Predicting the Popularity of User Generated Content

Jussara M. Almeida

Department of Computer Science
Federal University of Minas Gerais – Brazil

jussara@dcc.ufmg.br

User Generated Content (UGC)

- Unprecedented scale and growing rate
  - YouTube: over 72 hs of video uploaded per min
    more than 4 billion videos watched daily
  - Foursquare: over 40 million people worldwide
    over 4.5 billion check-ins
    millions more on a daily basis

User Generated Content (UGC)

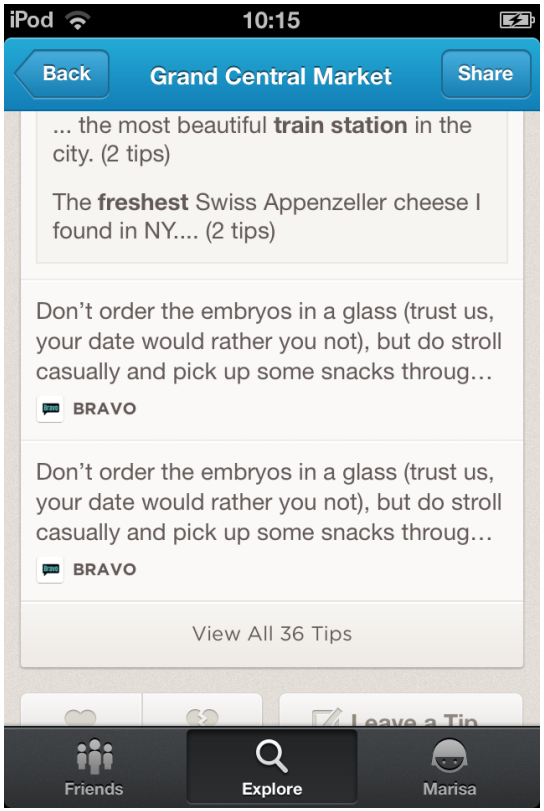
- Popularity distribution is highly skewed

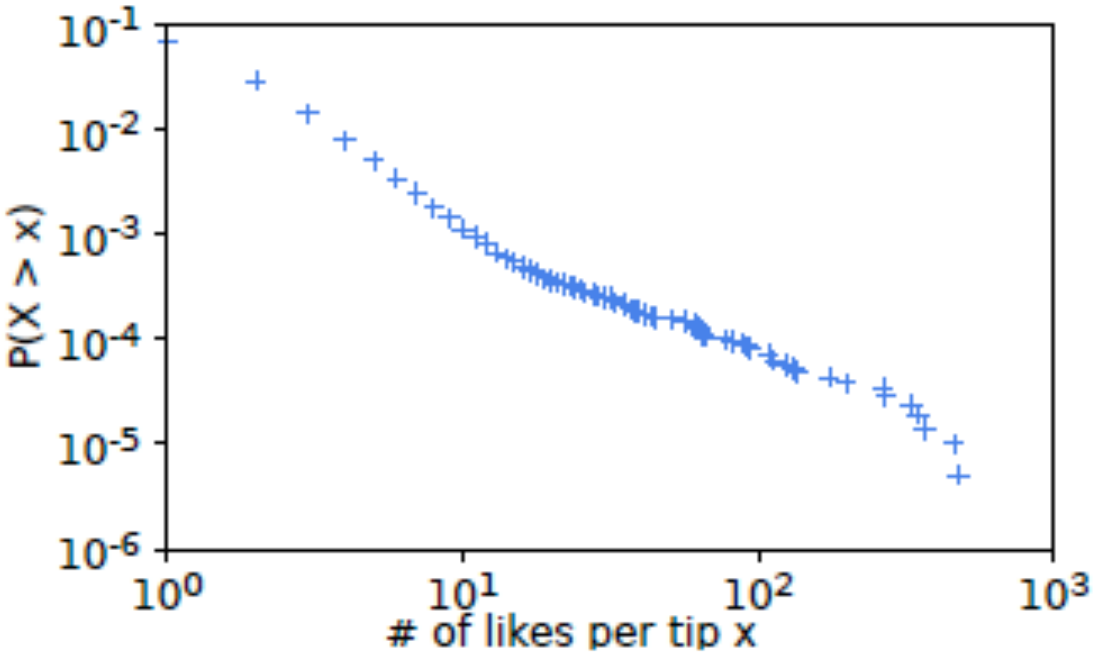
Most objects attract little attention

Few objects attract millions user views

User Generated Content (UGC)

- Popularity distribution is highly skewed





Ongoing Research

Which factors impact UGC popularity evolution and how can we exploit them to build simple yet reasonably accurate popularity prediction models?

Popularity Prediction: Why?

- Content distribution services (CDNs, caching)
- Searching services
- Advertising and marketing strategies
- Content filtering, ranking, and recommendation
- Customer feedback (Foursquare tips)
- Understand human dynamics of information consumption processes

Popularity Prediction: Challenges

- Multitude of factors with potential influence
  - Content itself
  - Social neighborhood or influence zone of user
  - Mechanisms available to drive users to content (search, recommendation, top lists)
  - Specific characteristics of application (ranking by creation time)
  - External factors

Popularity Prediction: Case Studies

- YouTube videos
- Foursquare tips

Popularity Prediction: Case Studies

- YouTube videos

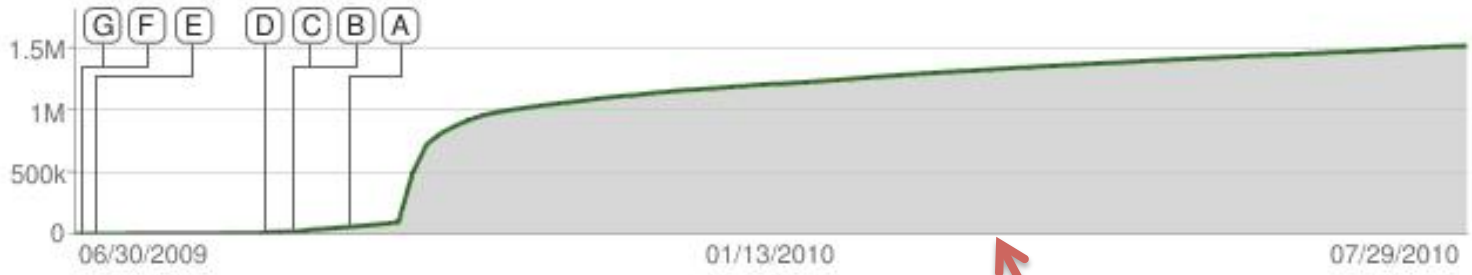
- Foursquare tips

YouTube Datasets

- **Top**: 18k videos that appeared on top lists
- **YouTomb**: 103k videos with copyright violation
- **Random**: 22k videos selected based on random queries

YouTube Datasets

Total Views: 1,517,107



Ratings: 4059

Likes: 3872

Dislikes: 187

Comments: 2,567

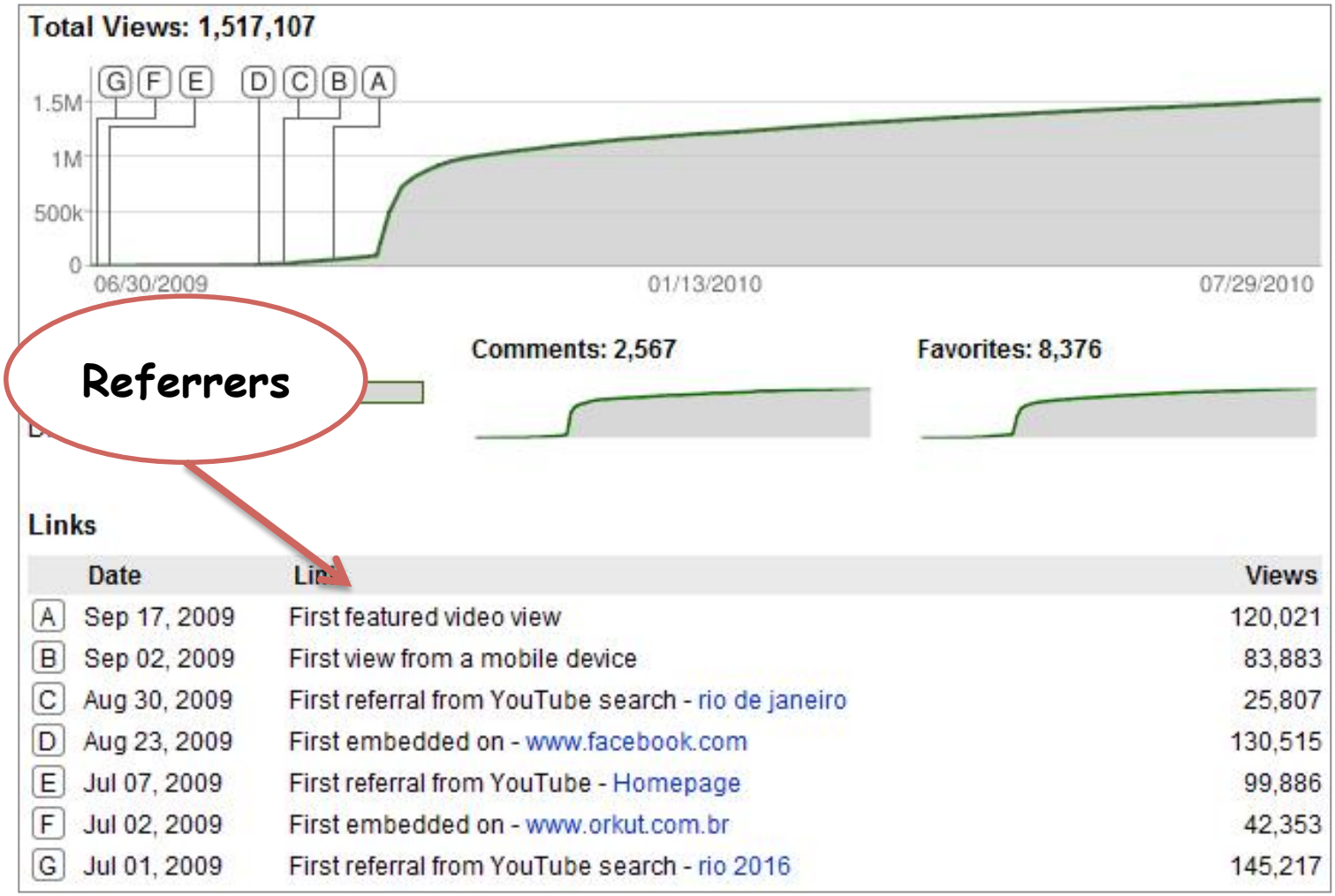
Favorites: 8,376

Popularity Evolution

Links

	Date	Link	Views
A	Sep 17, 2009	First featured video view	120,021
B	Sep 02, 2009	First view from a mobile device	83,883
C	Aug 30, 2009	First referral from YouTube search - rio de janeiro	25,807
D	Aug 23, 2009	First embedded on - www.facebook.com	130,515
E	Jul 07, 2009	First referral from YouTube - Homepage	99,886
F	Jul 02, 2009	First embedded on - www.orkut.com.br	42,353
G	Jul 01, 2009	First referral from YouTube search - rio 2016	145,217

YouTube Datasets

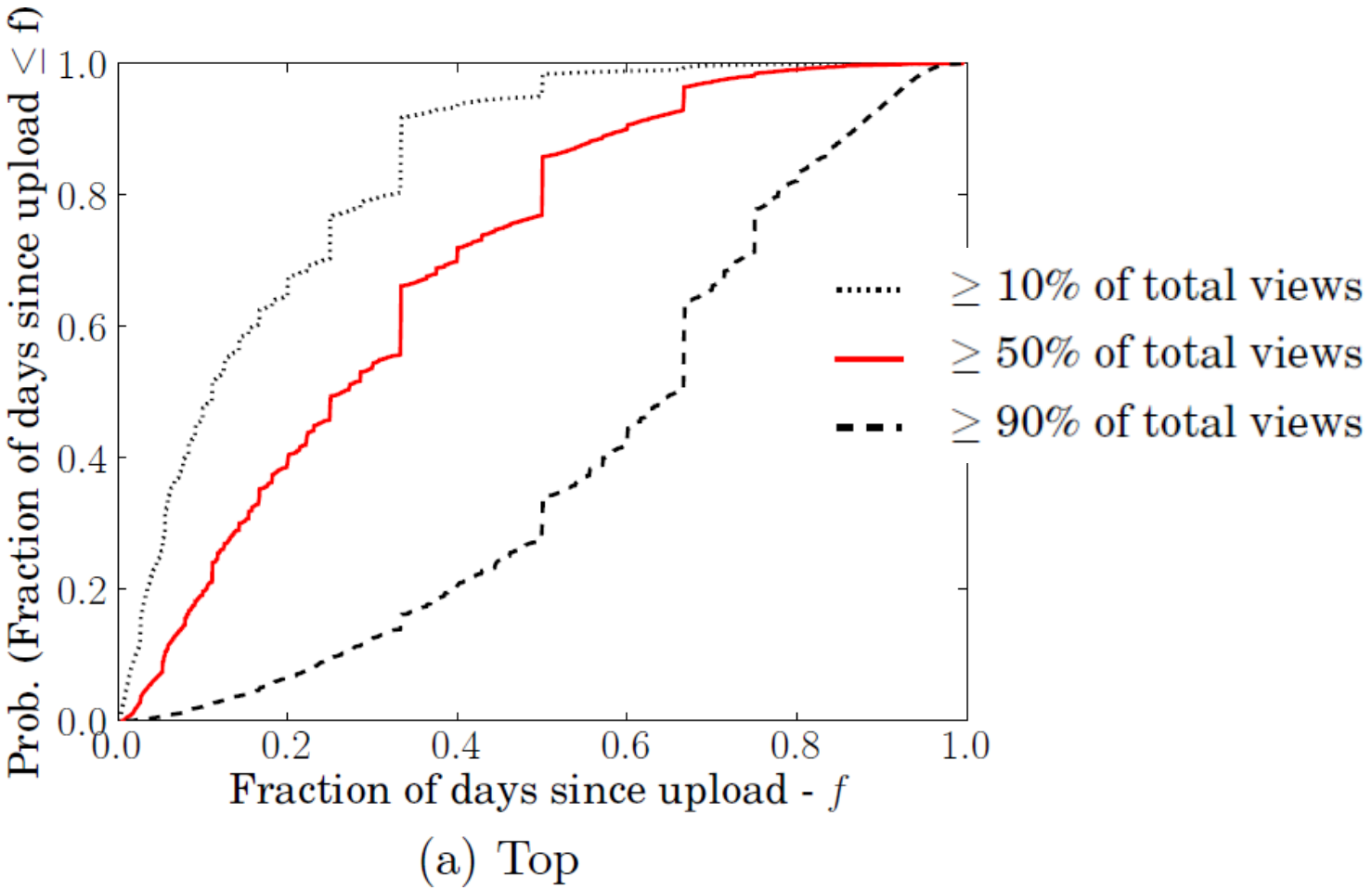


Popularity Evolution: Analysis

- How fast does a video become popular?
- How concentrated is popularity?
- Are there clear popularity trends?
- How content/referrer features correlate with trends?

How Fast?

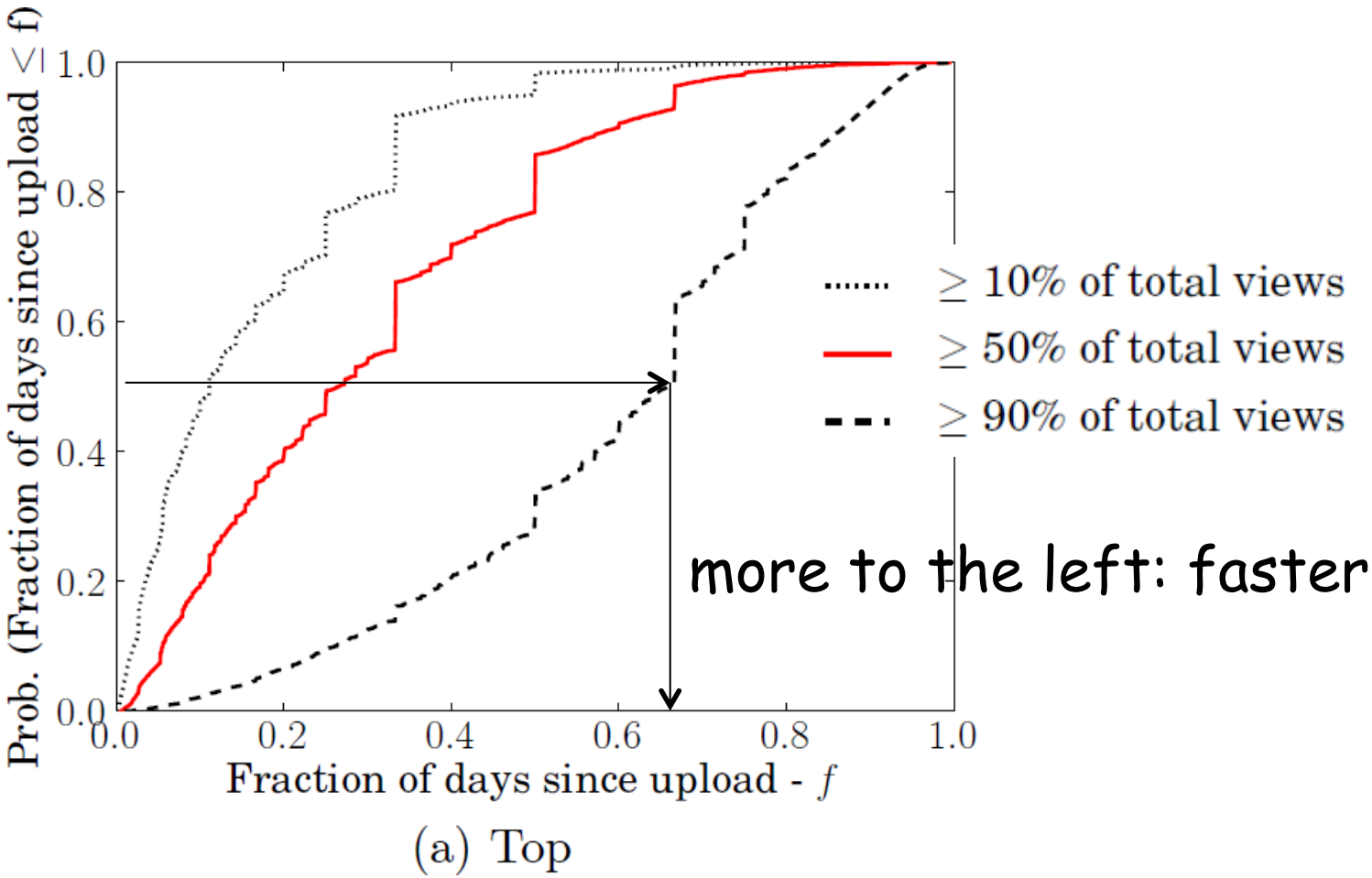
Fraction of time until X% of popularity reached



(a) Top

How Fast?

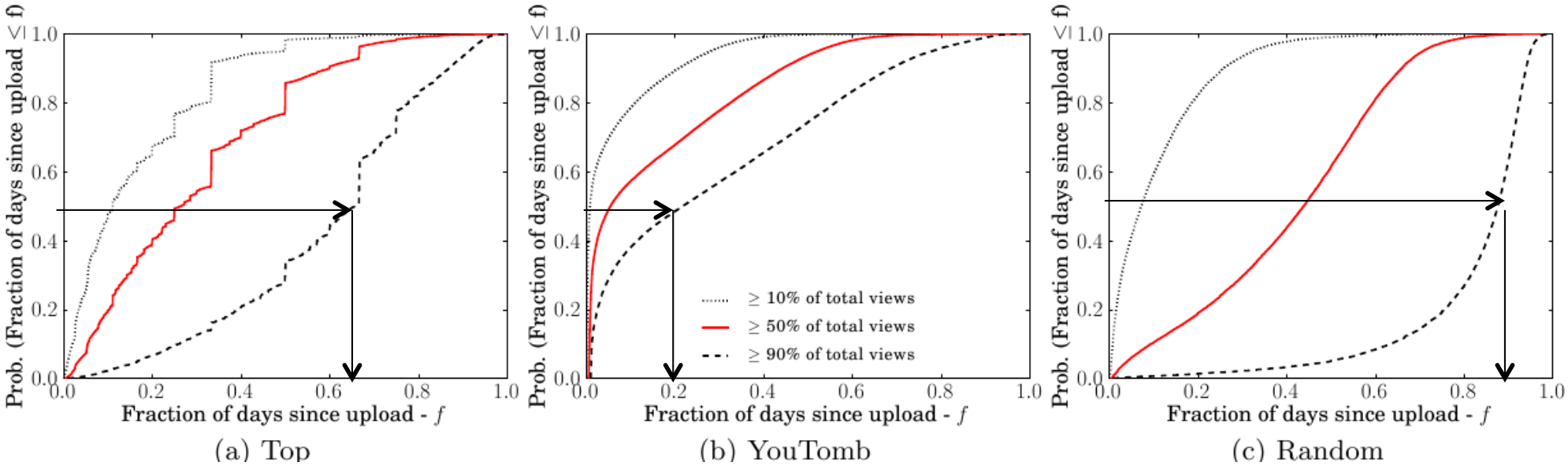
Fraction of time until X% of popularity reached



(a) Top

50% of videos take at most 65% of lifetime to reach 90% of views

How Fast?

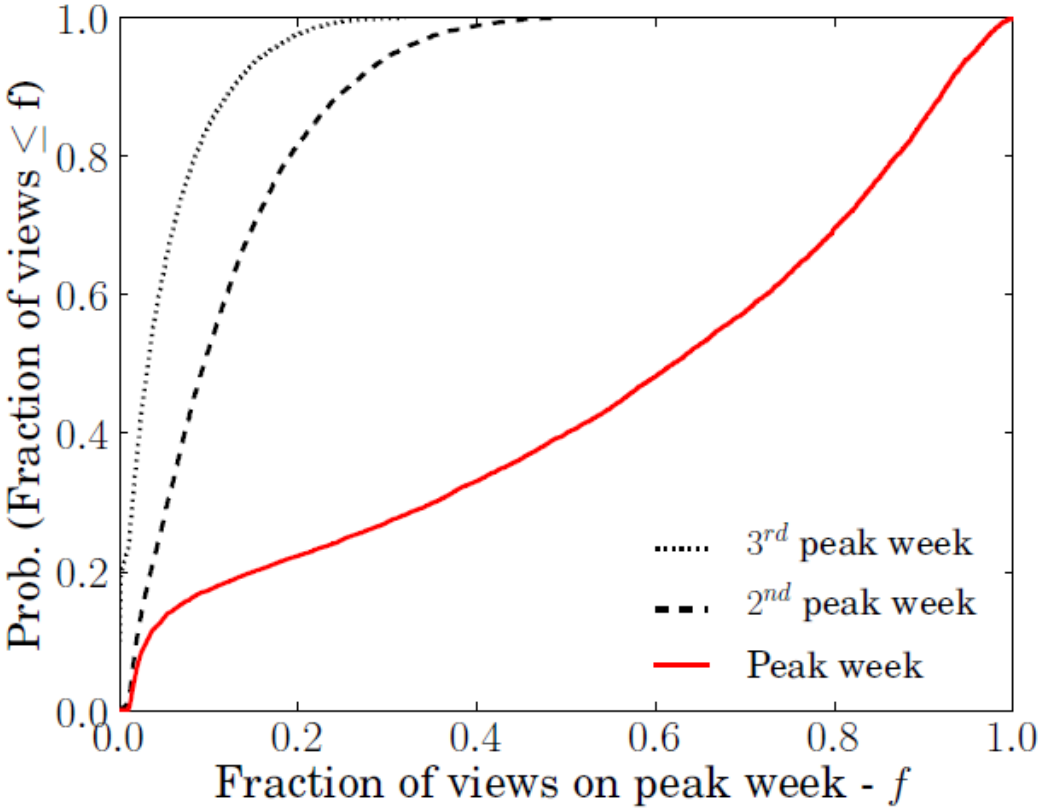


For 50% of the videos:

- YouTomb: ≤ 21% of lifetime to reach 90% of views
- Top: ≤ 65% of lifetime for same 90%
- Random: ≤ 87% for same 90%

How Concentrated?

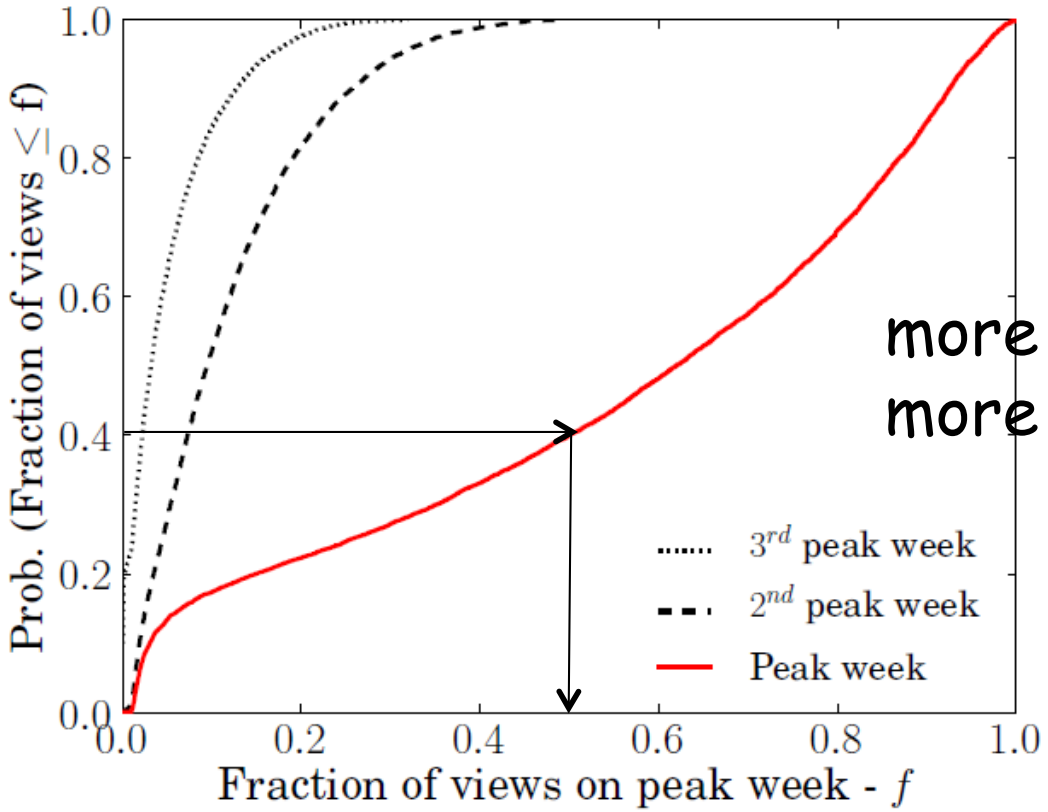
Fraction of views on peak week



(a) Top

How Concentrated?

Fraction of views on peak week

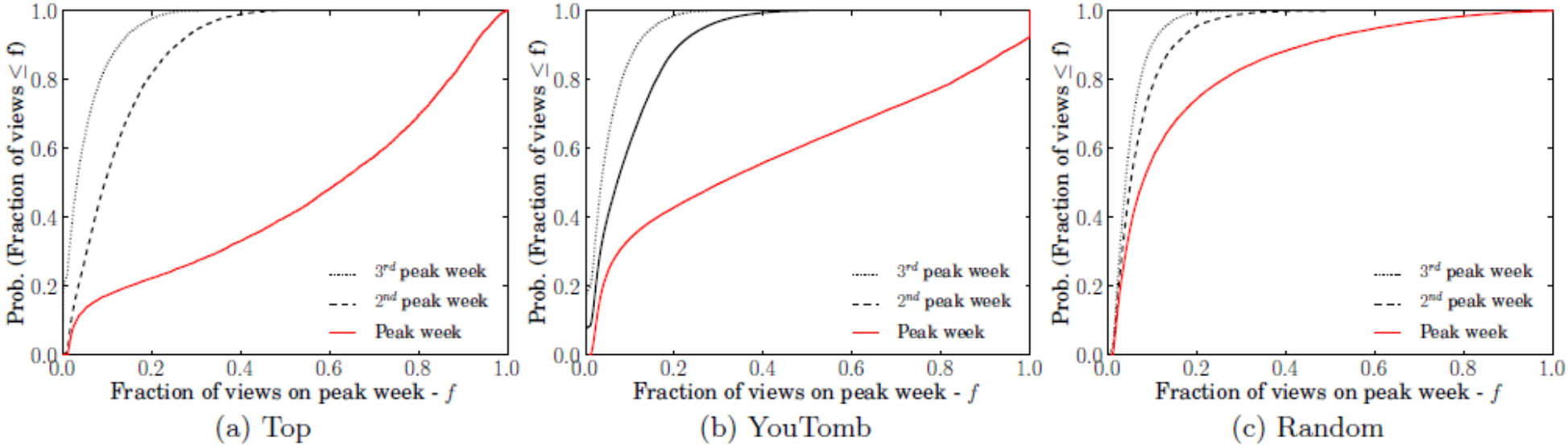


(a) Top

more to the right:
more concentrated

At least 50% of views on peak week for 60% of videos

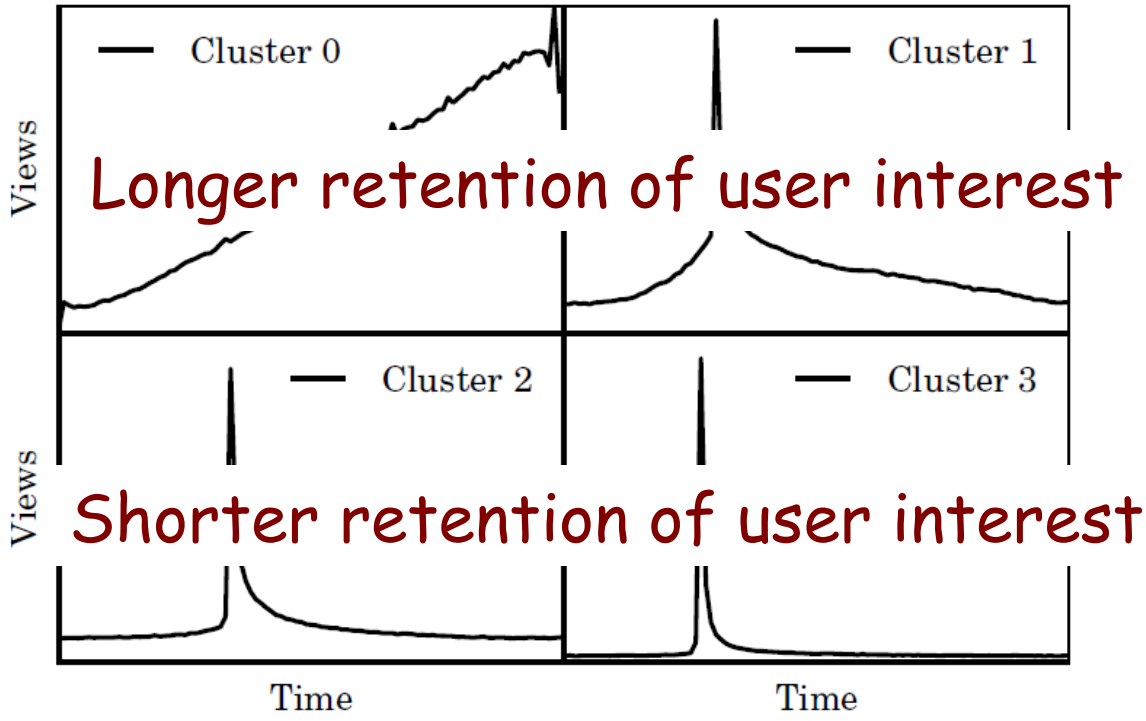
How Concentrated?



For 60% of videos, most popular week consists of:

- At least 50% of views for Top
- At least 40% of views for YouTomb
- At least 5% of views for Random

Are There Popularity Trends?

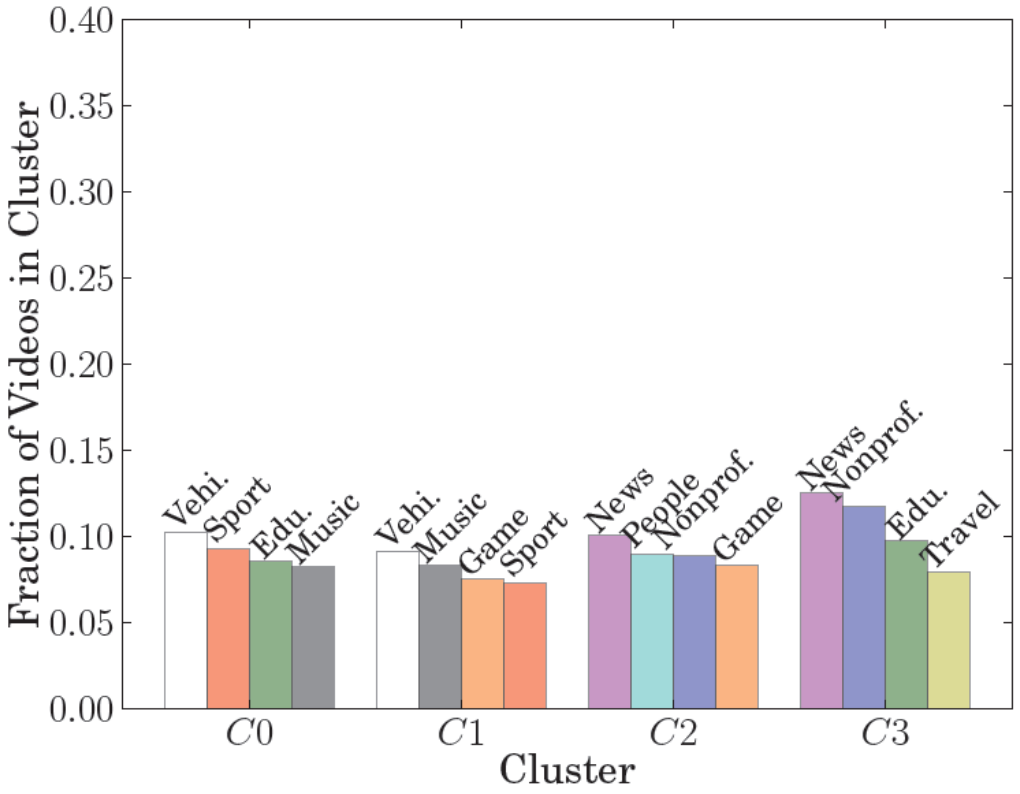


KSC Clustering: time shift and scale invariants [Yang2011]

- 4 Clusters in all datasets

Types of Content per Cluster

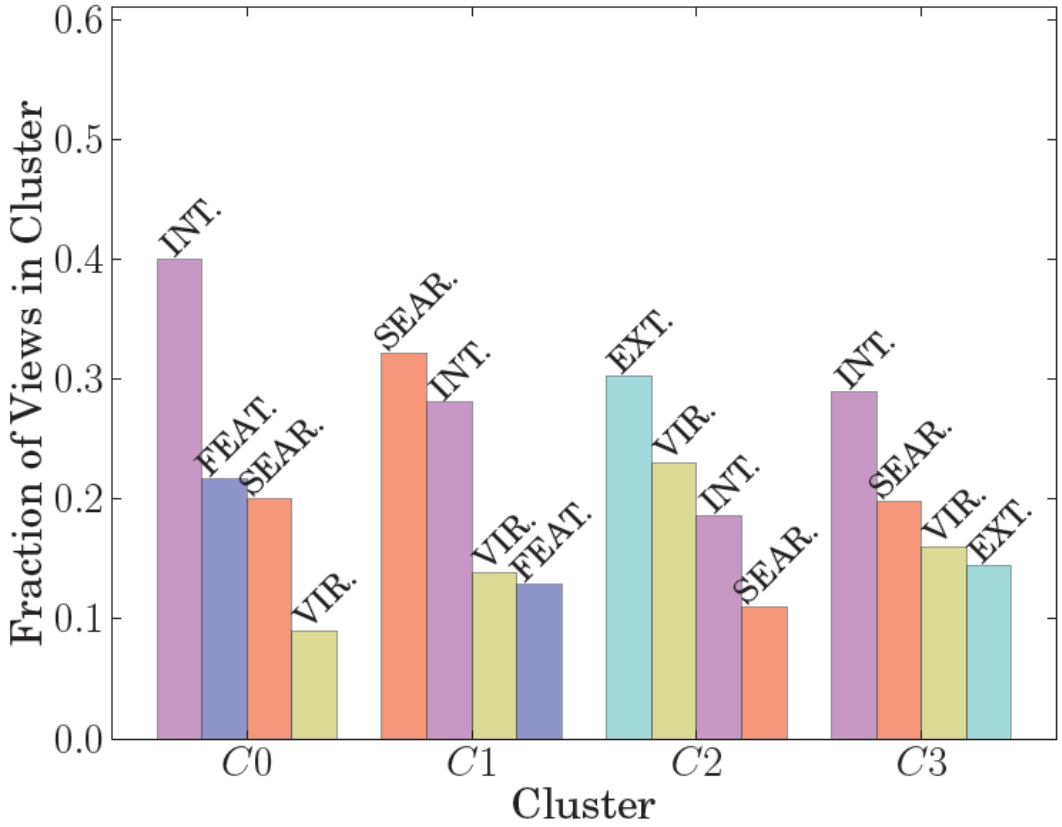
Fraction of videos per YouTube category



Distribution within each cluster differs from whole dataset (chi-squared test)

How Do Users Find This Content?

Fraction of views per type of referrer



Search is very important, but also internal browsing

Featured is important for videos in C0 and C1 (user retention)

Different distributions depending on cluster

Popularity Prediction

- Most previous work: linear regression models [Szabo2008,Pinto2013,Radinsky2012]
  - Fixed target dates
  - Fixed monitoring periods

$$\widehat{N}(t_f) = \alpha(t_f, t_r)N(t_r)$$

Popularity Prediction

- Most previous work: linear regression models [Szabo2008,Pinto2013,Radinsky2012]
  - Fixed target dates
  - Fixed monitoring periods

$$\widehat{N}(t_f) = \alpha(t_f, t_r)N(t_r)$$

Future Popularity

Popularity Prediction

- Most previous work: linear regression models [Szabo2008,Pinto2013,Radinsky2012]
  - Fixed target dates
  - Fixed monitoring periods

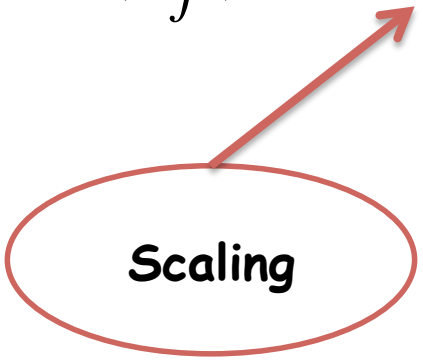
$$\widehat{N}(t_f) = \alpha(t_f, t_r) N(t_r)$$

Current Popularity

Popularity Prediction

- Most previous work: linear regression models [Szabo2008,Pinto2013,Radinsky2012]
  - Fixed target dates
  - Fixed monitoring periods

$$\widehat{N}(t_f) = \alpha(t_f, t_r)N(t_r)$$



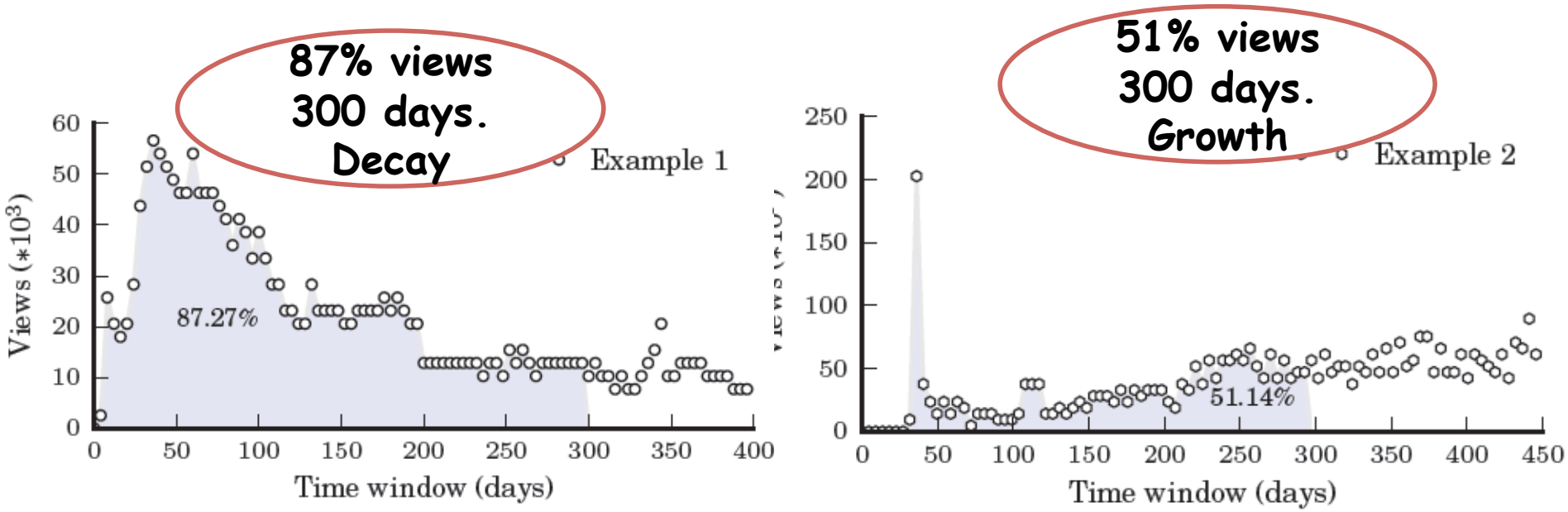
Popularity Prediction

- Most previous work: linear regression models
  [Szabo2008,Pinto2013,Radinsky2012]
  - Fixed target dates
  - Fixed monitoring periods

$$\widehat{N}(t_f) = \alpha(t_f, t_r) N(t_r)$$

- Specialized models: accuracy improvements
  [Pinto2013]
  - Predict popularity trend (our goal)

Popularity Prediction
Fixed Monitoring Periods?



Shortest monitoring period required for accurate predictions varies across videos

How Early Can We Predict the Popularity of a Video?

- Time is not the metric
  - Remaining interest: fraction of remaining views after prediction

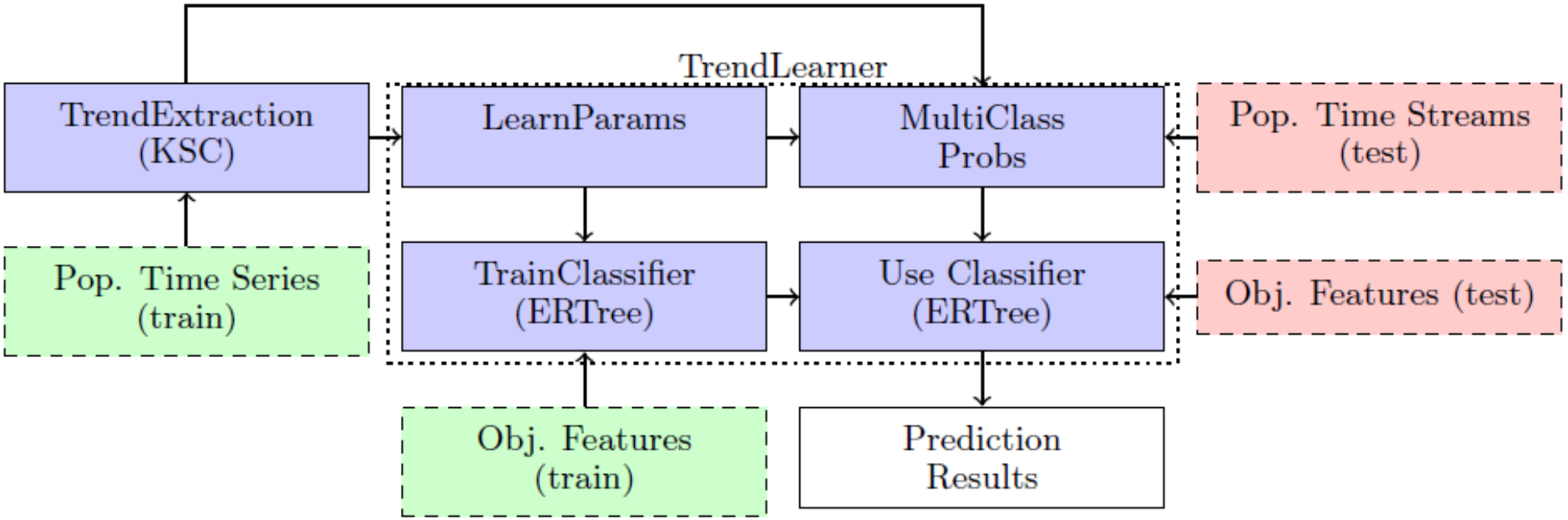
- Our solution:
  - Predict popularity trend: classification task
    - Trend = class = cluster
    - Clustering and classification algorithms
  - Prediction accuracy x remaining interest
    - Solves tradeoff on a per-video basis

Our Popularity Trend Prediction Strategy

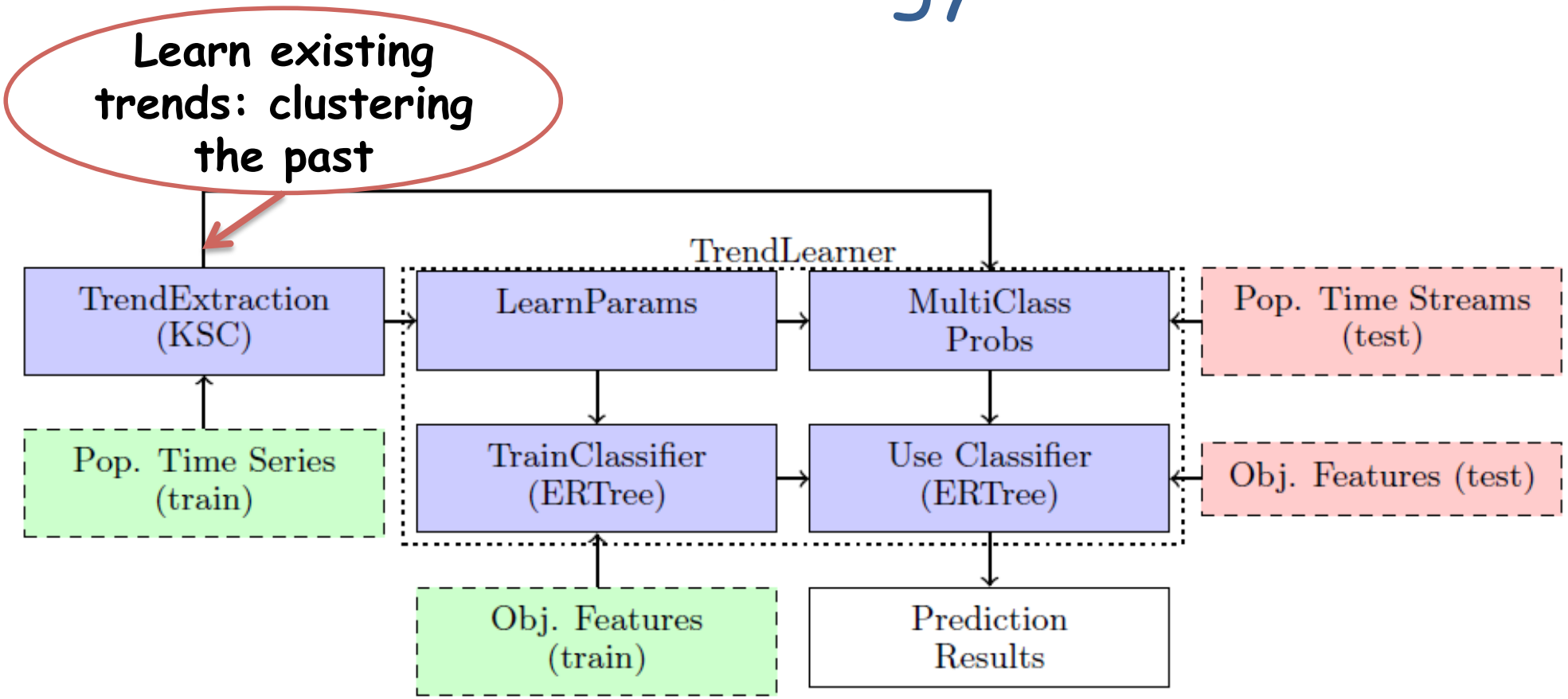
- Given a newly uploaded video:
  - Iterate over possible monitoring periods t_r:
    - Compute probability of video following each trend/class based on early popularity measures (up to t_r)
    - Take largest probability p and associated class C_i
      - If p exceeds minimum confidence of C_i, stop

- Parameters (learned from training set):
  - Minimum and maximum monitoring periods per trend/class
  - Minimum confidence per trend/class.

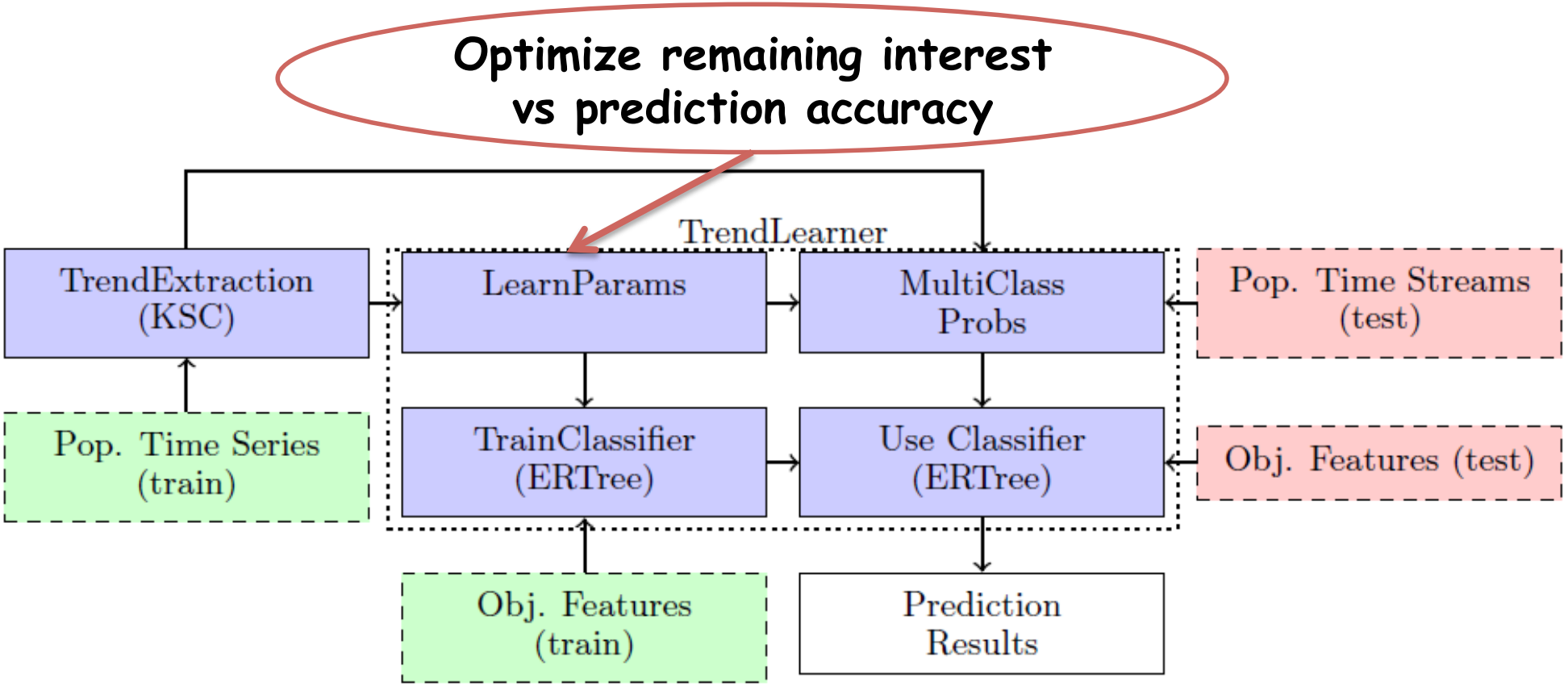
Our Popularity Trend Prediction Strategy



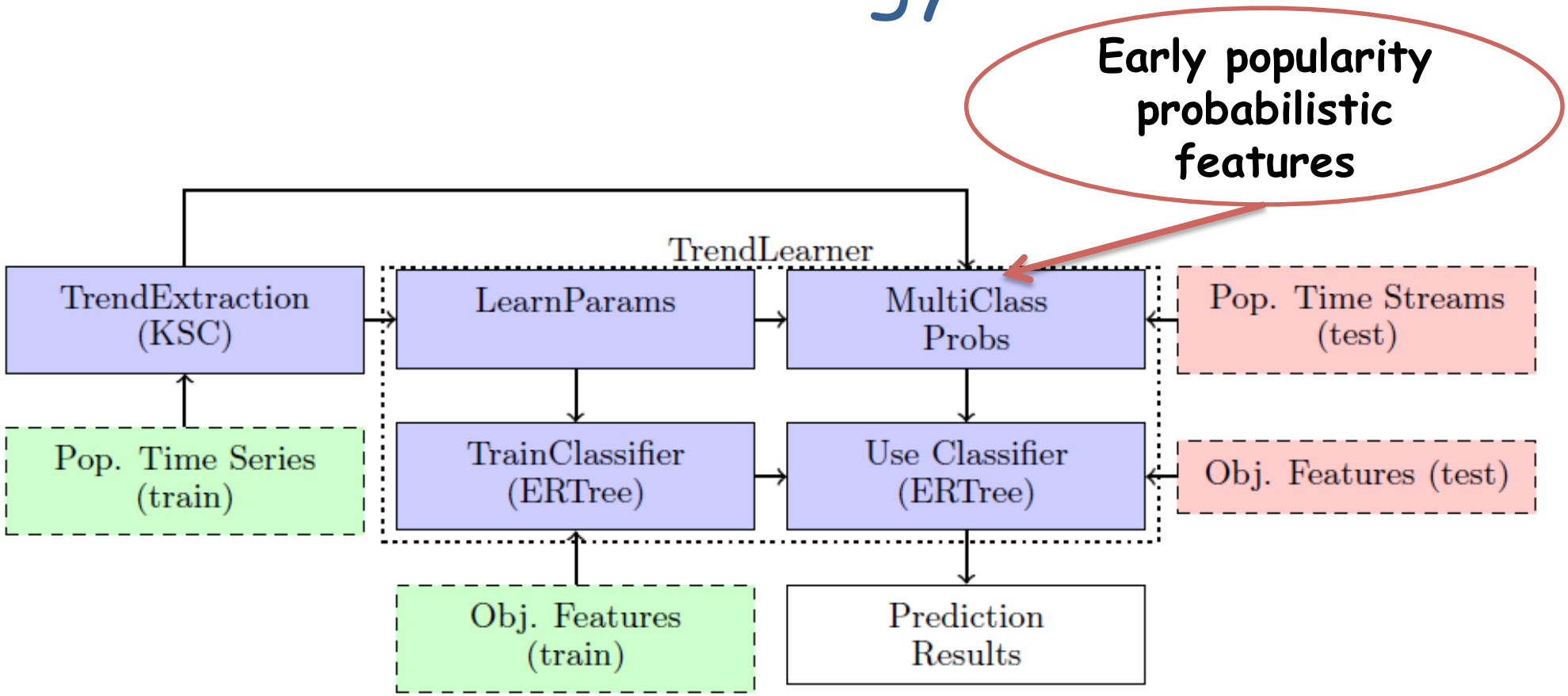
Our Popularity Trend Prediction Strategy



Our Popularity Trend Prediction Strategy

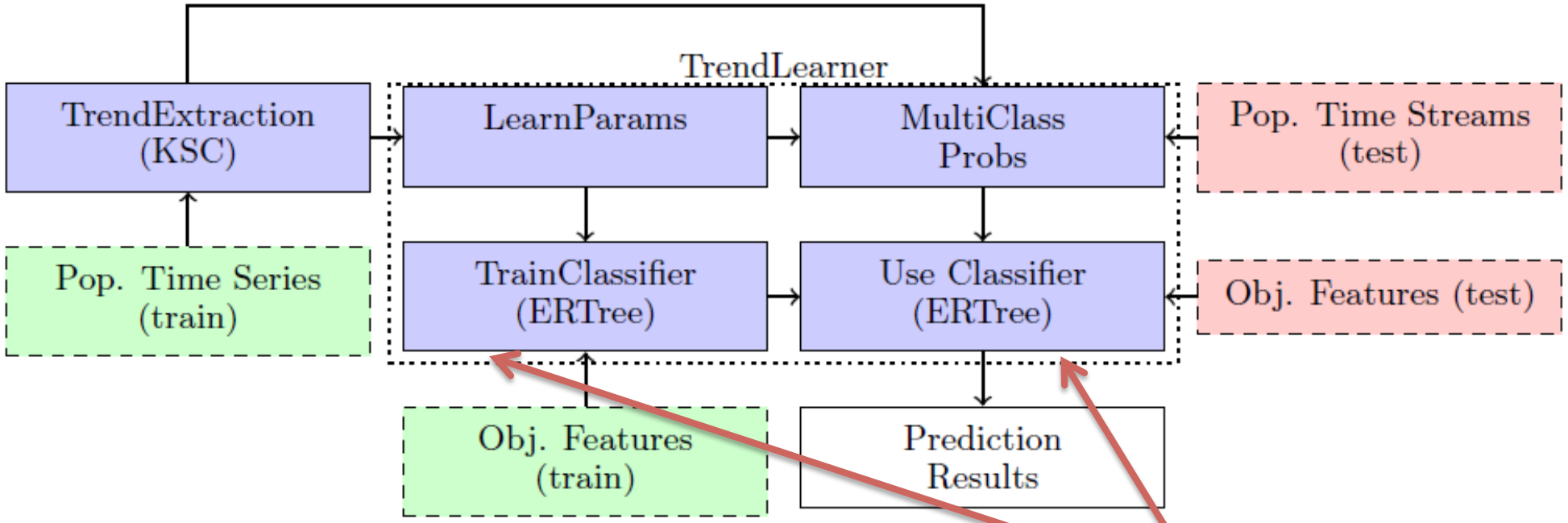


Our Popularity Trend Prediction Strategy



Prob. features: shortest monitoring time for prediction probabilities of video belonging to each class

Our Popularity Trend Prediction Strategy

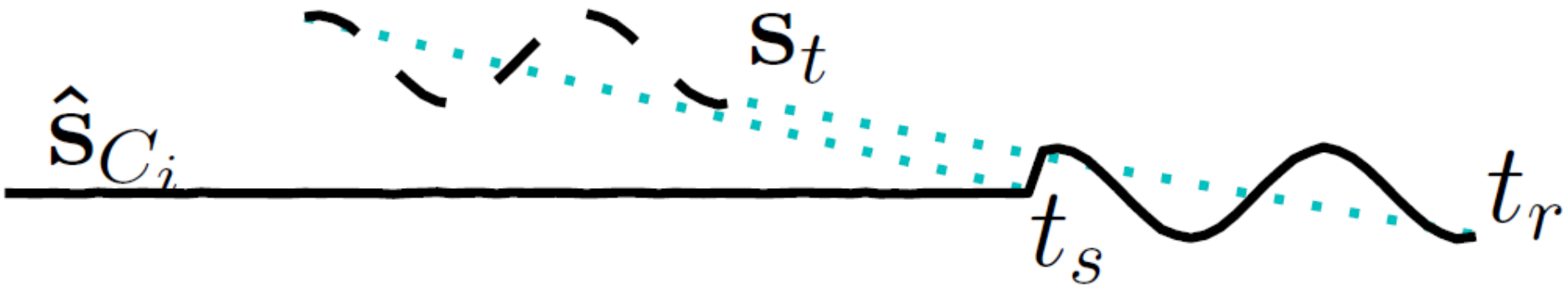


Object features: category, upload time, referrers, etc

Combine probability and object features: ERTree classifier

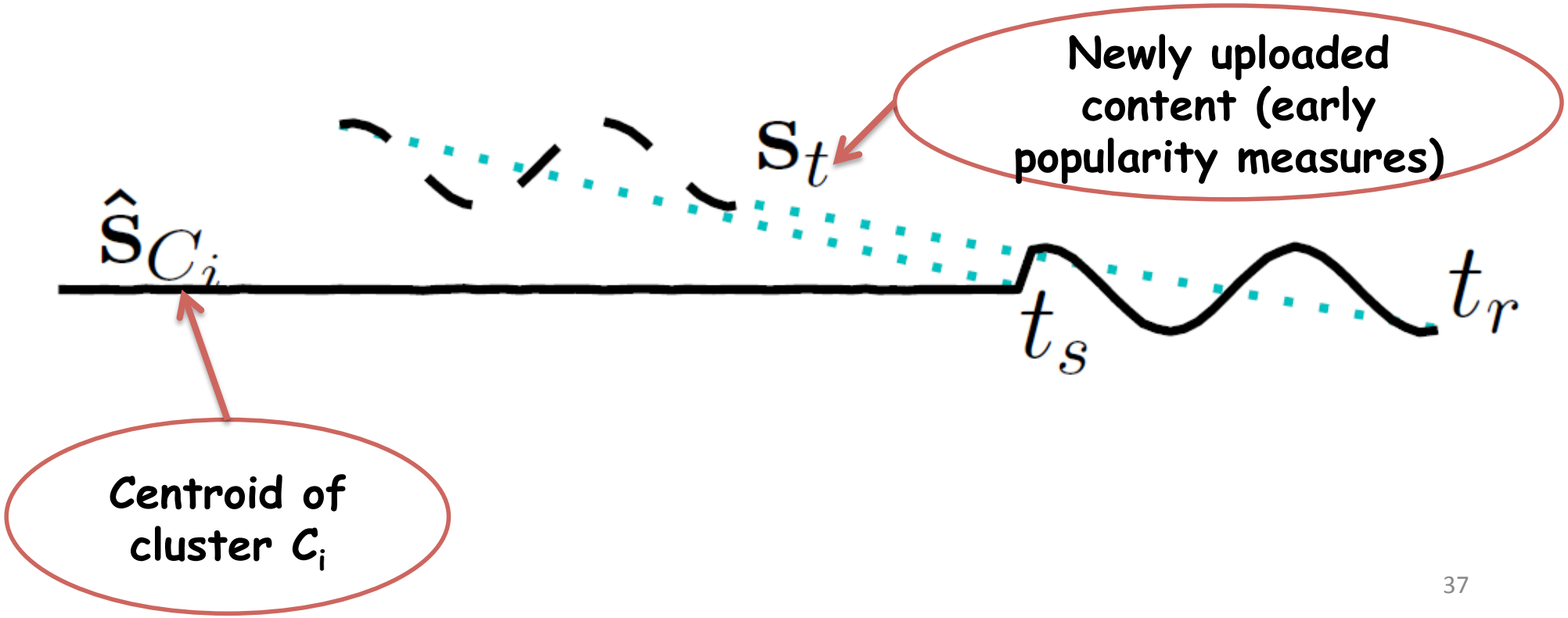
Trend Learner: Probability Features

- Probability of time series belonging to a trend/class is proportional to inverse distance to centroid of class
  - Distance metric: scale and shift invariants [Yang2011]



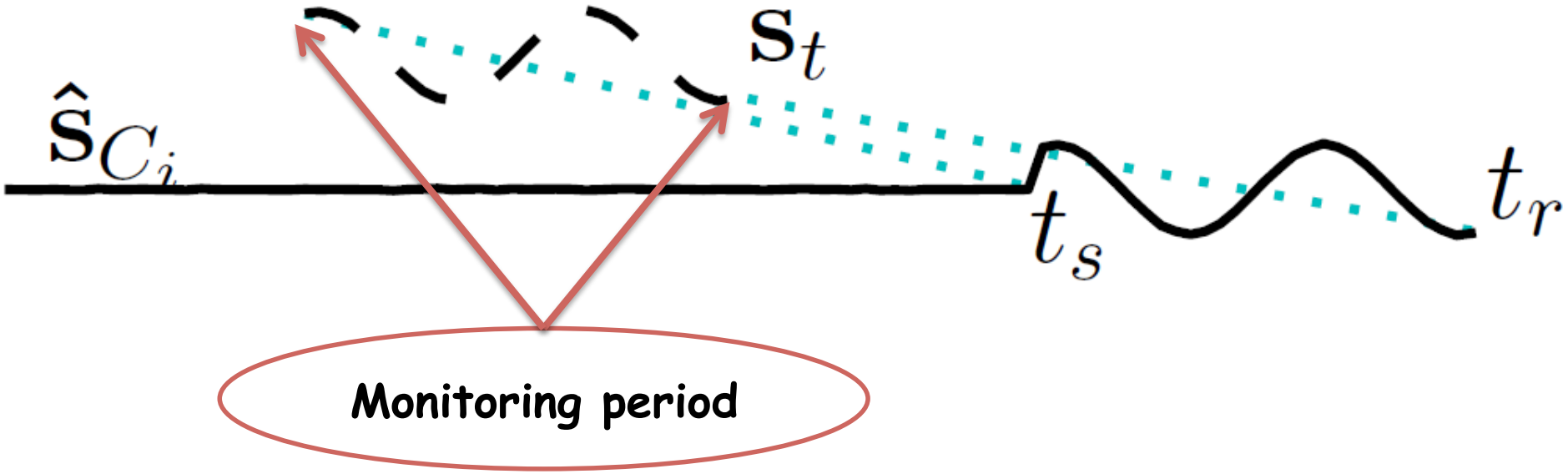
Trend Learner: Probability Features

- Probability of time series belonging to a trend/class is proportional to inverse distance to centroid of class
  - Distance metric: scale and shift invariants [Yang2011]



Trend Learner: Probability Features

- Probability of time series belonging to a trend/class is proportional to inverse distance to centroid of class
  - Distance metric: scale and shift invariants [Yang2011]



General Algorithm

1: **function** TRENDEXTRACTION(D^{train})
2: $k \leftarrow 1$
3: **while** β_{CV} is not stable **do**
4: $k \leftarrow k + 1$
5: $\mathbf{C}_D \leftarrow KSC(D^{train}, k)$
6: **end while**
7: Store centroids in $\mathbf{C}_D$
8: **end function**
9: **function** TRENDLEARNER($\mathbf{C}_D$, D^{train}, D^{test})
10: $\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{P}^{train} \leftarrow LearnParams(D^{train}, \mathbf{C}_D)$
▷ learn classifier from $\mathbf{P}^{train}$ and object features of D^{train}
11: $TrainERTrees(D^{train}, \mathbf{P}^{train} \bigcup$ object features)
12: $\mathbf{t}, \mathbf{P} \leftarrow MultiClassProbs(D^{test}, \mathbf{C}_D, \boldsymbol{\theta}, \boldsymbol{\gamma})$
13: **return** $\mathbf{t}, PredictERTrees(D^{test}, \mathbf{P} \bigcup$ object features)
14: **end function**

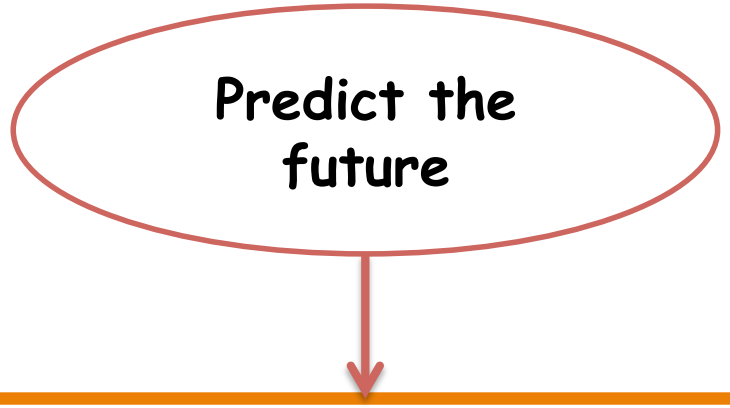
General Algorithm

1: **function** TRENDEXTRACTION(D^{train})
2: $k \leftarrow 1$
3: **while** β_{CV} is not stable **do**
4: $k \leftarrow k + 1$
5: $\mathbf{C}_D \leftarrow KSC(D^{train}, k)$
6: **end while**
7: Store centroids in $\mathbf{C}_D$
8: **end function**
9: **function** TRENDLEARNER($\mathbf{C}_D$, D^{train}, D^{test})
10: $\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{P}^{train} \leftarrow LearnParams(D^{train}, \mathbf{C}_D)$
▷ learn classifier from $\mathbf{P}^{train}$ and object features of D^{train}
11: $TrainERTrees(D^{train}, \mathbf{P}^{train} \bigcup$ object features)
12: $\mathbf{t}, \mathbf{P} \leftarrow MultiClassProbs(D^{test}, \mathbf{C}_D, \boldsymbol{\theta}, \boldsymbol{\gamma})$
13: **return** $\mathbf{t}, PredictERTrees(D^{test}, \mathbf{P} \bigcup$ object features)
14: **end function**

Learn trends by clustering the past

General Algorithm

1: **function** TRENDEXTRACTION(D^{train})
2: $k \leftarrow 1$
3: **while** β_{CV} is not stable **do**
4: $k \leftarrow k + 1$
5: $\mathbf{C}_D \leftarrow KSC(D^{train}, k)$
6: **end while**
7: Store centroids in $\mathbf{C}_D$
8: **end function**
9: **function** TRENDLEARNER($\mathbf{C}_D$, D^{train}, D^{test})
10: $\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{P}^{train} \leftarrow LearnParams(D^{train}, \mathbf{C}_D)$
▷ learn classifier from $\mathbf{P}^{train}$ and object features of D^{train}
11: $TrainERTrees(D^{train}, \mathbf{P}^{train} \bigcup$ object features)
12: $\mathbf{t}, \mathbf{P} \leftarrow MultiClassProbs(D^{test}, \mathbf{C}_D, \boldsymbol{\theta}, \boldsymbol{\gamma})$
13: **return** $\mathbf{t}, PredictERTrees(D^{test}, \mathbf{P} \bigcup$ object features)
14: **end function**



Experimental Evaluation: F1 Metric

Use probabilities directly: take largest probability

		Classifier			
		P	**P**+ERTree	ERTree	TrendLearner
Top	Micro F1	.48 ± .06	.48 ± .06	.58 ± .01	.62 ± .01
	Macro F1	.44 ± .06	.44 ± .06	.57 ± .01	.61 ± .01
Random	Micro F1	.67 ± .02	.62 ± .01	.65 ± .01	.71 ± .01
	Macro F1	.69 ± .02	.63 ± .01	.63 ± .01	.70 ± .01

Experimental Evaluation: F1 Metric

Use probabilities as features of ERTree

		Classifier			
		P	**P**+ERTree	ERTree	TrendLearner
Top	Micro F1	.48 ± .06	.48 ± .06	.58 ± .01	.62 ± .01
	Macro F1	.44 ± .06	.44 ± .06	.57 ± .01	.61 ± .01
Random	Micro F1	.67 ± .02	.62 ± .01	.65 ± .01	.71 ± .01
	Macro F1	.69 ± .02	.63 ± .01	.63 ± .01	.70 ± .01

Experimental Evaluation: F1 Metric

Use only object features

		Classifier			
		P	**P**+ERTree	ERTree	TrendLearner
Top	Micro F1	.48 ± .06	.48 ± .06	.58 ± .01	.62 ± .01
	Macro F1	.44 ± .06	.44 ± .06	.57 ± .01	.61 ± .01
Random	Micro F1	.67 ± .02	.62 ± .01	.65 ± .01	.71 ± .01
	Macro F1	.69 ± .02	.63 ± .01	.63 ± .01	.70 ± .01

Experimental Evaluation: F1 Metric

Our solution

		Classifier			
		P	**P**+ERTree	ERTree	TrendLearner
Top	Micro F1	.48 ± .06	.48 ± .06	.58 ± .01	.62 ± .01
	Macro F1	.44 ± .06	.44 ± .06	.57 ± .01	.61 ± .01
Random	Micro F1	.67 ± .02	.62 ± .01	.65 ± .01	.71 ± .01
	Macro F1	.69 ± .02	.63 ± .01	.63 ± .01	.70 ± .01

Results: Summary

- Promising results: accuracy and remaining interest
  - median of 68% of views remaining after prediction (Top)
  - median of 32% of views remaining after prediction (Random)

- Specialized models to predict popularity at future date
  1. Predict popularity trend
  2. Apply specialized regression model for predicted trend

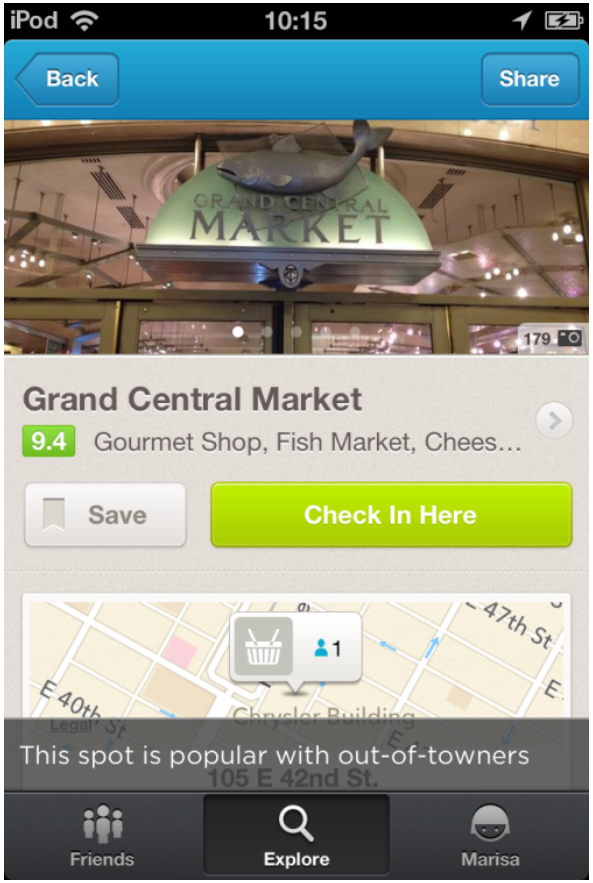
  ➢ improve prediction accuracy by 30% (median)

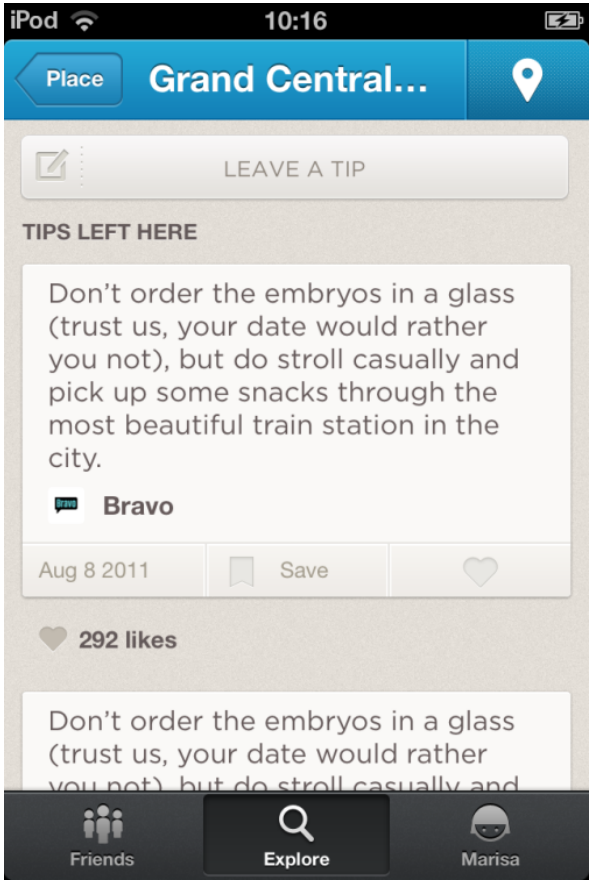
Popularity Prediction: Case Studies

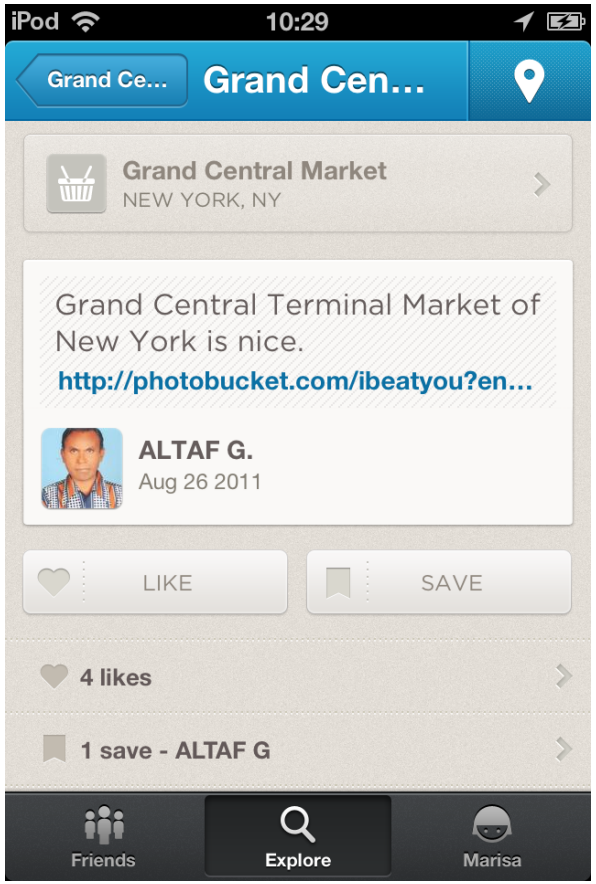
- YouTube videos

- Foursquare tips

Foursquare: Tips and Likes







Tip = Micro-review

Popularity = total number of likes

Popularity ≈ helpfulness ≈ quality

Reviews v.s. Micro-Reviews

Epinions

Canon EOS 1100D / Rebel T3 Digital Camera

11 consumer reviews

Average Rating: Excellent

5 stars 6
4 stars 5
3 stars
2 stars
1 star

Absolutely amazing camera

User Rating: Excellent

Ease of Use:
Durability:
Battery Life:
Photo Quality:
Shutter Lag

Pros: great photo quality
Cons: The LCD feels like it's blurry sometimes
The Bottom Line: A camera of high quality

Brought on Amazon.com,this camera is absolutely amazing! My boyfriend has the T4i and although that has some pretty interesting features, I can't help but to love this baby more. I went to a hockey game the other night and even though we sat pretty far away the pictures that I took with this came out excellent! I find this camera to be more lightweight than the Nikon 3100 (I used to rent out my University's camera) and easier to use as well. But that's just my opinion! I'm a Canon girl through and through.

Also, comparing it still to the 3100, I find the T3 has sharper, brighter colors especially when you toy around with the "Creative Auto" setting and change up some of the settings there. I believe the grip for the 3100 was rubberized and I will admit that it does take a little adjusting to for the non-rubberized grip, but now after using it for almost two months it's just something natural to me and whenever I use my boyfriend's T4i, the rubberized grip just feels weird to me now! The battery life, for me and how I use it, is excellent! I even left the switch "on" for a whole weekend by accident after taking pictures all day, and picked it right back up and continued to shoot more pictures. I've only had to charge it two or three times so far, and I tend to take at least 20 to 30 pictures a day at the very least. Some days I'll take up to 300 or more!

This camera was the best purchase I've made, and I am not disappointed one bit by the performance, the body, and the picture quality!

foursquare

Balthazar Bakery

12 tips

Sort: Popular / Recent

Rachel May 12, 2010
What a hidden treasure — this is the WHOLESALE location of the bakery! You can even get a peek into their operation. Amazing.
Save · Like · 7 likes · Flag

Philip L. December 10, 2012
Is this heaven?
Save · Like · 2 likes

Christine G. March 9, 2013
The goat cheese tart - a must try!!!
Save · Like · 1 like

Vin L. August 4, 2012
Fantastic bakery
Save · Like · 1 like

IC C. April 22, 2011
Monkey bread!!'
Save · Like · 1 like

Christopher H. April 1, 2011
Whole wheat croissant! So GOOD!
Save · Like · 1 like

Oz A. November 23, 2012
Best Chocolate Croissants ever!
Save · Like

Eujin M. April 28, 2012
the bagged chocolate chip cookies are addictive! olive rolls are delicious.
Save · Like

Tip Popularity Prediction Tasks

- Predict the popularity of a tip at a given future date
- Predict the *popularity level* of a tip at a given future date
- Rank tips based on predicted popularity at future date

Tip Popularity Prediction Tasks

- Predict the popularity of a tip at a given future date
- Predict the *popularity level* of a tip at a given future date
  - Classification task
- Rank tips based on predicted popularity at future date

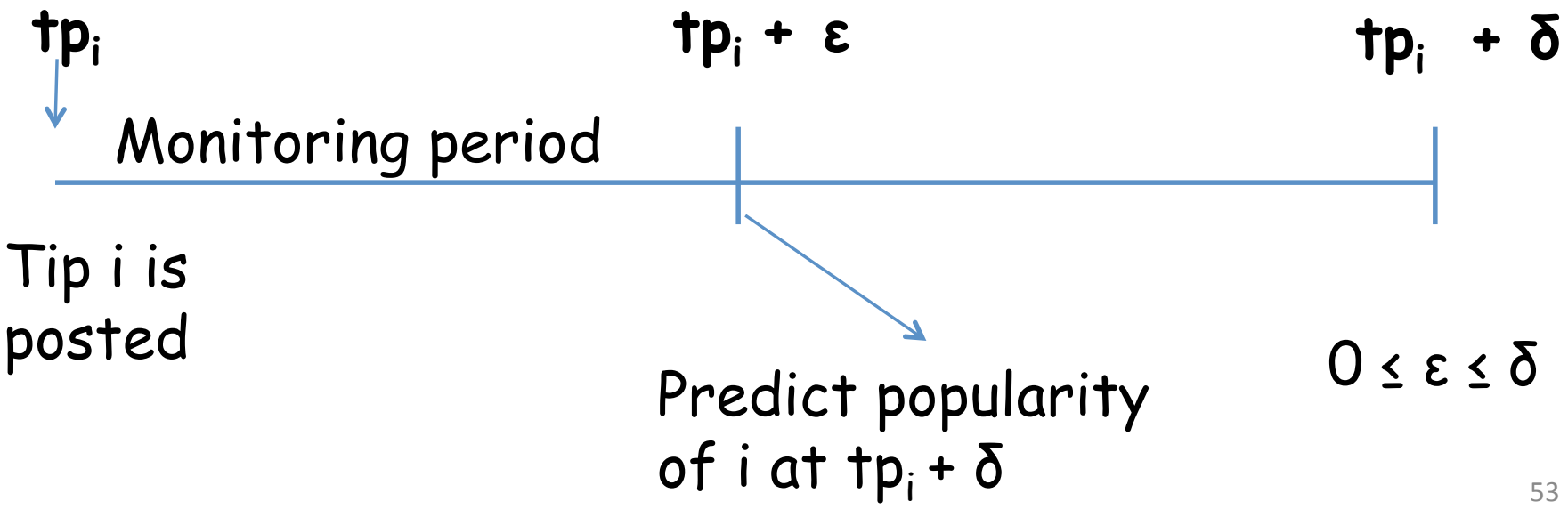
Popularity Levels

Popularity level	# of likes/tip	# tips
Low	< 5	703,827
High	≥ 5	3,427

- High class imbalance: severe impact on prediction!!!
- Similar for other definitions of popularity levels

Problem Statement

- A tip is a tuple (p,u,v) where
  - p: features extracted from the tip's content
  - u: features associated with the tip's author
  - v: features of venue where tip was posted



Prediction Algorithms

- Support Vector Machine (SVM) classifier
  - Linear e RBF kernels
- Regression
  - Support Vector Regression (SVR)
    - Linear e RBF kernels
  - Simpler linear regression (OLS)
- Median number of likes of tips previously posted by the user (baseline)

Features: Tip's Author

- Number of tips
- Number of likes received
- Number of likes given
- Number of distinct venues
- Number of mayorships
- Is user mayor of venue?

Activities in the system

- Number of friends/followers
- Number of likes from SN
- Number of tips posted by SN
- Number of likes given by SN

Social network

- Visibility of user in venue
- Type of user

Others

Features - Venue

- Number of tips
- Number of likes
- Number of checkins
- Number of unique visitors

Activities

- Is venue verified?
- Venue category

Characteristics

- Position of tip in ranking by # likes
- Position in ranking by posting time

Others

Features - Tip

- Number of characters
- Number of words
- Number of urls or emails

Amount of content

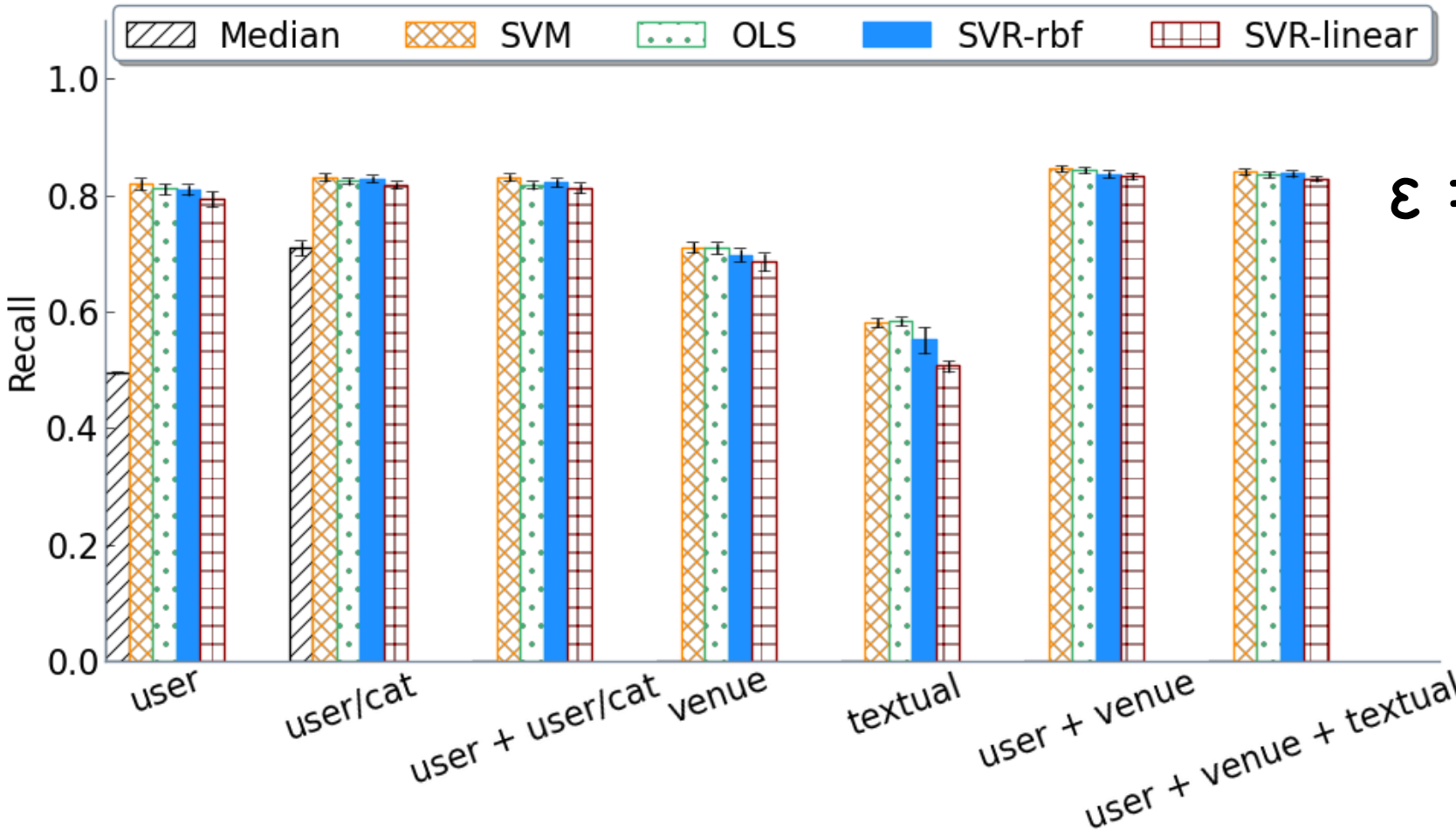
- % nouns
- % adjectives
- % adverbs
- % verbs
- % punctuation marks

Part-of-speech tags

- Positive scores (average)
- Negative scores (average)
- Neutral scores (average)

Polarity/Sentiment

Experimental Evaluation: Macro-Recall

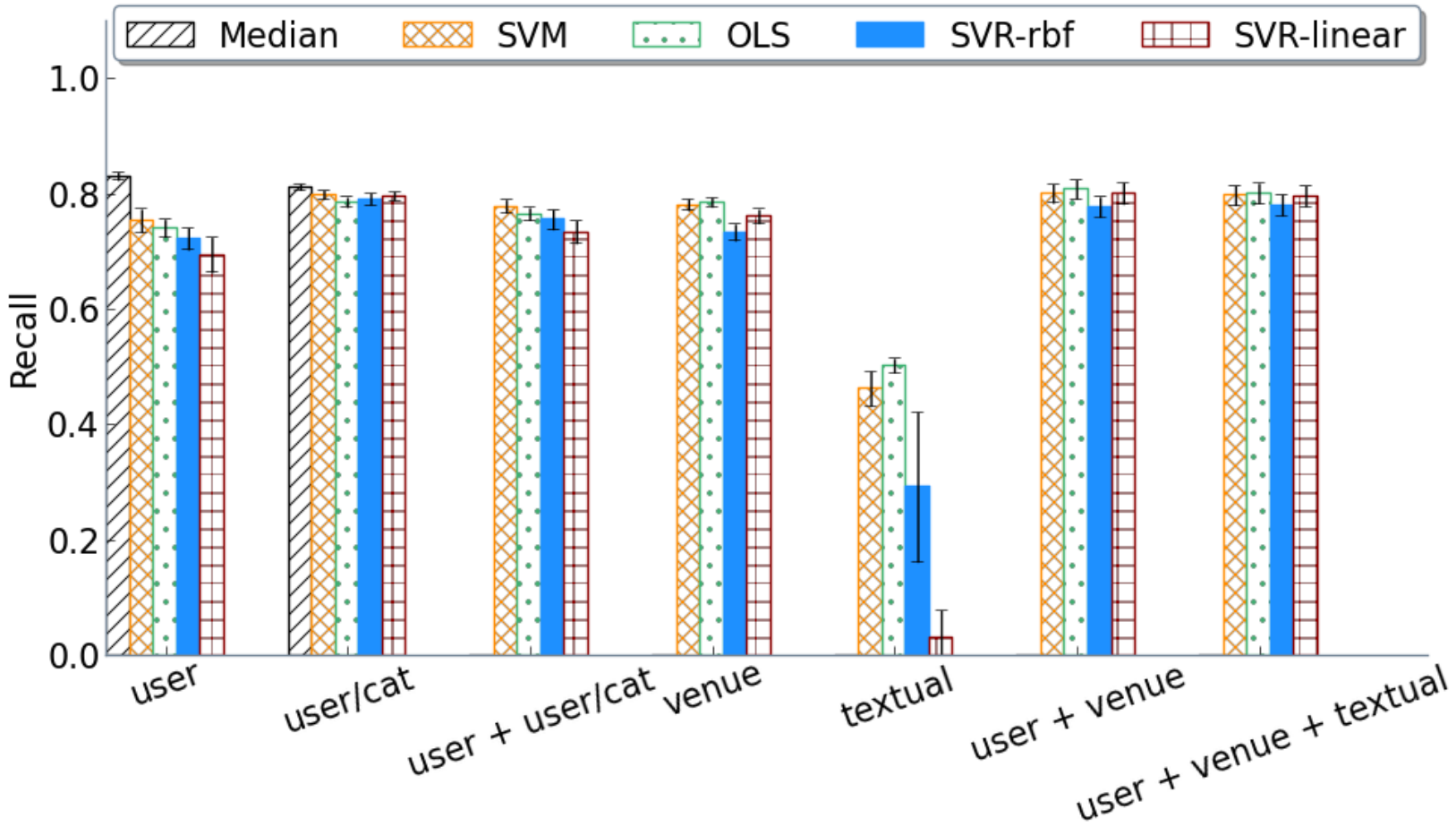


ε = 0

Best approach: combine user + venue features

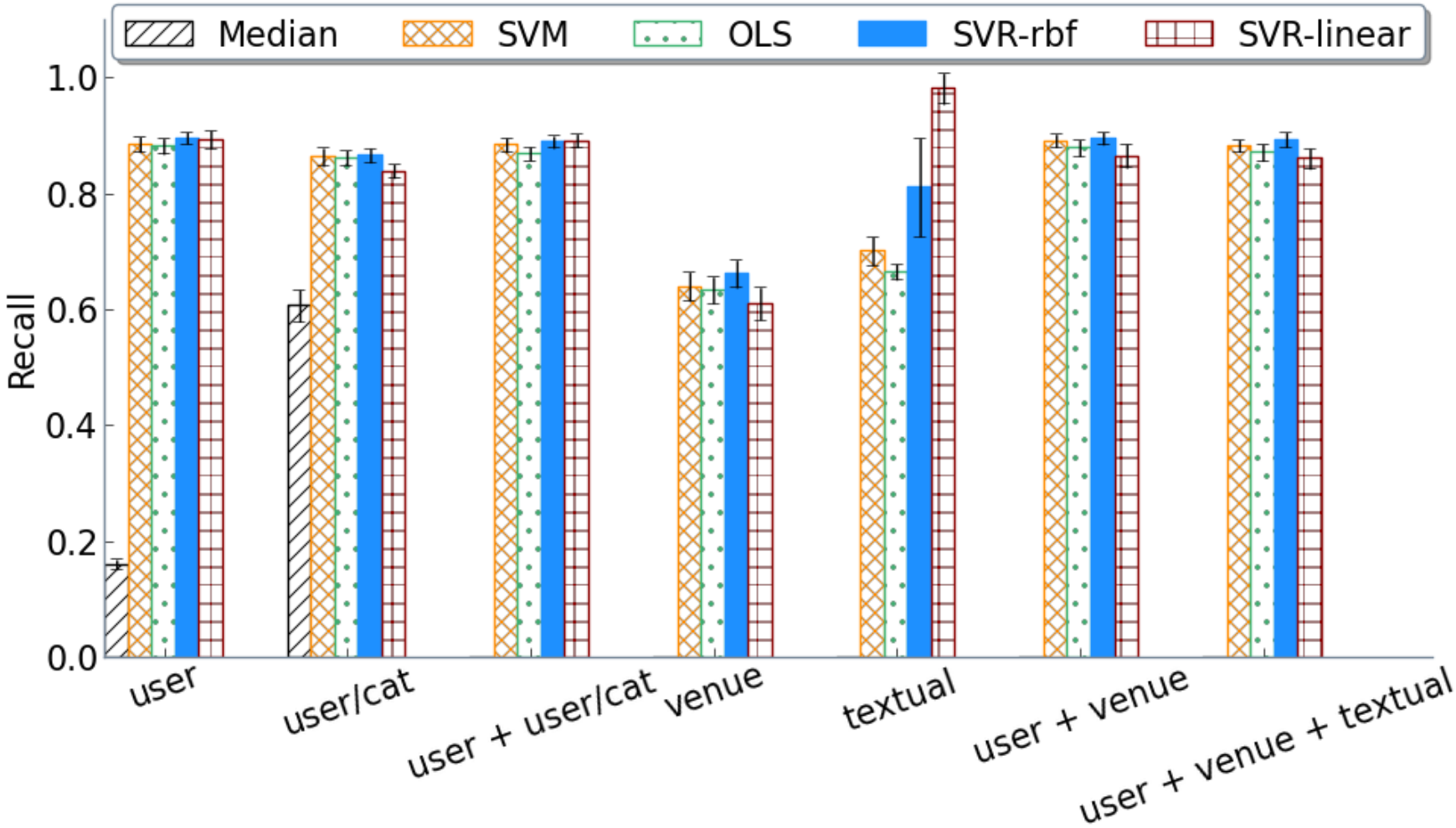
OLS, SVM and SVR: similar results (but OLS is simpler)

Experimental Evaluation: Recall of Low Popularity



Median of likes: best results (class imbalance)

Experimental Evaluation: Recall of High Popularity



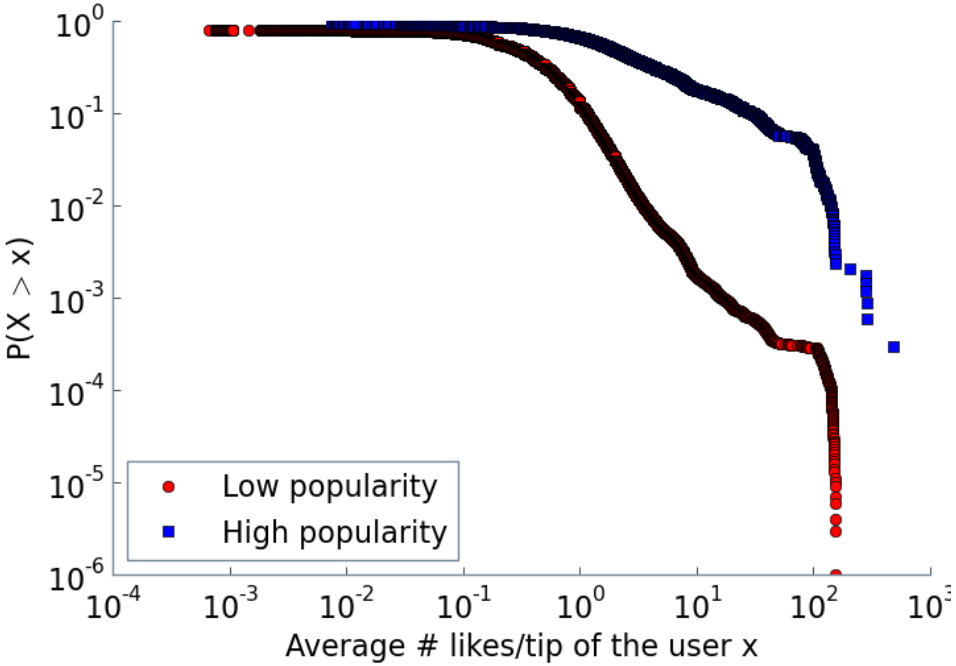
SVR (RBF): slightly better than OLS and SVM

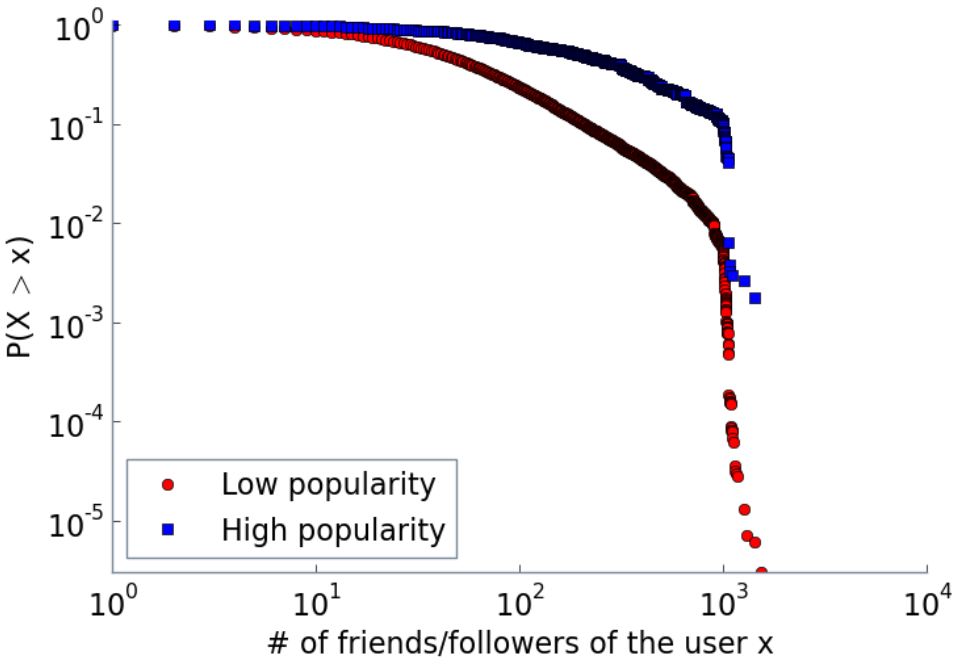
Results: Summary

- Best results:
  - OLS: similar to SVM and SVR, but simpler
  - combination of user and venue features as inputs

- Ranking of features by importance (Information Gain)
  - Top-3 related to user: number of likes in previous tips
  - 4th: size of social network of user
  - 6th-7th: popularity of venue (# visitors, # checkins)

Important User Features





Average # likes

friends/followers

Summary

- Popularity prediction of UGC: challenging task
  - Multitude of external and internal factors
  - Inherent characteristics of application
  - Highly skewed popularity distribution: severe imbalance impacts efficacy of regression/classification
- Current work:
  - YouTube: popularity trends and measures
  - Foursquare: popularity measures, levels and ranking
- Next steps: new features
  - YouTube: user and social network, content
  - Foursquare: geographical aspects

Foursquare Dataset

- Almost 7 million tips
- 5,7 million likes
- 1,8 million users
- 3,2 million venues