Recommendation Systems
part 1

School for advanced sciences of Luchon 2015

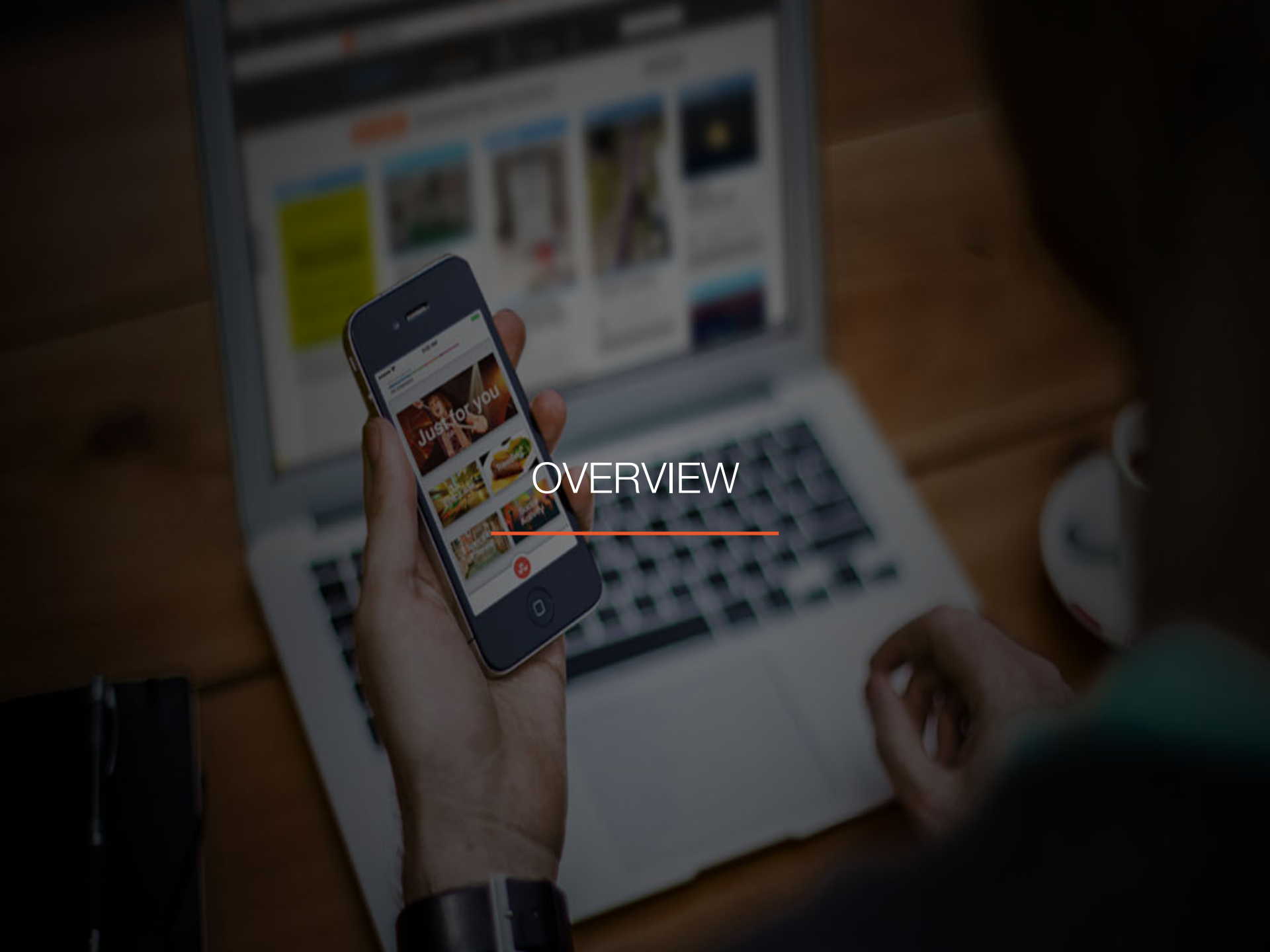
Debora Donato
debora@stumbleupon.com

Today presentation

- Recommendation systems overview
  - Challenges
- Definitions
- Problem space
  - Metrics

About me

- Senior Director of Engineering and Principal Data Scientist at StumbleUpon
  - Leading the Personalization and the Data Science team
- Previously Senior Applied Scientist in the Search Team at Yahoo! Lab
  - Web Mining and Web Information Retrieval
  - User Modeling
  - Spam Detection and Demotion



OVERVIEW

NETFLIX

- Single item type
- Few K items
- 276 categories
- Hand-labeled

- Hand-labeled
- Item-item similarity based methods

amazon

- +200M items
- ~30M recs/mo.
- Auto features
- ~200 methods

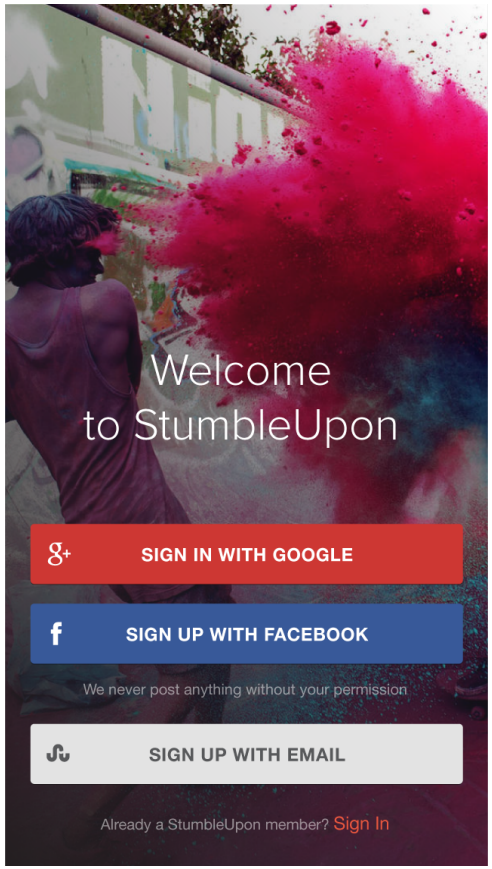
Flipboard

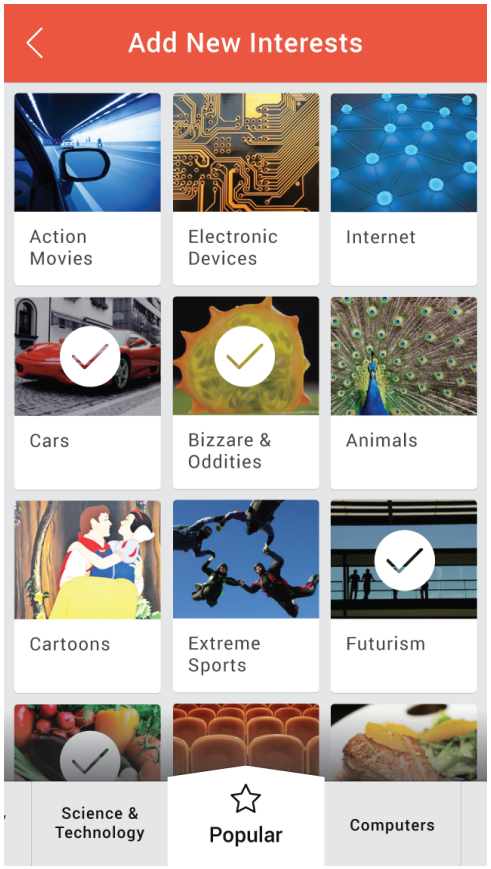
- Mostly about presentation
- Social recs only
- 10 million recs/ month

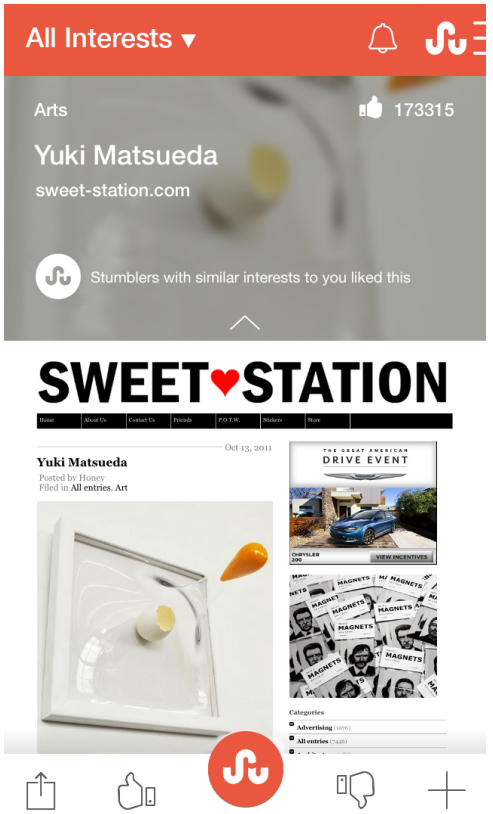
- No serendipity
- Many at a time
- Not personalized*
- Repeats

Google™

StumbleUpon – Choose Topics, Discover Content

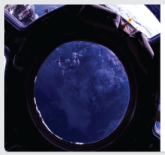






Bookmark, Organize and Share

History



Science & Technology
View From the ISS at Night

Photography
Plants in Space

Football
This is a Really Long Title. So Long in Fact ...

Bizzare & Oddities
Moonbows and Other Rarities

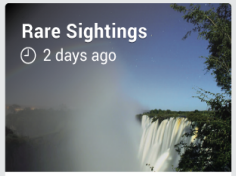
Architecture
What Now, Winchester?

Space & Technology

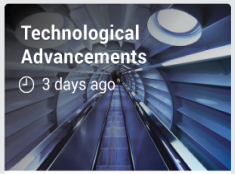
Lists

Lists Created | Lists Following

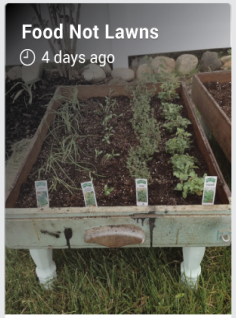
Strange Architecture
30 min ago
13 Pages 3 Following



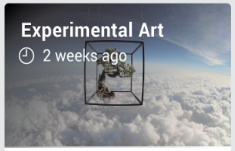
Rare Sightings
2 days ago
24 Pages 14 Following



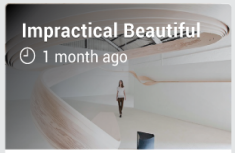
Technological Advancements
3 days ago
6 Pages 7 Following



Food Not Lawns
4 days ago
39 Pages 3 Following

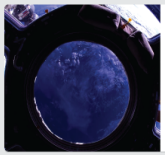


Experimental Art
2 weeks ago
6 Pages 7 Following



Impractical Beautiful
1 month ago

Share



View From the ISS at Night
thisistheurl.com



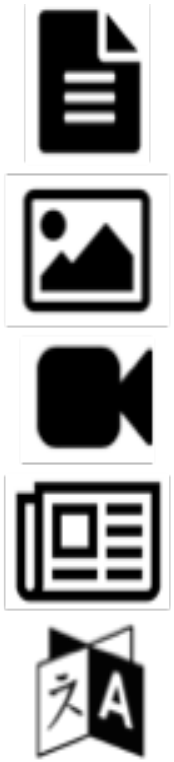

3 more

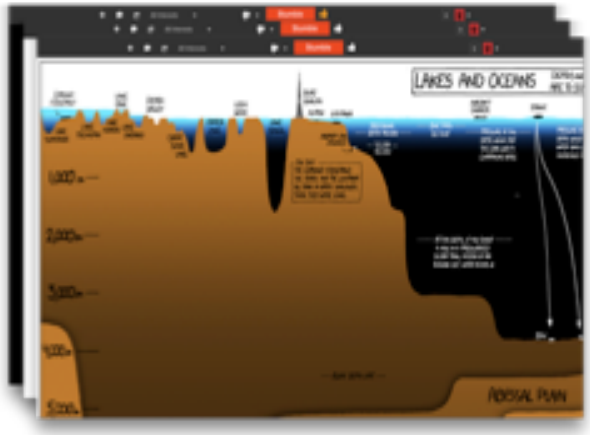
Your Message

Cancel

Send

Recommendations: Matching User With Content





1. Understand User 2. Understand Content 3. Recommend 4. Get Feedback

Item Lifecycle in a Recommender System

Ingestion

Entry point for items;
Feature extraction

Initial Recs

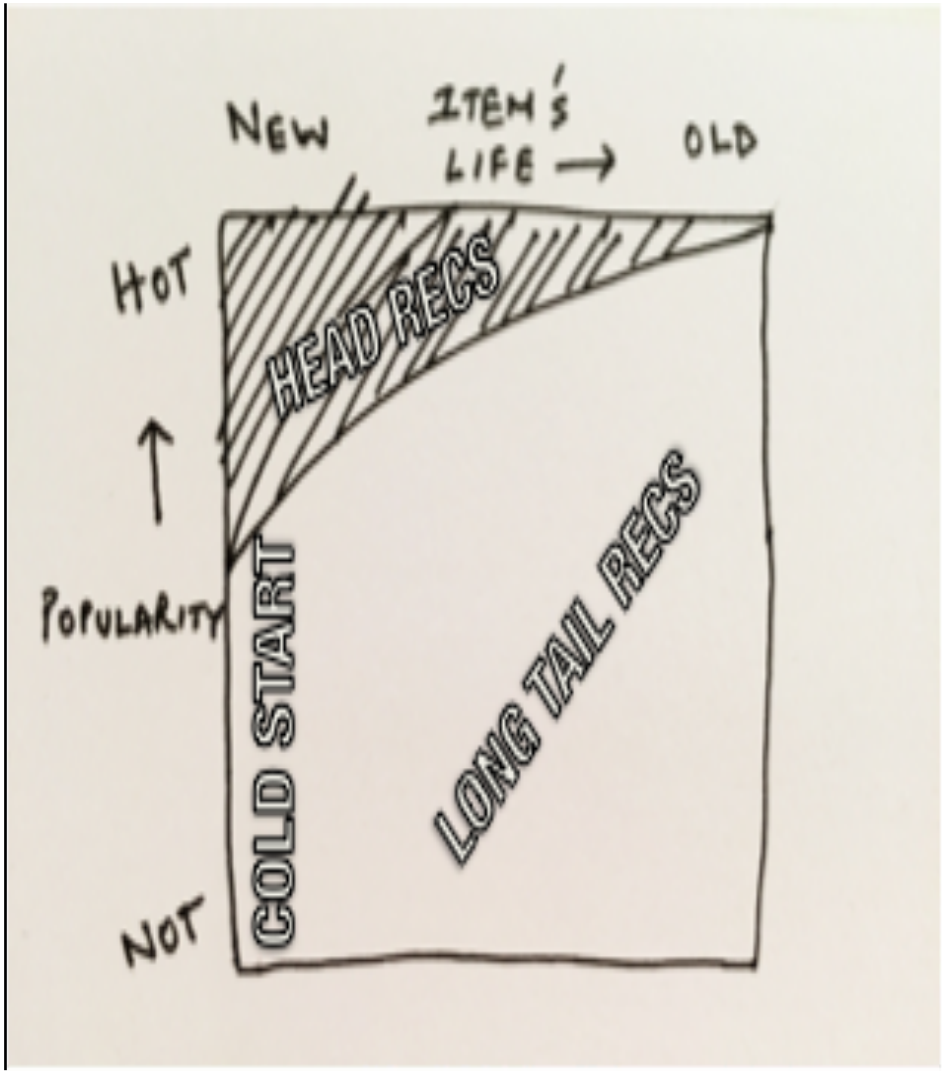
Cold start Optimize for maximum expected positive ratings and satisfy item demand

Head Recs

Trending Popular in the short run
Timeless Popular in the long run

Tail Recs

Collaborative Filtering Nearest neighbors based on user signals
Serendipitous Recs Unexpected but relevant



Content discovery (a.k.a. serendipitous search)

DEFINITION

- Recommend the “unexpected but personally relevant”
- “Go beyond relevance” and look for “interestingness”

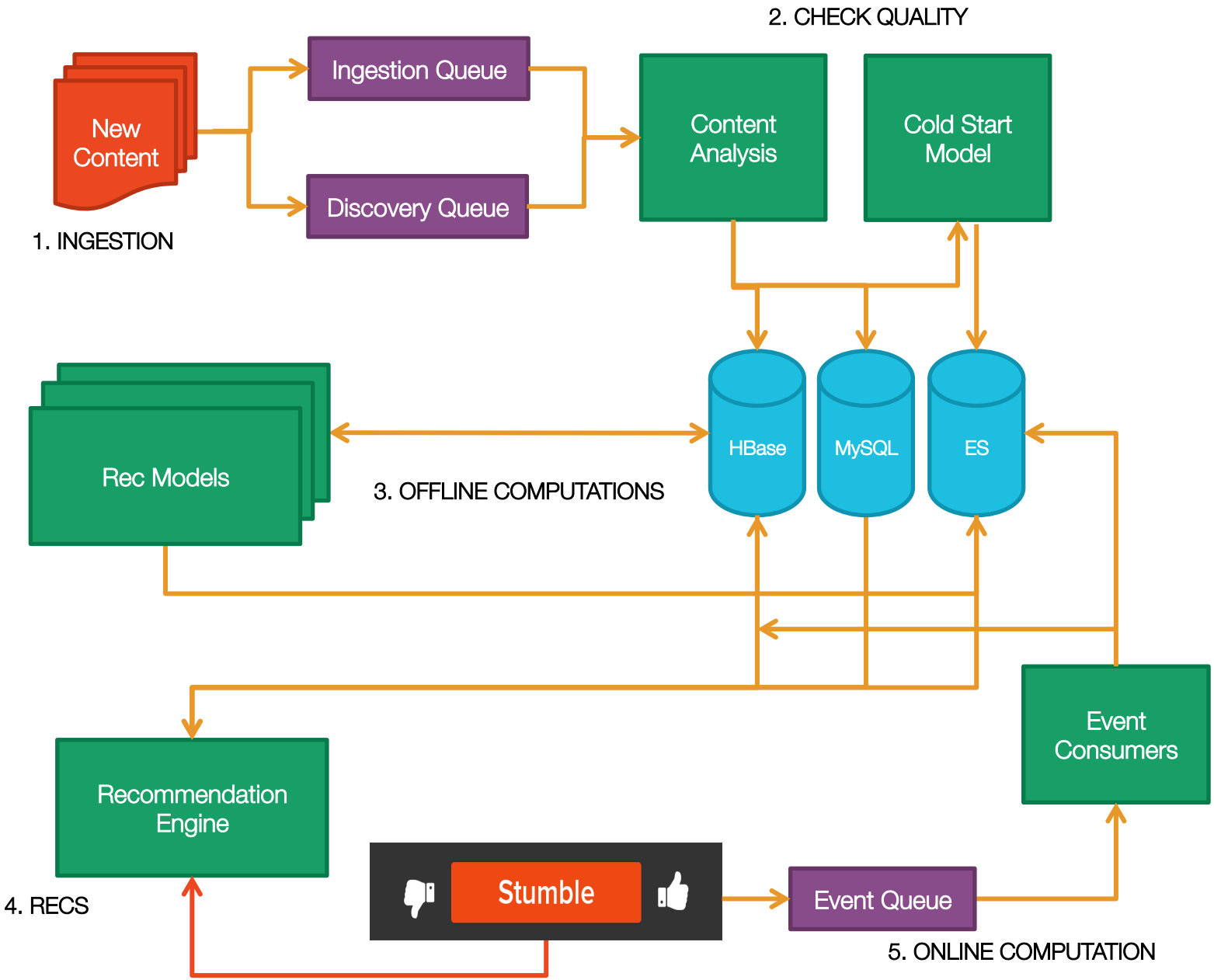
MOTIVATION

- Avoiding “tunnel vision/ filter bubble“
- Allows exploration enabling true discovery

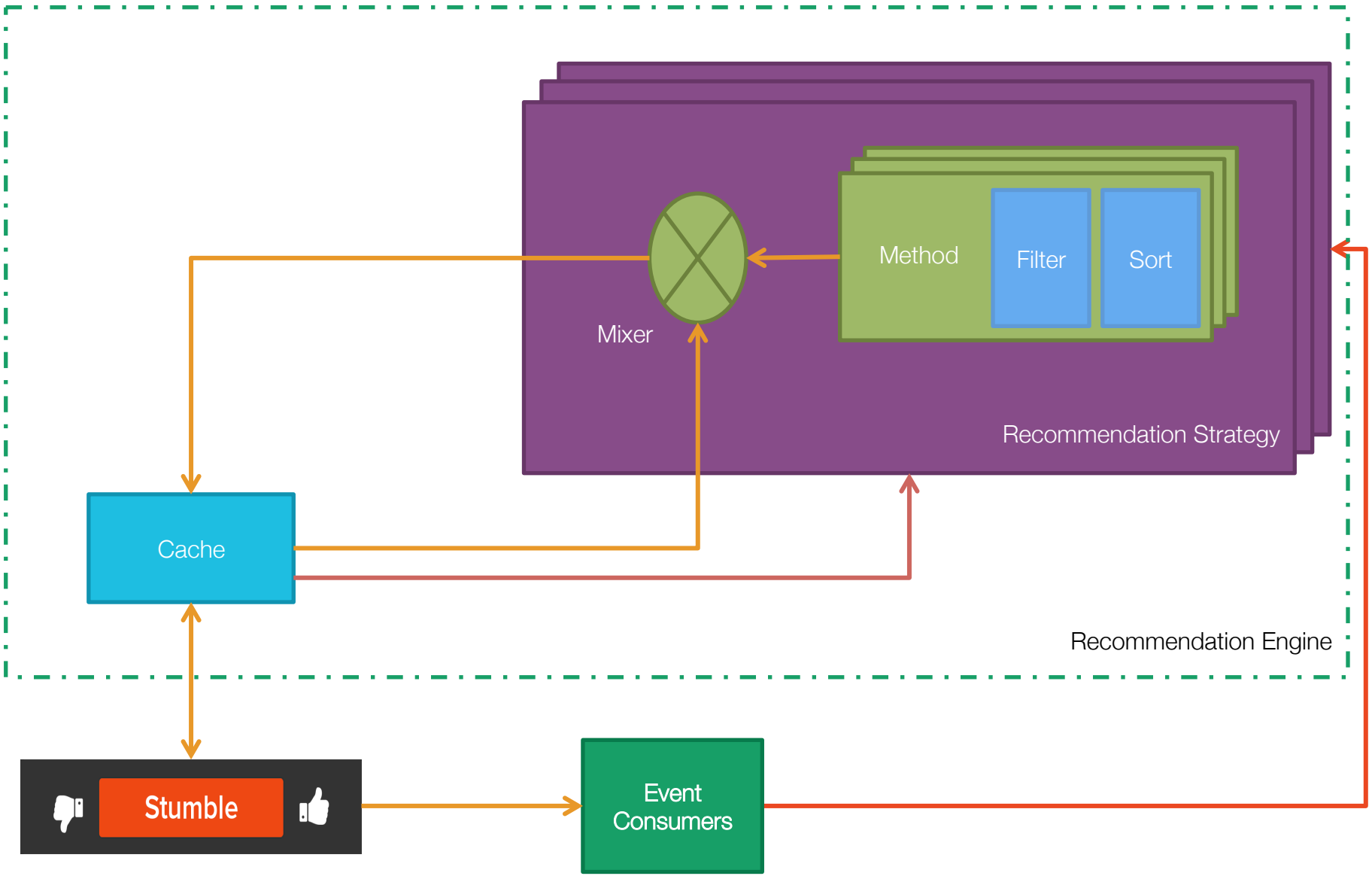
CHALLENGE

- Serving fantastic content that is not random, is unexpected but still useful
- Measuring/Controlling Serendipity

Architecture I

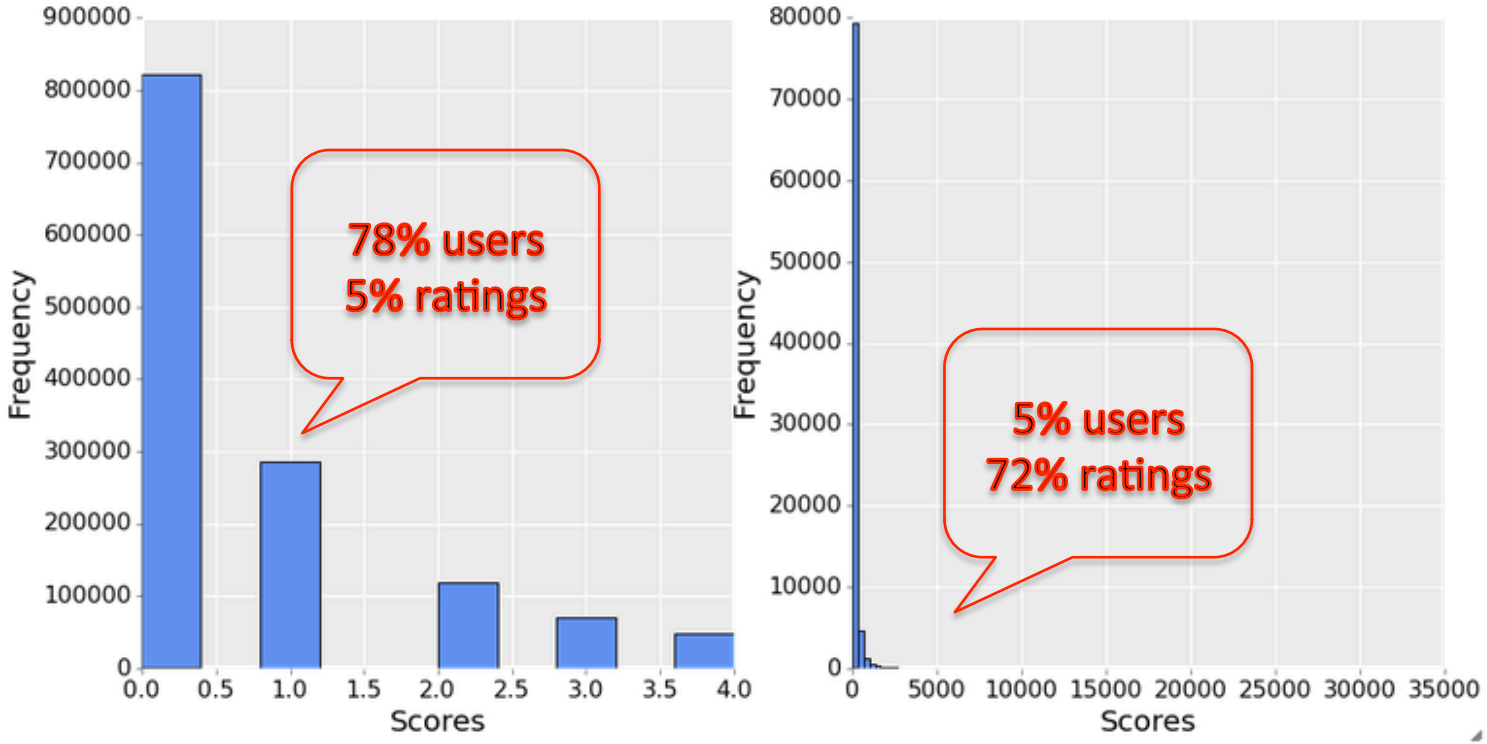


Architecture II



Major Challenges: Data Sparsity

- Power law curse
  - Tendency of the users to perform some actions (ratings)
  - Data sparsity



Sparsity … in practice

- Random sampling
  - High probability to sample from the tail of the distribution.
  - MF accuracy ~ 97%
  - Learning about users who have tendency to rate is not real useful

- Bucketing across all the possible behavioral and popularity segments
  - Density of the user/item matrix ~ 1% (same of the Netflix dataset)
  - MF accuracy ~63%
  - Not “productionizable”

Major Challenges: Scalability

- 3,000,000,000 STUMBLES PER YEAR
- 200,000,000 USER SUBMITTED PAGES
- 120,000,000 LIKES PER YEAR
- 35,000,000 REGISTED USERS

Major Challenges: Cold Start

- New users:
  - None or basic information
  - Popularity-based recommendation
  - Segment-based recommendation

- New items:
  - Content understanding:
    - Classification
    - Filtering
  - Sampling:
    - Targeting
    - Exploration - exploitation.

Major Challenges: Diversity vs Accuracy

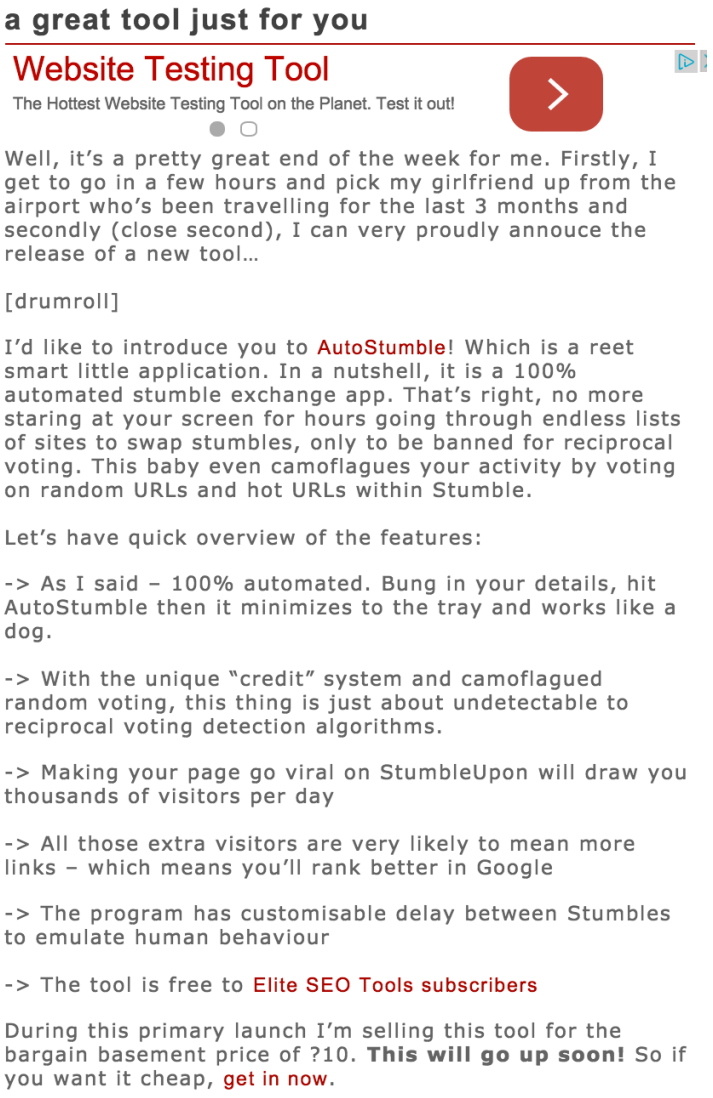
- High accuracy recommendations have usually little value:
  - You liked Star War I, II, III… not surprisingly you will like IV, V, VI or whatever follows
  - More importantly, you do not need the system to output a recommendation for you.

- The value of Serendipity:
  - Content that is still relevant for you but it is somehow surprisingly.
  - Understanding the effect of diversity in “recommendation sessions”

Major Challenges: Vulnerability

Reasons of malicious behavior:

- E-commerce impact of recommendations
- Social networks and media can drive huge amount of traffic to publishers, bloggers and brands.
  - SU is the 4th source of social referrals (FB, Pinterest, Twitter)
  - Incredible opportunity for arbitrage.



a great tool just for you

Website Testing Tool

The Hottest Website Testing Tool on the Planet. Test it out!

Well, it's a pretty great end of the week for me. Firstly, I get to go in a few hours and pick my girlfriend up from the airport who's been travelling for the last 3 months and secondly (close second), I can very proudly annouce the release of a new tool...

[drumroll]

I'd like to introduce you to AutoStumble! Which is a reet smart little application. In a nutshell, it is a 100% automated stumble exchange app. That's right, no more staring at your screen for hours going through endless lists of sites to swap stumbles, only to be banned for reciprocal voting. This baby even camoflagues your activity by voting on random URLs and hot URLs within Stumble.

Let's have quick overview of the features:

-> As I said – 100% automated. Bung in your details, hit AutoStumble then it minimizes to the tray and works like a dog.

-> With the unique "credit" system and camoflagued random voting, this thing is just about undetectable to reciprocal voting detection algorithms.

-> Making your page go viral on StumbleUpon will draw you thousands of visitors per day

-> All those extra visitors are very likely to mean more links – which means you'll rank better in Google

-> The program has customisable delay between Stumbles to emulate human behaviour

-> The tool is free to Elite SEO Tools subscribers

During this primary launch I'm selling this tool for the bargain basement price of ?10. **This will go up soon!** So if you want it cheap, get in now.

Major Challenges: Time

- Most recommendation algorithms neglect the time stamps of evaluations

- Two different aspects to be kept into consideration:
  - Stages of life:

    Pregnancy -> Baby -> Parenting ->Teen
  - Different temporal patterns of relevance

    News (few hours – few days)

    Technology (few months – few years)

    …

    History, Math,Cooking

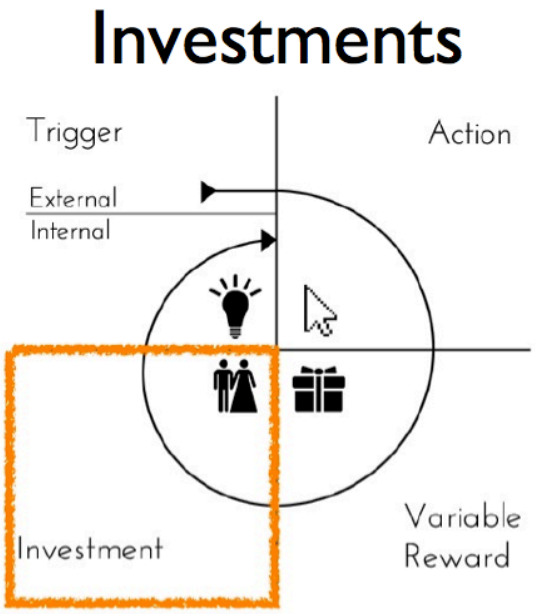
Major Challenges: Evaluation

- Plethora of off-line metrics
  - How to choose the right one?
- Rec system comparison is difficult
  - they may solve different problems
- Online user ‘reaction’ can not be evaluated offline
  - Trust
  - Biases
  - Context

Major Challenges: Engagement

"Accurate" personalized recommendation algorithms are sufficient to drive engagement and retain users.

Usability, UI visual appeal, saliency, social interactions, etc. may influence engagement.



http://www.scribd.com/doc/106584363/Investment-Phase-of-Desire-Engine

Challenges we solve

CONTENT UNDERSTANDING

- Automatic categorization
- Keyword extraction
- Language extraction
- Evergreen vs ephemeral content
- Video and photo mode
- News clustering
- Social signals
- Device-based preferences

USER MODELING

- Likeminded users
- Expert model
- Explicit and implicit feedback
- Demographic and behavioral characterization
- Churn model
- Interest identification
- Measuring engagement

RECOMMENDATIONS

- Cold start problem
- Sparse signals
- Heterogenous content
- Personalization
- Contextualization
- Session strategy
- Ad targeting & yield management

ABUSE PREVENTION

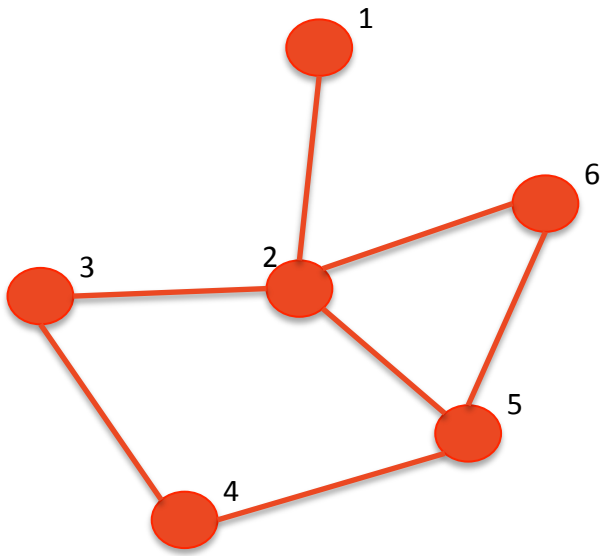
- Bot Detection
- Spammer Detection
- Self-promoters
- Dupe detection

DEFINITIONS AND PROBLEM SPACE

A network G is a ordered pair of disjoint sets (V, E) where:

V is the set of nodes

$E \subseteq V \times V$ is the set of edges.



- Undirected network
- Directed network
- Self-loop, multi-edge, multinetworks
- Adjacent nodes,
- Neighborhood Γ_x of a node x
- Degree $k_x = |\Gamma_x|$ of a node x
- Out-degree k_x^{out} and In-degree k_x^{in}

A network $G(V,E)$ is a *bipartite network* if there exists a partition $(V_1,V_2): V_1 \cup V_2 = V, V_1 \cap V_2 = \varnothing$ and $E \subseteq V_1 \times V_2$

Bipartite networks which represent interactions between users and objects in online service sites, describe the fundamental structure of recommender systems.

Tripartite network has been used to represent collaborative tagging systems (also called folksonomies).

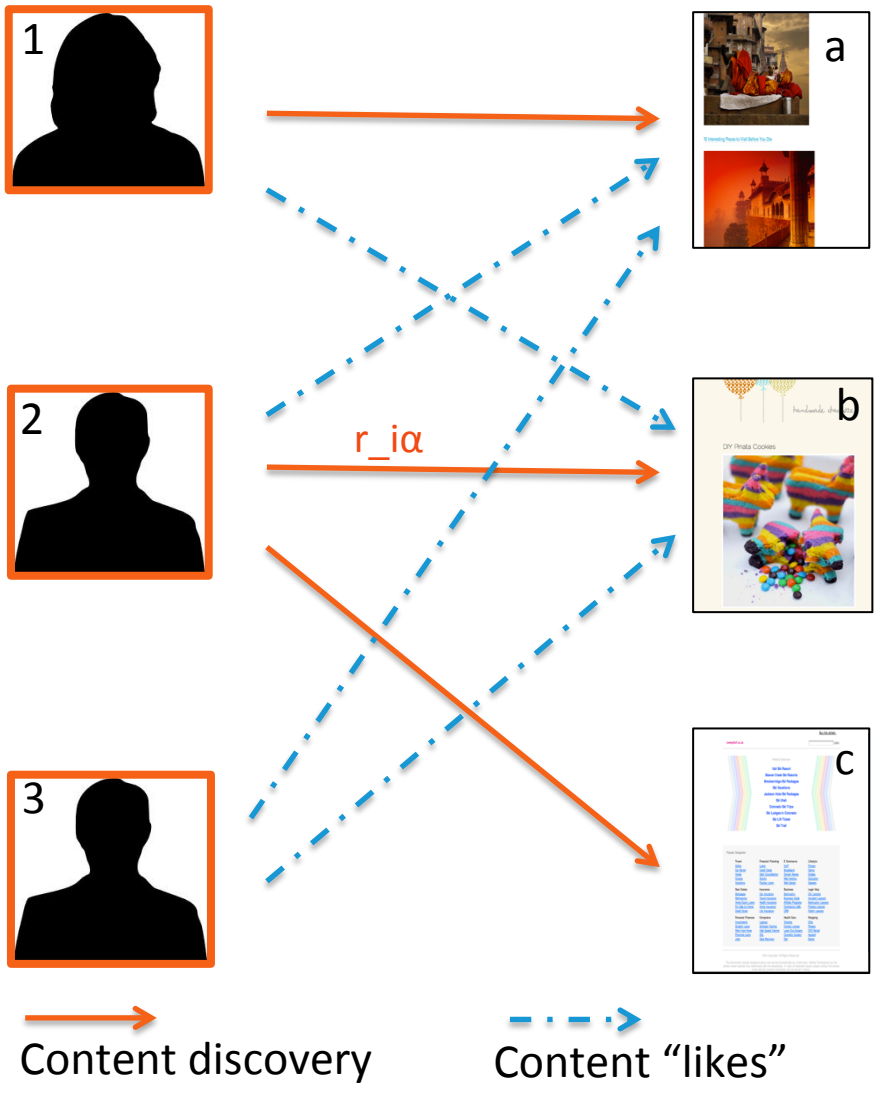
M is the set of the users

i = <age, gender, interests, language, location>

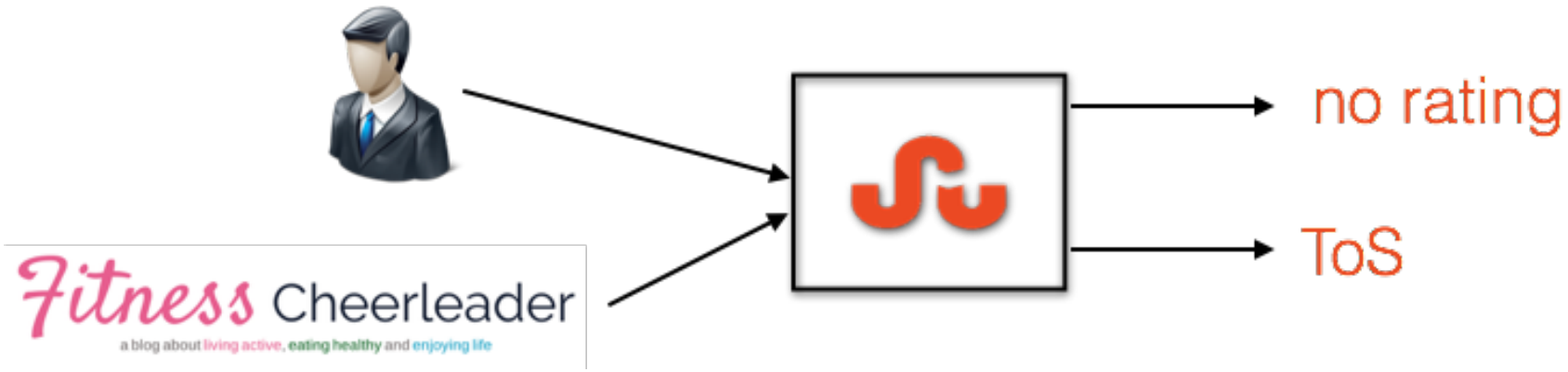
N is the set of items

α =<category, age, news, tags,…>

$r_{ij} = \{-1, 0, 1\}$ represent the rating of the object j by the user i

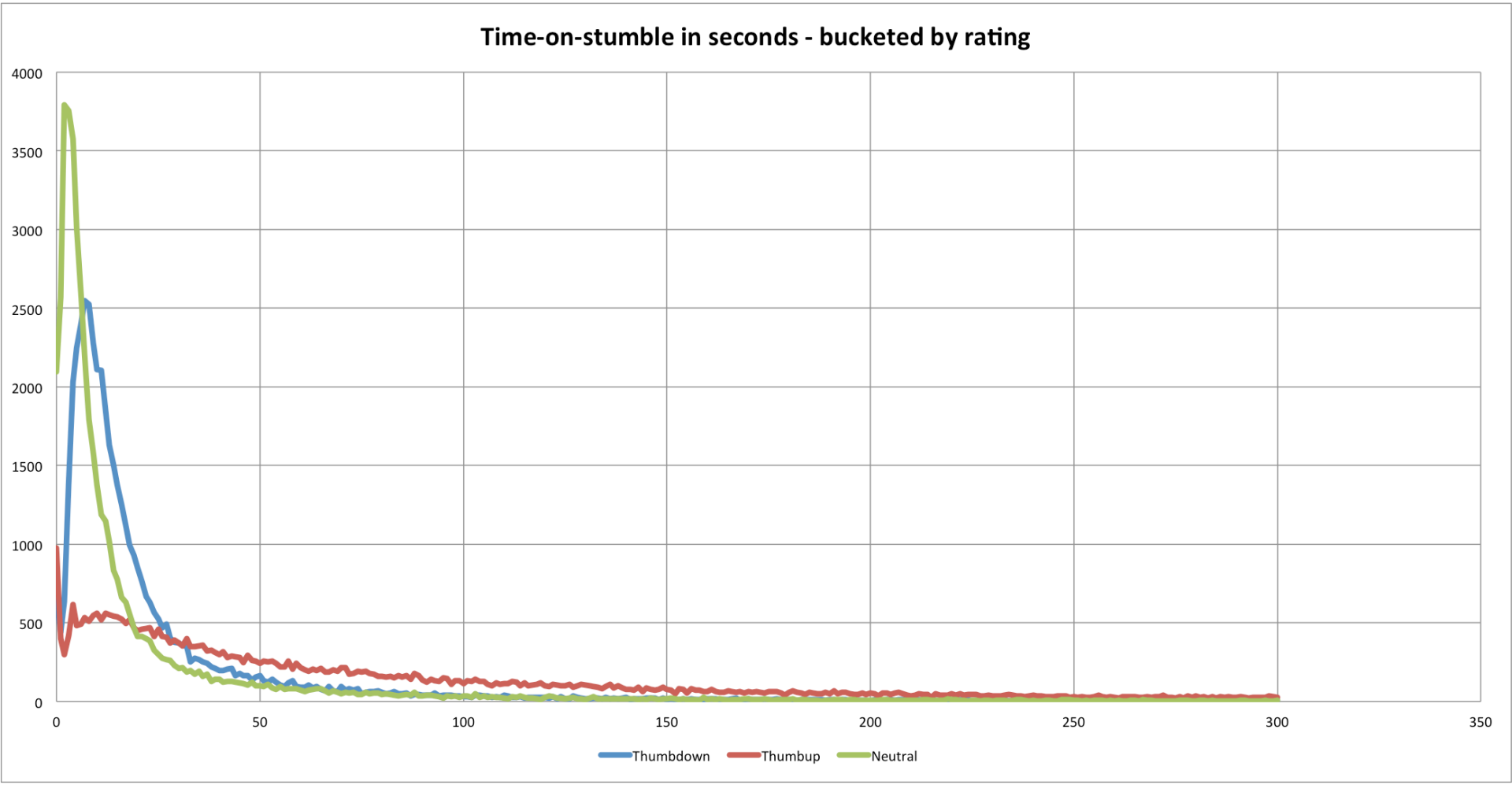


Implicit signals



- Signal correlated with user involvement with current url
- ToS, #clicks, scrolling, #playback
- Classification of unrated transaction into likes/ dislikes based on implicit feedback

Positive and Negative feedback vs ToS



Evaluation of recommendation systems

For a given user i, a recommendation system can

- predict single items ratings

or

- rank the top-k relevant item

The rating matrix is then partitioned in two set:

- Training set E^P
- Testing set E^T

In the following:

- $r_{i\alpha}$ is the real rating or relevance given to the item α by the user I
- $\tilde{r}_{i\alpha}$ is the predicted rating or relevance

Rating Accuracy Metrics

Main rating accuracy metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

$$MAE = \frac{1}{\left|E^P\right|} \sum_{(i,\alpha)\in E^P} \left|r_{i\alpha} - \tilde{r}_{i\alpha}\right|$$

- Pros:
  - Simplicity
  - Appropriate for explicit ratings

$$RMSE = \left(\frac{1}{\left|E^P\right|} \sum_{(i,\alpha)\in E^P} \left(r_{i\alpha} - \tilde{r}_{i\alpha}\right)^2 \right)^{1/2}$$

- Cons:
  - item rank is not considered
  - RMSE penalize large errors more heavily

Rating and ranking correlation metrics:

- The Pearson correlation (PC) measures the linear correlation between two set of ratings

$$PC = \frac{\sum_{\alpha}\left(\tilde{r}_{\alpha} - \bar{\tilde{r}}\right)\left(r_{\alpha} - \bar{r}\right)}{\sqrt{\sum_{\alpha}\left(\tilde{r}_{\alpha} - \bar{\tilde{r}}\right)^2}\sqrt{\sum_{\alpha}\left(r_{\alpha} - \bar{r}\right)^2}}$$

- The Spearman correlation (ρ) is defined as PC using ranks instead of ratings.

- The Kendall's Tau (τ) measures the extent to which the two rankings agree on the exact values of ratings.
  - Cons: it applies equal weight to any interchange of successively ranked objects, no matter where it occurs.

$$\tau = \frac{C - D}{C + D}$$

Classification Accuracy Metrics

Classification Accuracy metrics are appropriate

- For list of relevant objects
- When the rating are not explicit

The Area Under ROC Curve (AUC) measures the probability relevant items will be identified.

$$AUC = \frac{n' + 0.5n''}{n}$$

Computed* by performing n independent comparisons (choosing one relevant and one irrelevant object) and counting

- The number n' of times score(rel) > score (irrel)
- The number n" of times score (rel) = score (irrel)

*T. Zhou, L. Lu¨, Y.-C. Zhang, Predicting missing links via local information, The European Physical Journal B 71 (2009)

Other top-k evaluation metrics

- Recall @K: (# correct predicted)/(# relevant items)
  - $R(K) = \frac{1}{N} \sum_i R_i(K)$
  - where $R_i(K) = \frac{1}{|D_i|} \sum_{\alpha=1}^{K} rel_{i,\alpha}$

- Normalized Discounted Cumulative Gain @K
  - $NDCG(K) = \frac{1}{N} \sum_i NDCG_i(K)$
  - where $NDCG_i(K) = \frac{DCG_i(K)}{iDCG_i(K)}$
  - and $DCG_i(K) = rel_{i,\alpha} + \sum_{\alpha=2}^{K} \frac{rel_{i,\alpha}}{\log_2 \alpha}$

- Inter-user diversity
  - Given users i and j, the difference between the top-K recommendations can be measured by the Hamming distance

$$H_{ij}(K) = 1 - \frac{Q_{ij}(K)}{K}$$

- Intra-user diversity
  - Given the user I and the top-K ranked items, the average similarity of such items can be measured by

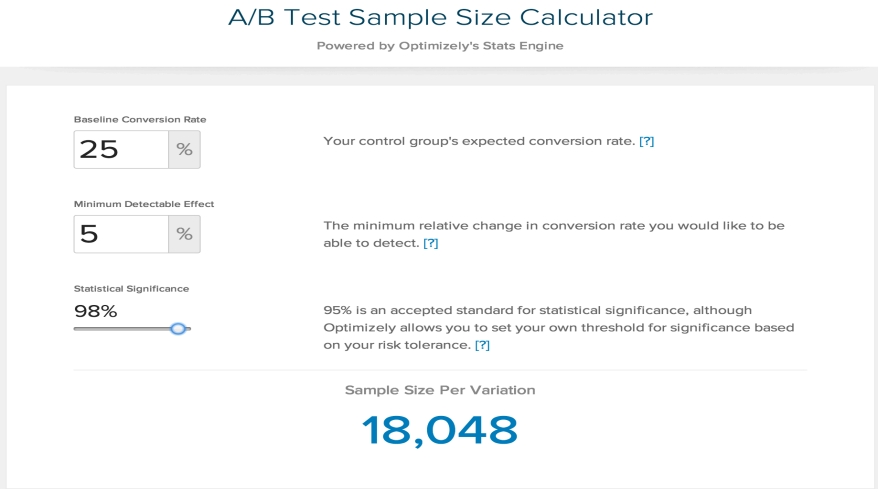
$$I_i(K) = \frac{1}{K(K-1)} \sum_{\alpha \neq \beta} sim\left(item_\alpha, item_\beta\right)$$

  - where $sim\left(item_\alpha, item_\beta\right)$ is usually defined using the item metadata

A/B Testing

- A/B tests are run to determines which one of two or more ‘variations’ of a page actually leads to more conversion on business objectives.
- Power Analysis is used to determine the sample analysis given a specified power or to calculate the power given a specific sample size.
- http://www.ats.ucla.edu/stat/r/dae/t_test_power2.htm

https://vwo.com/ab-testing/



https://www.optimizely.com/resources/sample-size-calculator