me Why? Tell me More! Explaining Predictions, Iterated Learning Bias, and Counter-Polarization in Big Data Discovery Models

CCS@Lexington, October 16, 2017

Olfa Nasraoui

This work is a Collaboration with:

Behnoush Abdollahi, Mahsa Badami, Sami Khenissi, Wenlong Sun, Gopi Nutakki, Pegah Sagheb: @UofL

& Patrick Shafto: @Rutgers-Newark

Knowledge Discovery & Web Mining Lab

Computer Engineering & Computer Science Dept.

University of Louisville

http://webmining.spd.louisville.edu/

olfa.nasraoui@louisville.edu

Acknowledgements:

National Science Foundation:

NSF INSPIRE (IIS)- Grant #1549981

NSF IIS - Data Intensive Computing Grant # 0916489

Kentucky Science & Engineering Foundation: KSEF-3113-RDE-017