

Extreme Classification

A New Paradigm for
Ranking & Recommendation

Manik Varma
Microsoft Research



Classification

Pick one

 Label 1
Label 2

Binary

Pick one

Label 1
Label 2
 Label 3
...
Label L

Multi-class

Pick all that apply

 Label 1
Label 2
 Label 3
...
Label L

Multi-label

Extreme Multi-label Learning

- Learning with millions of labels

Predict the set of monetizable Bing queries that might lead to a click on this ad



Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: [GO](#)

Get a quote by phone: [1-800-841-1588](#)

Find a local agent: [GO](#)

Need other insurance?

ATV	Homeowners	Overseas
Boat	ID Theft Protection	Renters
Commercial Auto	Life	RV
Condo/Ca-op	Mobile Home	Umbrella
Flood	Motorcycle	

Already a policyholder? [Log in.](#)

GEICO

© 1996-2012 GEICO

geico auto insurance

geico car insurance

geico insurance

www geico com

care geicos

geico com

need cheap auto insurance

wisconsin cheap car insurance quotes

cheap auto insurance florida

all state car insurance coupon code

Research Problems

- Defining millions of labels
- Obtaining good quality training data
- Training using limited resources
- Log time and log space prediction
- Obtaining discriminative features at scale
- Performance evaluation
- Dealing with tail labels and label correlations
- Dealing with missing and noisy labels
- Statistical guarantees
- Applications

Extreme Multi-label Learning - People

- Which people are present in this selfie?



Extreme Multi-label Learning – Wikipedia

Jeannette Wing - Wikipedia, the free encyclopedia - Windows Internet Explorer

http://en.wikipedia.org/wiki Jeannette Wing - Wikipedia...

NIPS FutureConferences Suggested Sites Web Slice Gallery

the conference will be held on 9th and 10th January 2015 at Jadavpur University
Registration for the conference is now open

Jeannette Wing

From Wikipedia, the free encyclopedia

Jeannette Marie Wing is Corporate Vice President of [Microsoft Research](#) with oversight of its core research laboratories around the world and Microsoft Research Connections.^{[2][3]} Prior to 2013, she was the President's Professor of [Computer Science](#) at [Carnegie Mellon University](#), Pittsburgh, Pennsylvania, United States. She also served as assistant director for Computer and Information Science and Engineering at the [NSF](#) from 2007 to 2010.^{[4][5][6][7][8][9][10][11][12][13]}



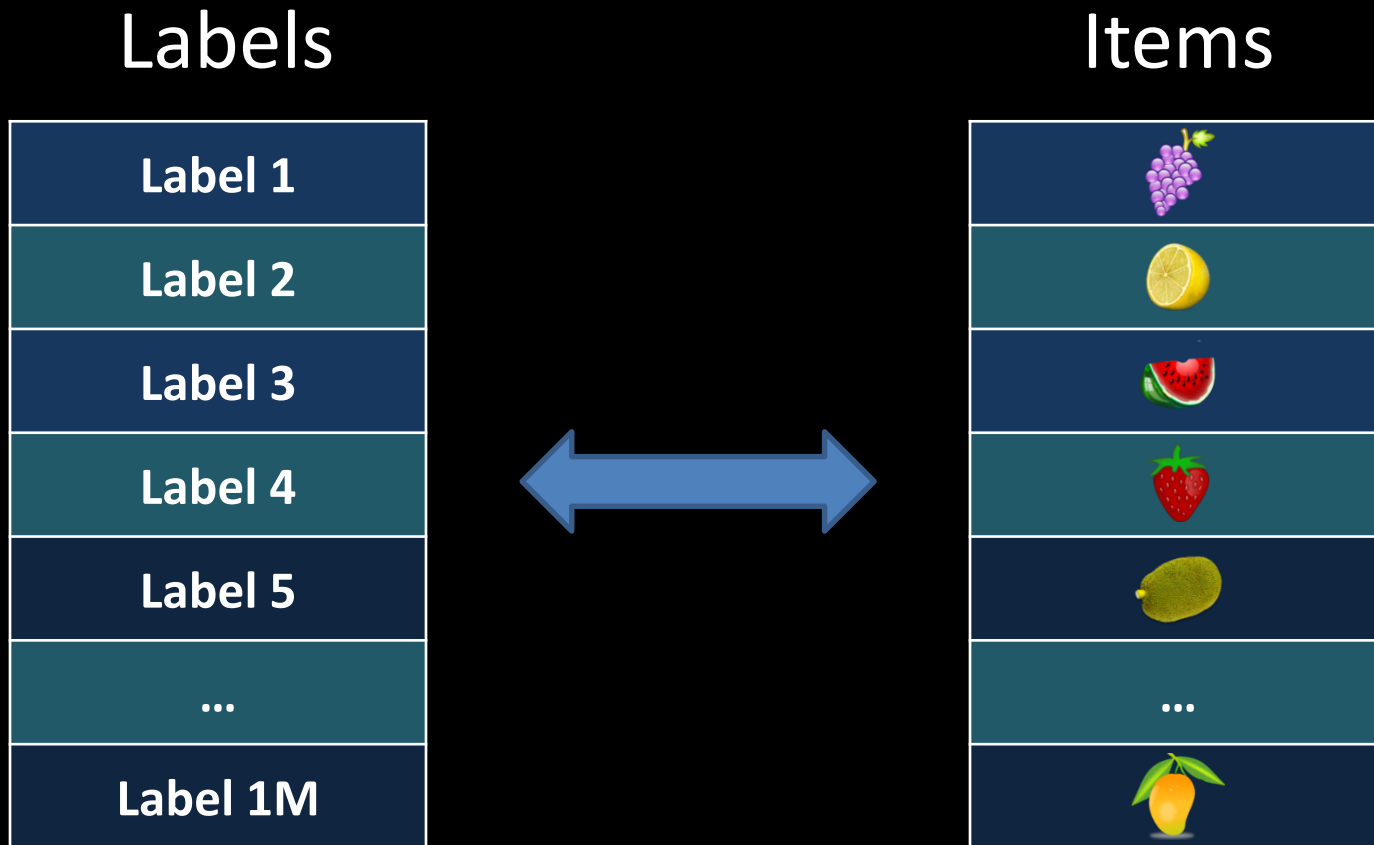
Jeannette Wing

Contents [hide]

Labels: Living people, American computer scientists, Formal methods people, Carnegie Mellon University faculty, Massachusetts Institute of Technology alumni, Academic journal editors, Women in technology, Women computer scientists.

Reformulating ML Problems

- Ranking or recommending millions of items





FastXML

A Fast, Accurate & Stable
Tree-classifier for eXtreme
Multi-label Learning

Yashoteja Prabhu (IIT Delhi)

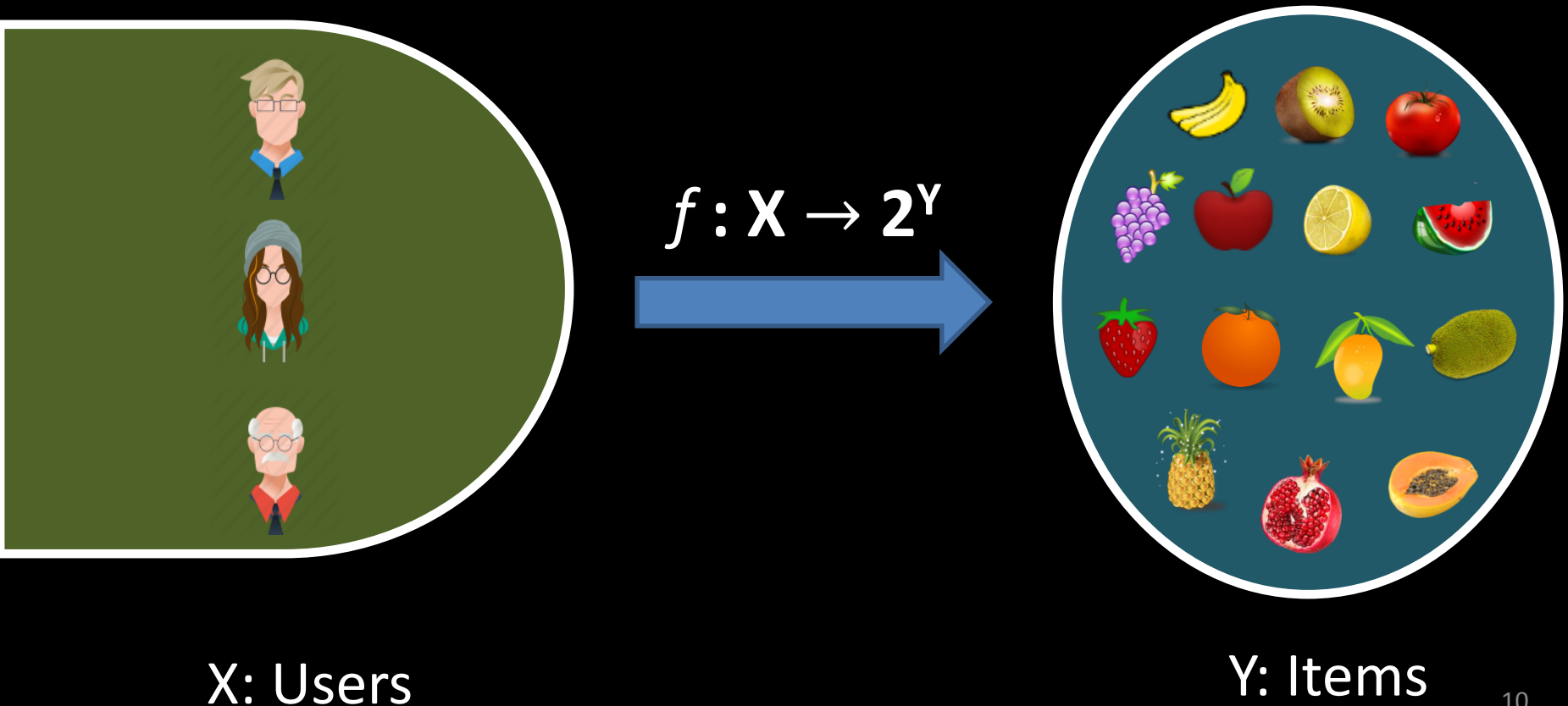
Manik Varma (Microsoft Research)

FastXML

- Logarithmic time prediction in milliseconds
 - Ensemble of balanced tree classifiers
- Accuracy gains upto 25% over competing methods
 - Nodes partitioned using nDCG
- Upto 1000x faster training over the state-of-the-art
 - Alternating minimization based optimization
 - Proof of convergence to a stationary point

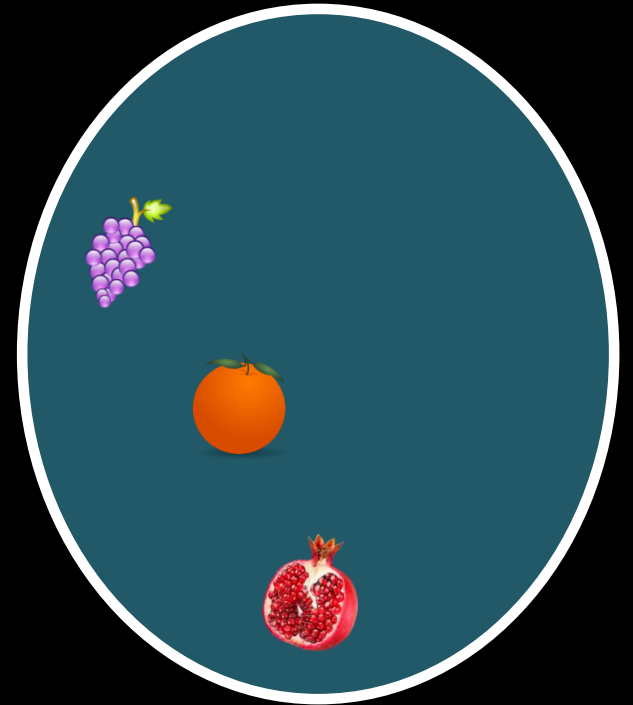
Extreme Multi-Label Learning

- Problem formulation



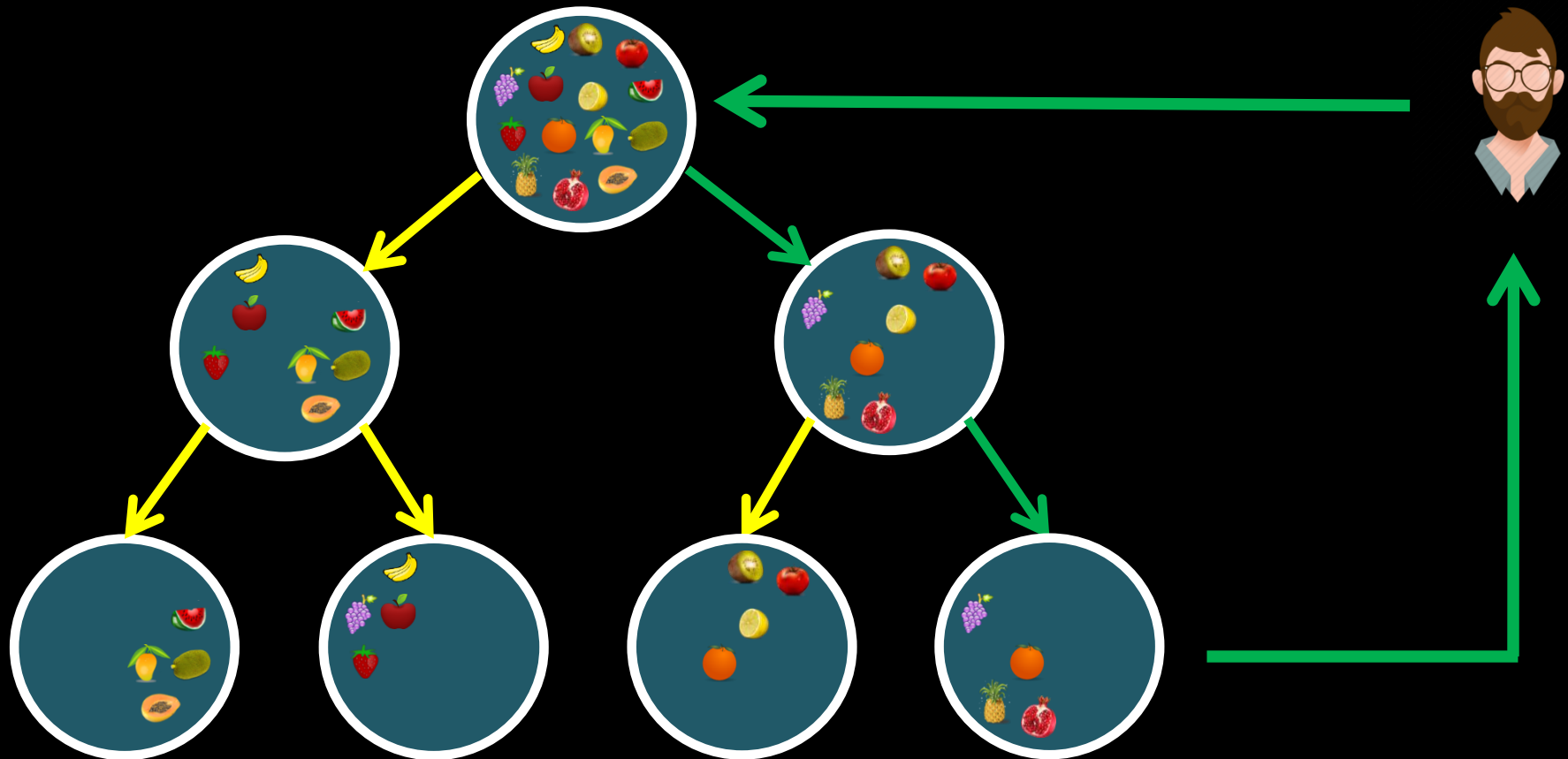
Extreme Multi-Label Learning

- Problem formulation



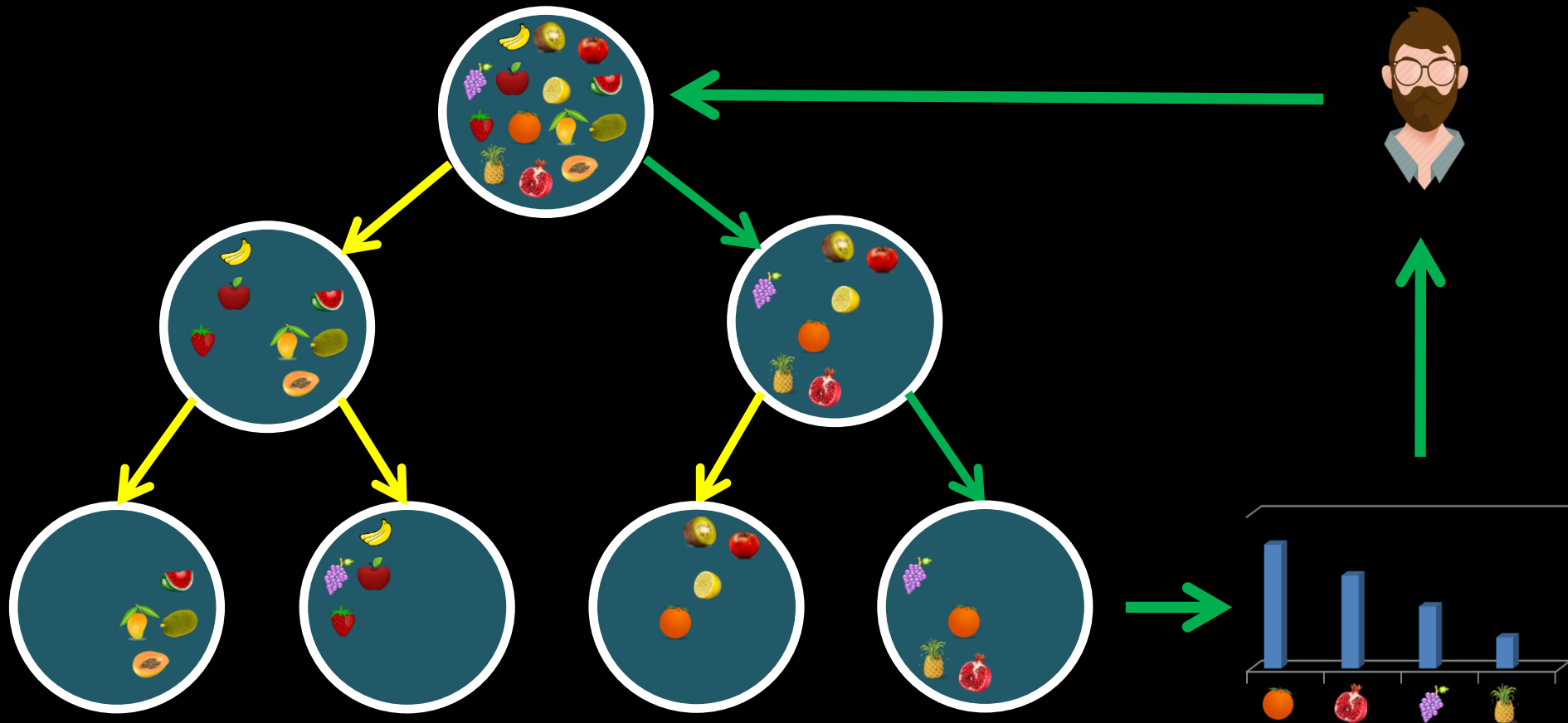
Tree Based Extreme Classification

- Prediction in logarithmic time

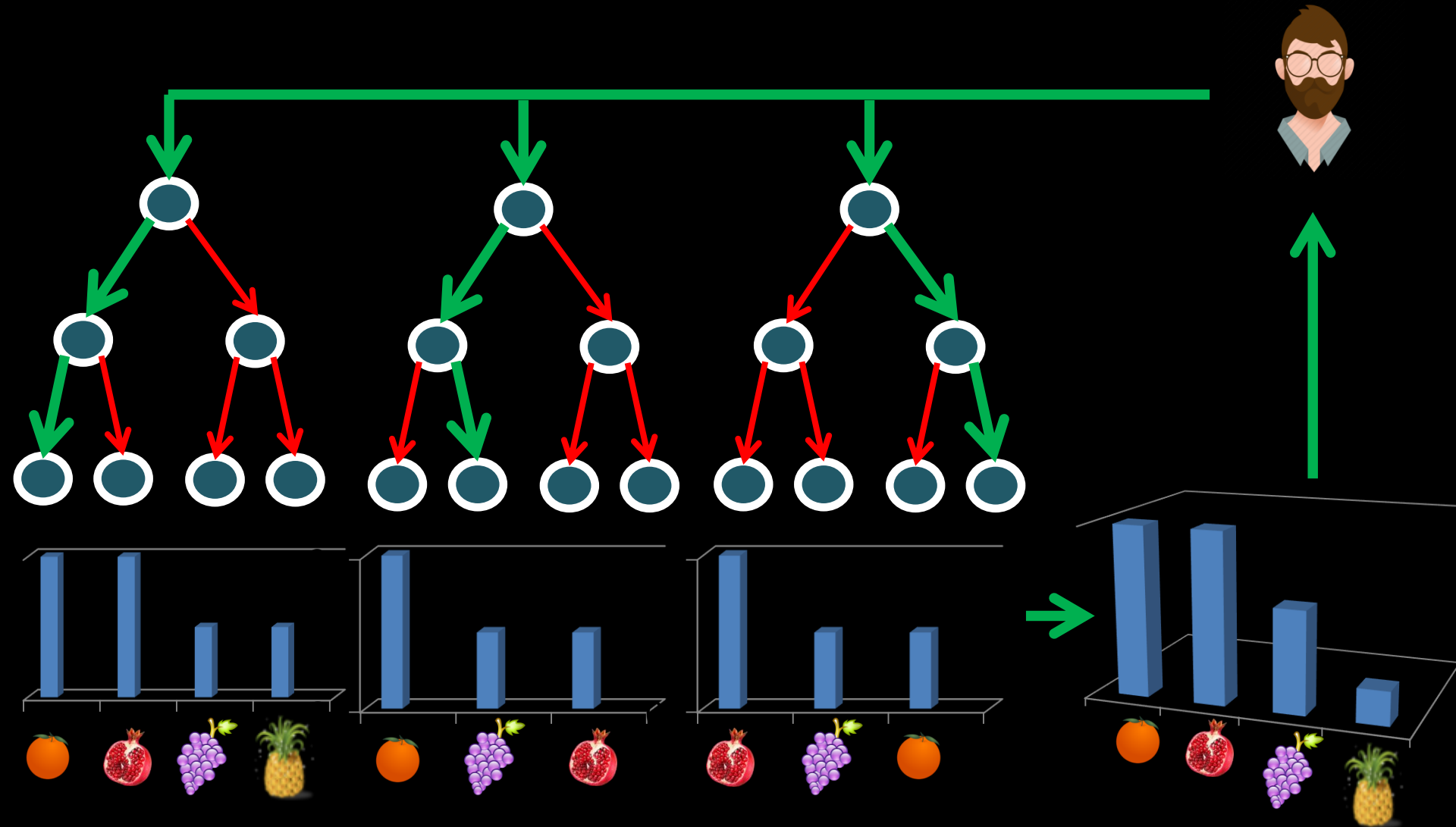


Tree Based Extreme Classification

- Prediction in logarithmic time



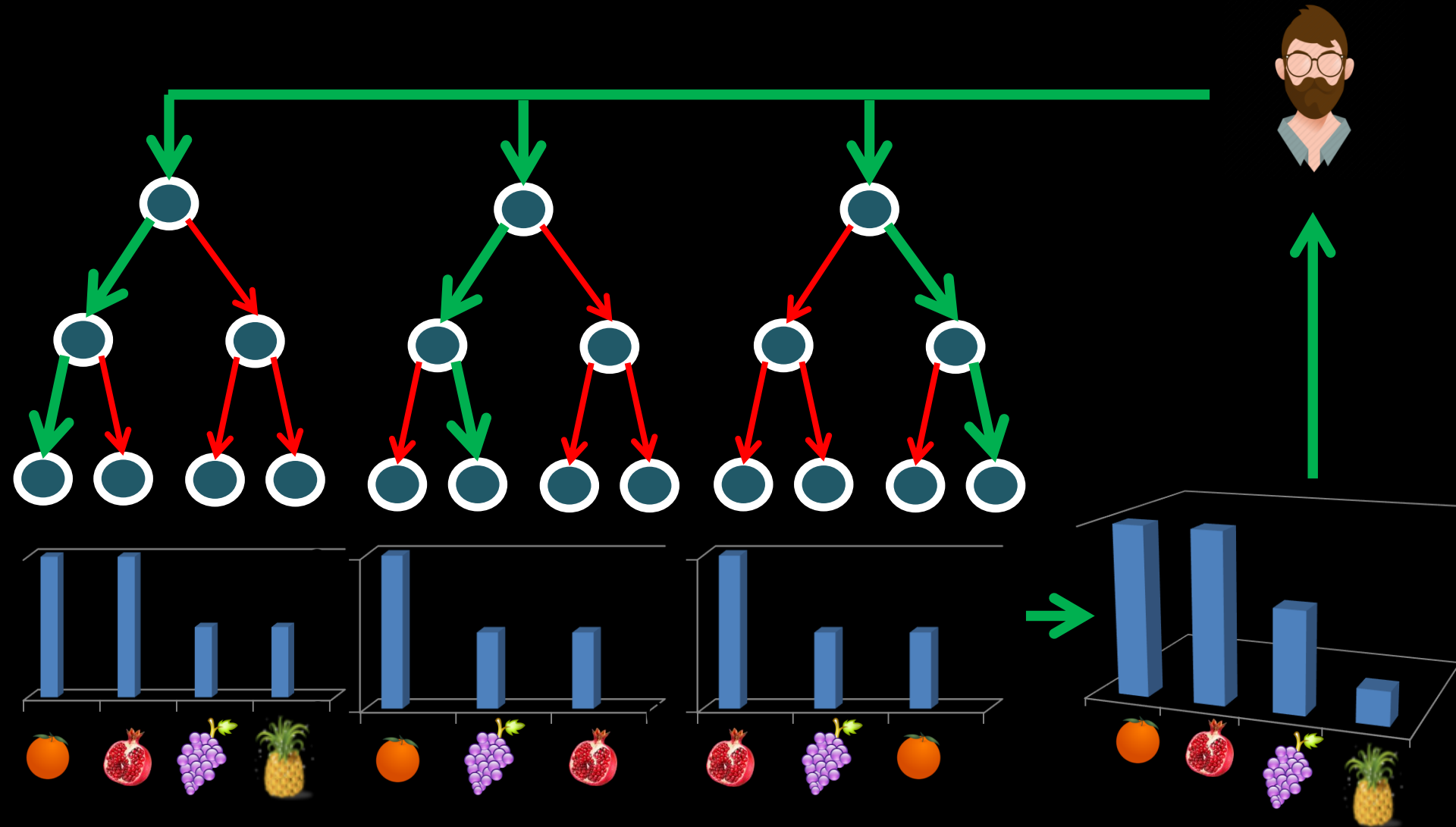
FastXML Architecture



FastXML

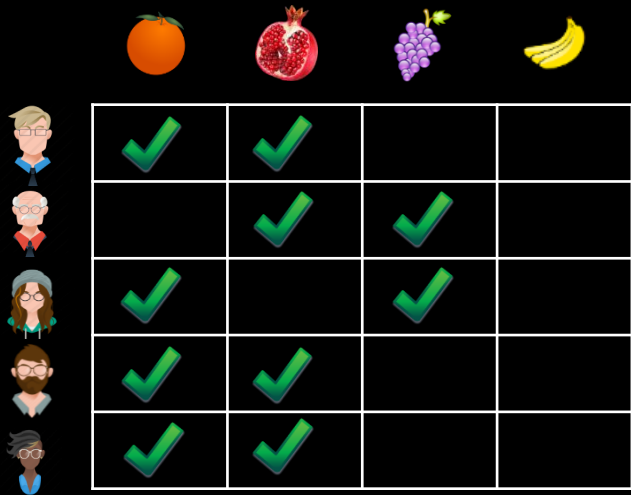
- Logarithmic time prediction in milliseconds
 - Ensemble of balanced tree classifiers
- Accuracy gains upto 25% over competing methods
 - Nodes partitioned using nDCG
- Upto 1000x faster training over the state-of-the-art
 - Alternating minimization based optimization
 - Proof of convergence to a stationary point

FastXML Architecture



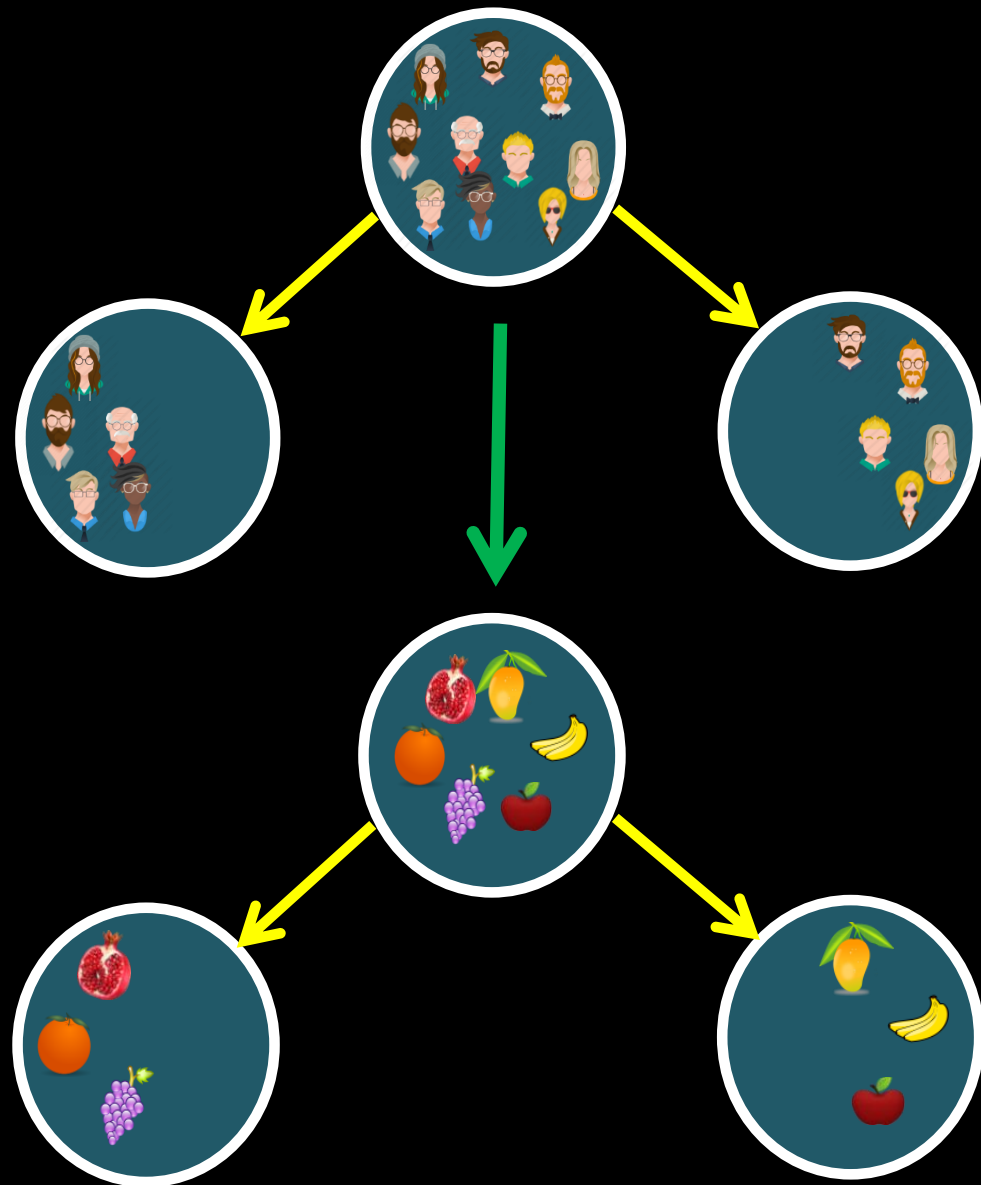
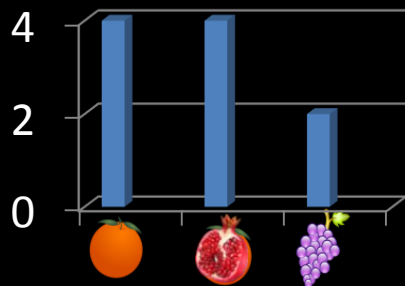
Learning to Partition a Node

Training data



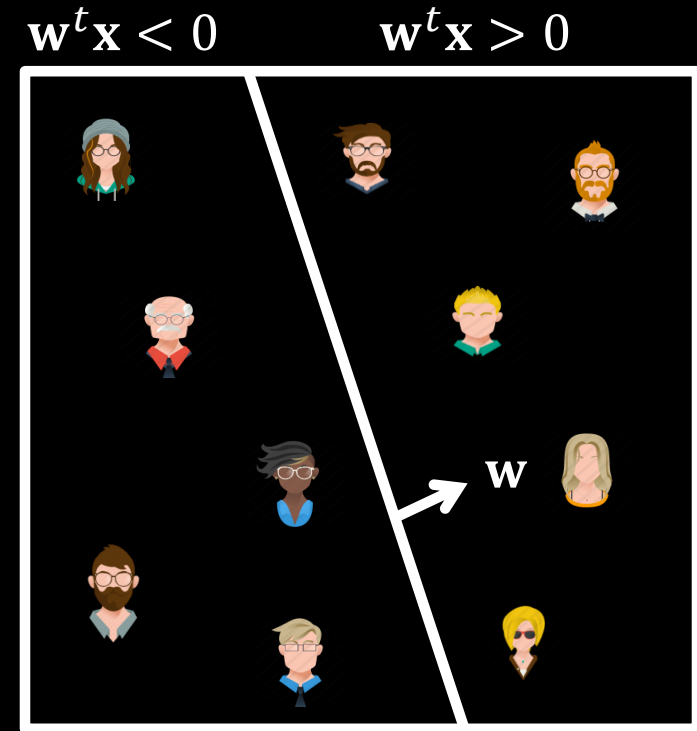
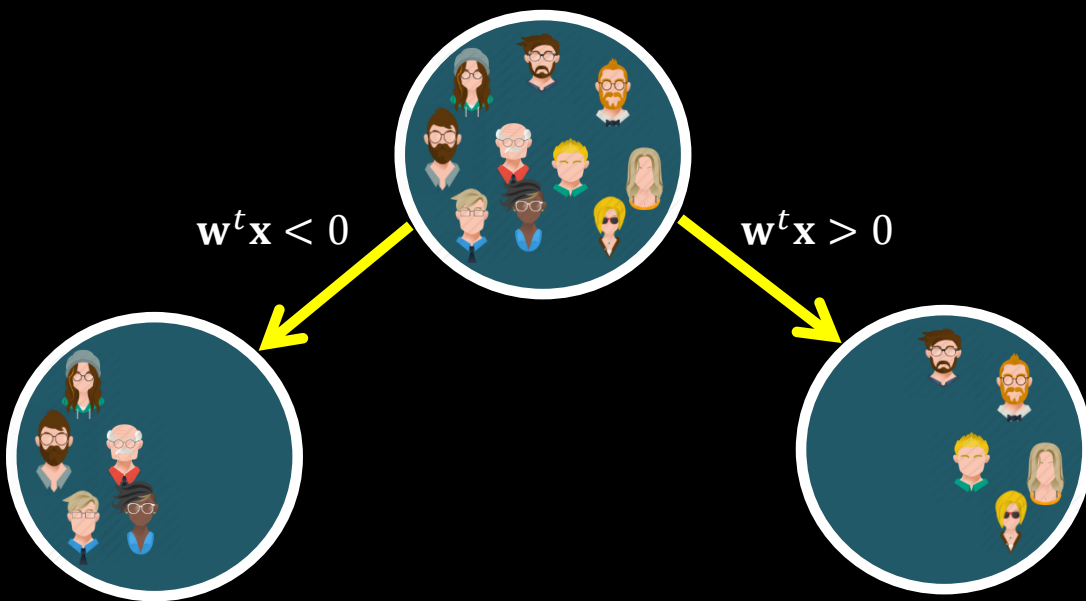
Training data table showing fruit preferences for five people (rows) across four fruit categories (columns): Orange, Pomegranate, Grapes, and Banana. Green checkmarks indicate preference.

	Orange	Pomegranate	Grapes	Banana
Person 1	✓	✓		
Person 2		✓	✓	
Person 3	✓		✓	
Person 4	✓	✓		
Person 5	✓	✓		



Learning to Partition a Node

$$\text{Min}_w \quad \|w\|_1 - C \sum_{i \in \text{Users}} \text{nDCG}(x_i, y_i, w)$$



X : Space of Users

FastXML

- Logarithmic time prediction in milliseconds
 - Ensemble of balanced tree classifiers
- Accuracy gains upto 25% over competing methods
 - Nodes partitioned using nDCG
- Upto 1000x faster training over the state-of-the-art
 - Alternating minimization based optimization
 - Proof of convergence to a stationary point

Optimizing nDCG

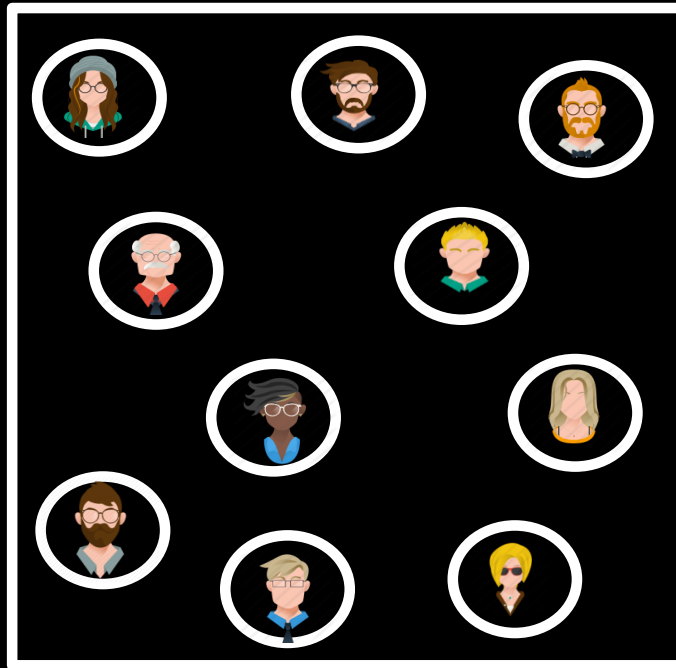
- nDCG is hard to optimize
 - nDCG is non-convex and non-smooth
 - Large input variations → No change in nDCG
 - Small input variations → Large changes in nDCG

$$\text{nDCG} \propto \text{like}(i, \mathbf{r}_1) + \sum_{l=2}^L \frac{\text{like}(i, \mathbf{r}_l)}{\log(l+1)}$$

$$\text{like}(i, \mathbf{r}_l) = \begin{cases} 1 & \text{If user } i \text{ likes the item with rank } \mathbf{r}_l \\ 0 & \text{otherwise} \end{cases}$$

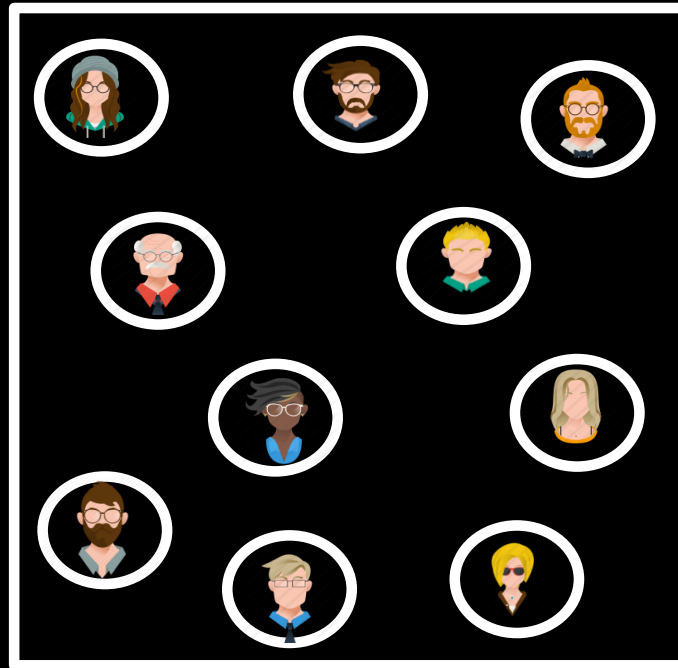
Optimizing nDCG

$$\text{Min}_{\mathbf{w}} \quad \|\mathbf{w}\|_1 - C \sum_{i \in \text{Users}} \text{nDCG}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$$



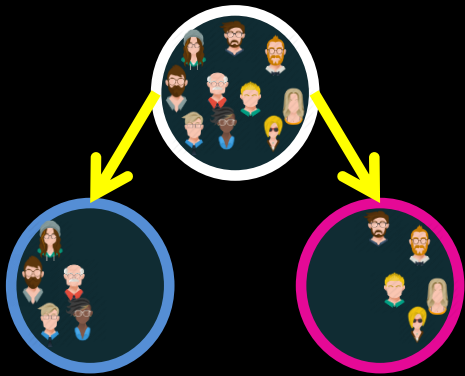
Optimizing nDCG – Reformulation

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \quad \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{\mathbf{y}_i} \mathbf{y}_i$$

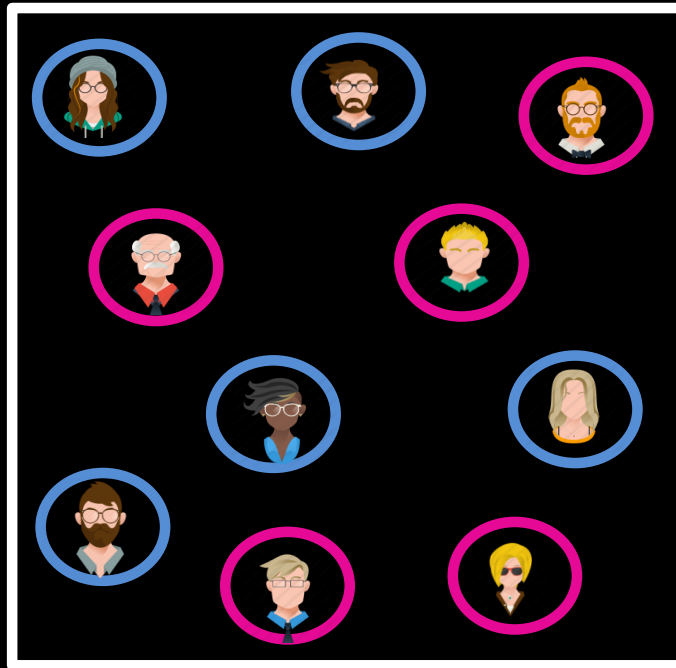


Optimizing nDCG – Initialization

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \quad \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{\mathbf{y}_i} \mathbf{y}_i$$

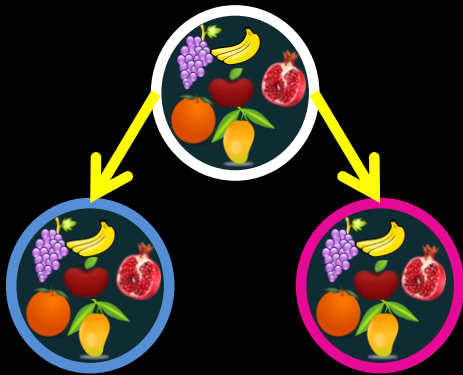


$$\delta_i \sim \text{Bernoulli}(0.5), \forall i$$

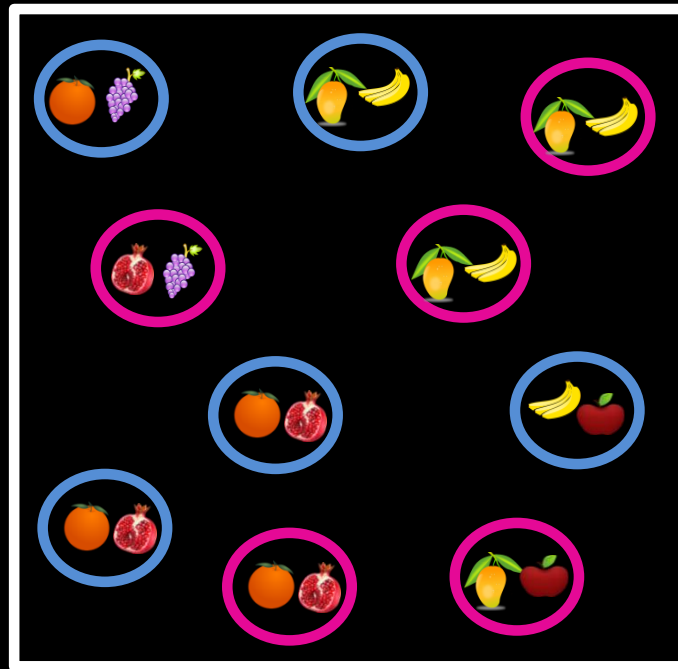


Optimizing nDCG – Initialization

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \quad \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{\mathbf{y}_i} \mathbf{y}_i$$



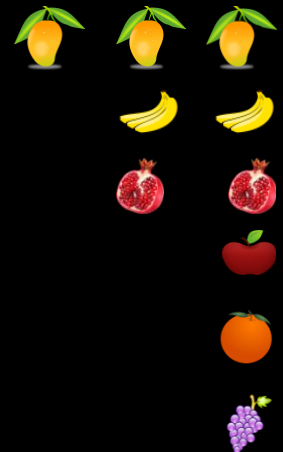
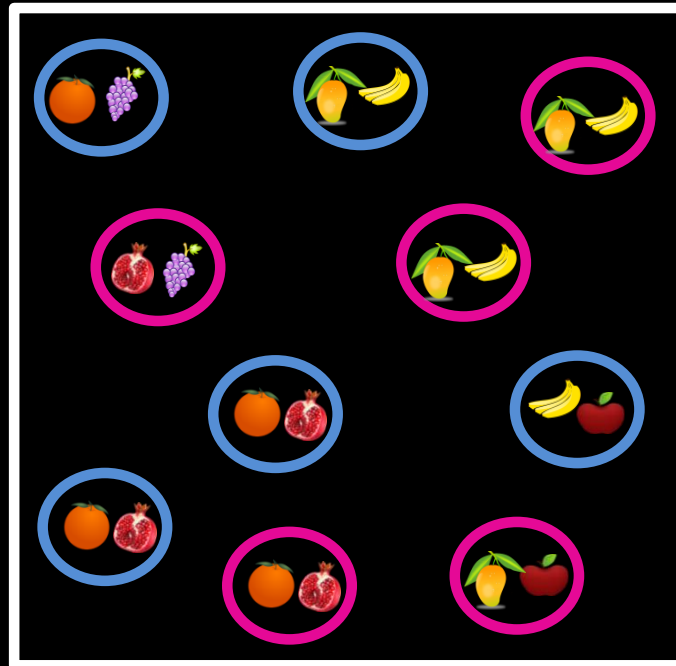
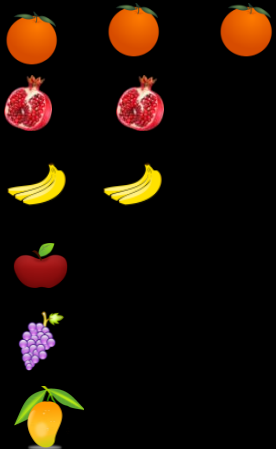
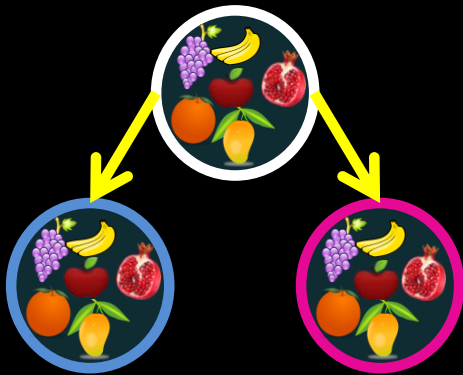
$$\delta_i \sim \text{Bernoulli}(0.5), \forall i$$



Optimizing nDCG – Initialization

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \quad \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{\mathbf{y}_i} \mathbf{y}_i$$

$$\mathbf{r}^{\pm*} = \text{rank} \left(\sum_{i: \delta_i = \pm 1} N_{\mathbf{y}_i} \mathbf{y}_i \right)$$

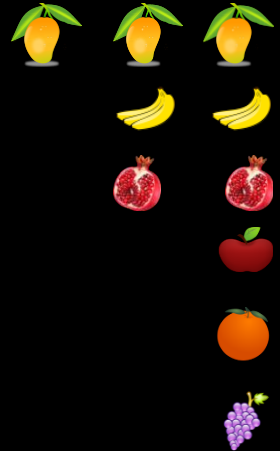
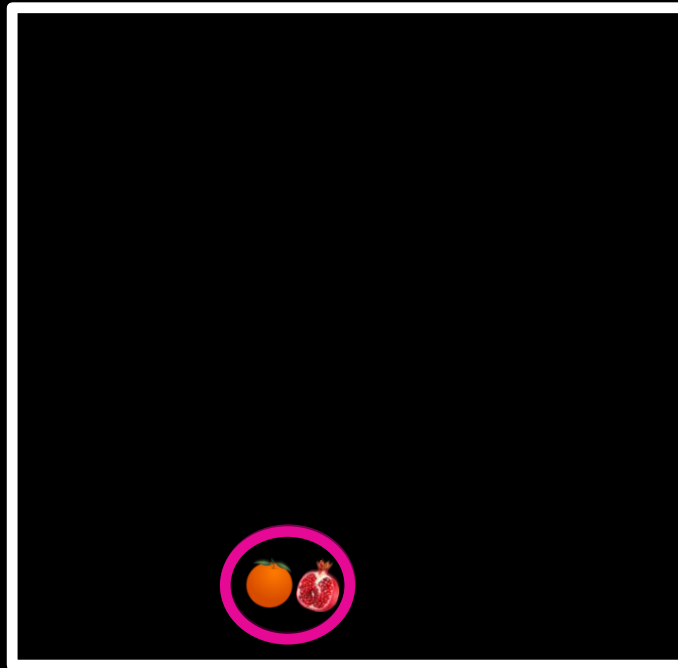
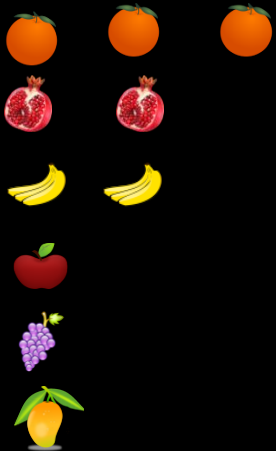
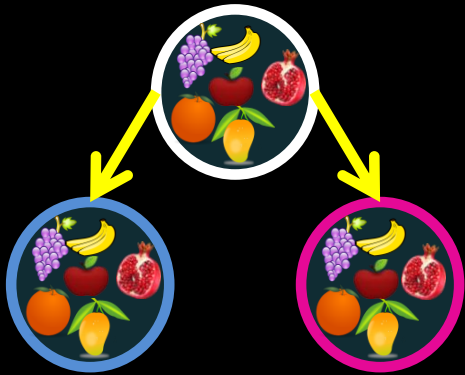


Optimizing nDCG – Repartitioning Users

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{\mathbf{y}_i} \mathbf{y}_i$$

$$\delta_i^* = \text{sign}(v_i^- - v_i^+)$$

$$v_i^{\pm} = C_{\delta}(\pm 1) \log(1 + e^{\mp \mathbf{w}^t \mathbf{x}_i}) - C_r \text{nDCG}(\mathbf{r}^{\pm})^t N_{\mathbf{y}_i} \mathbf{y}_i$$

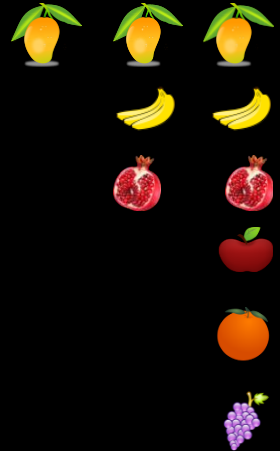
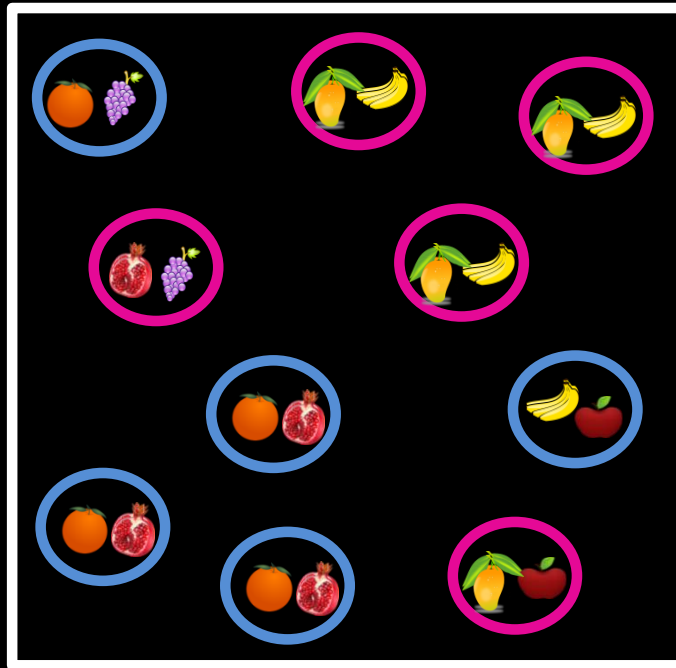
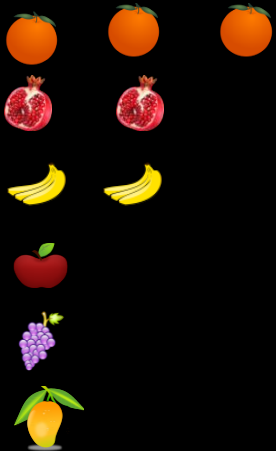
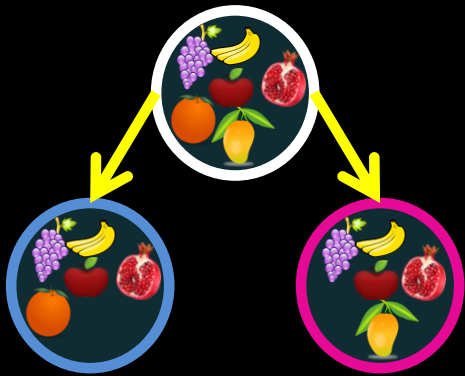


Optimizing nDCG – Repartitioning Users

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \quad \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{y_i} \mathbf{y}_i$$

$$\delta_i^* = \text{sign}(v_i^- - v_i^+)$$

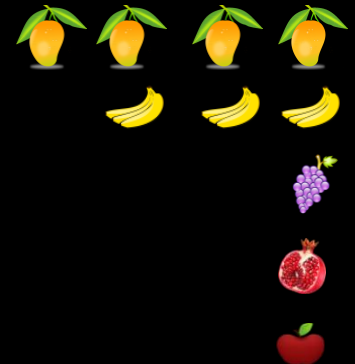
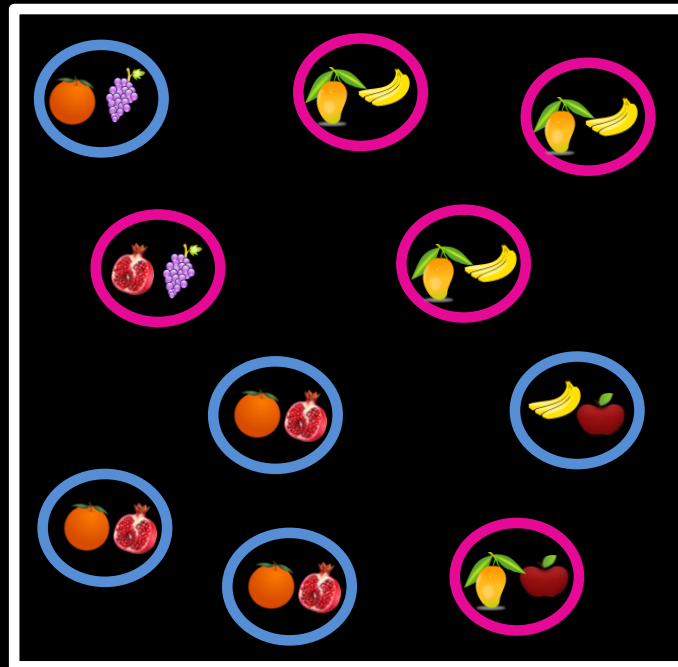
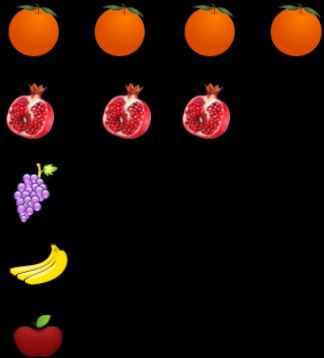
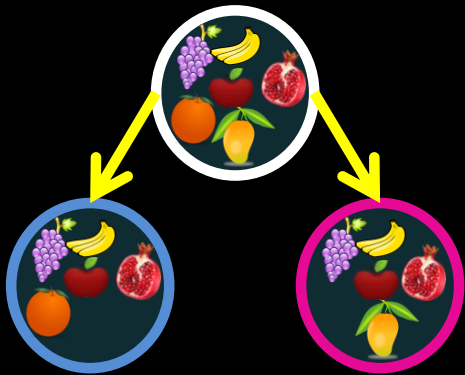
$$v_i^{\pm} = C_{\delta}(\pm 1) \log(1 + e^{\mp \mathbf{w}^t \mathbf{x}_i}) - C_r \text{nDCG}(\mathbf{r}^{\pm})^t N_{y_i} \mathbf{y}_i$$



Optimizing nDCG – Reranking Items

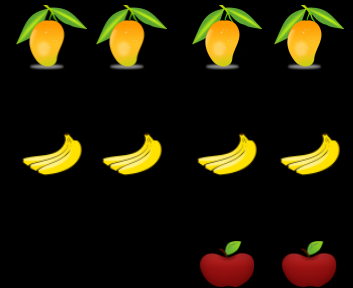
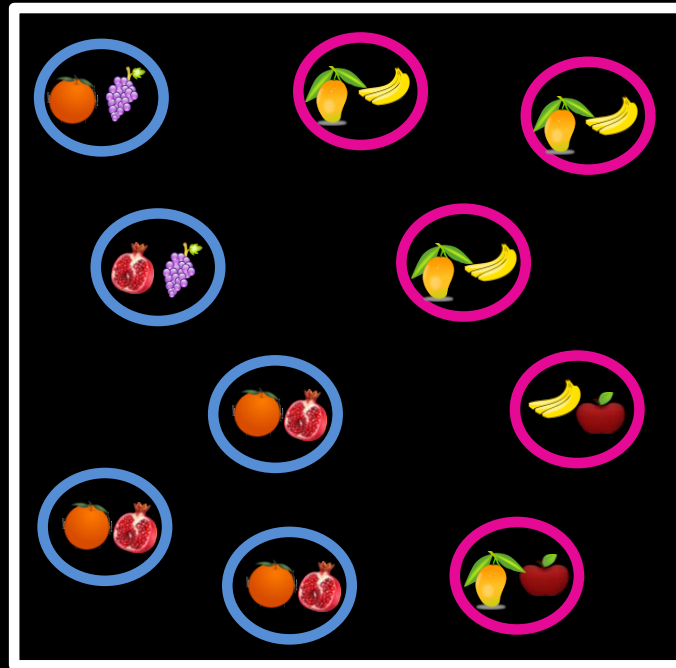
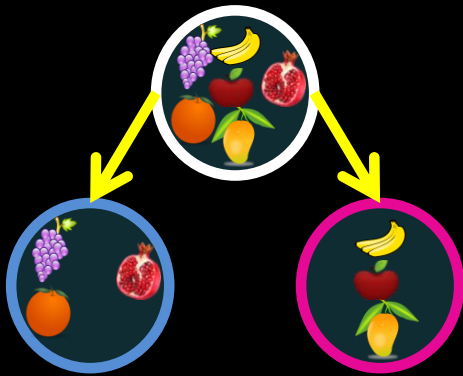
$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{y_i} \mathbf{y}_i$$

$$\mathbf{r}^{\pm*} = \text{rank} \left(\sum_{i: \delta_i = \pm 1} N_{y_i} \mathbf{y}_i \right)$$



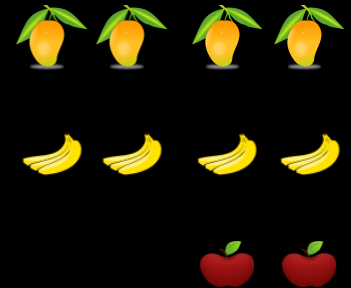
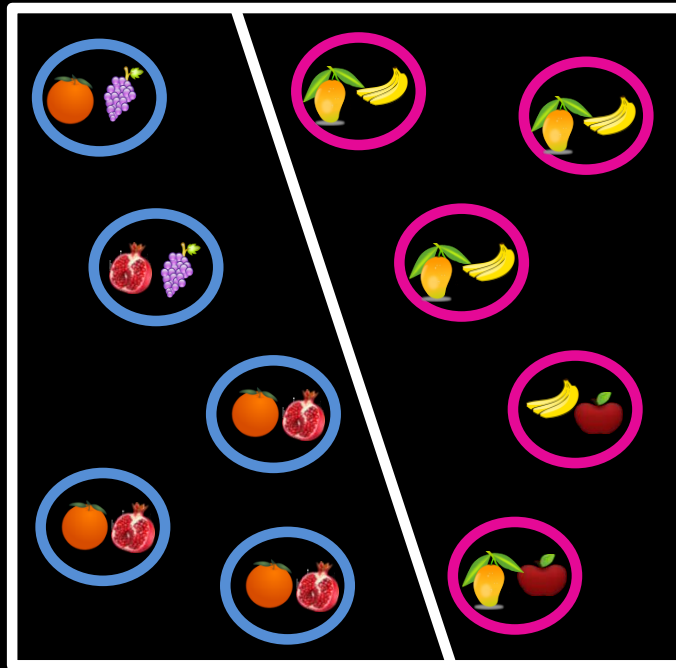
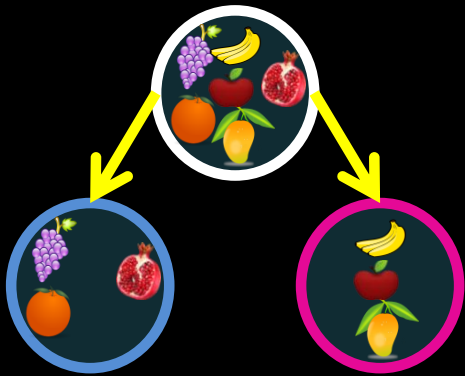
Optimizing nDCG

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{\mathbf{y}_i} \mathbf{y}_i$$



Optimizing nDCG – Hyperplane Separator

$$\text{Min}_{\mathbf{w}, \delta, \mathbf{r}^{\pm}} \quad \|\mathbf{w}\|_1 + \sum_i C_{\delta}(\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}) - C_r \sum_i \text{nDCG}(\mathbf{r}^{\delta_i})^t N_{y_i} y_i$$



Data Set Statistics

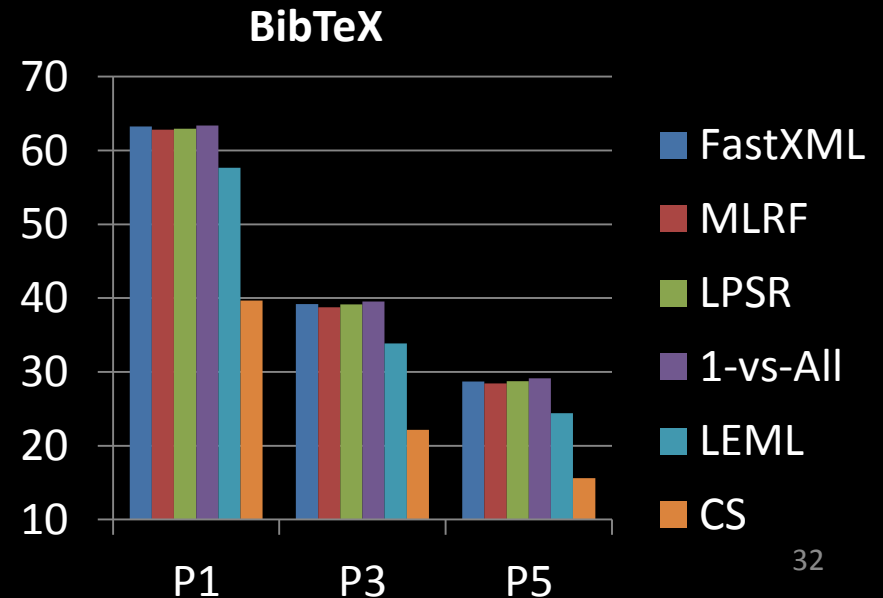
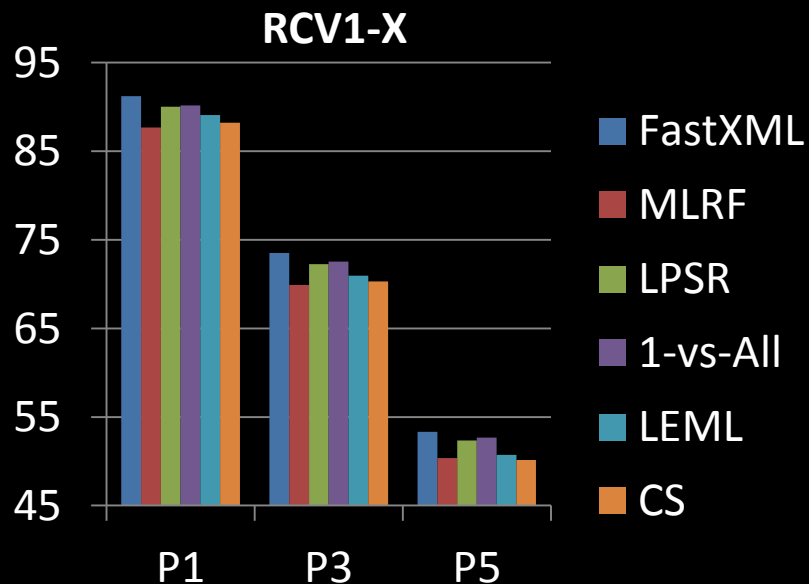
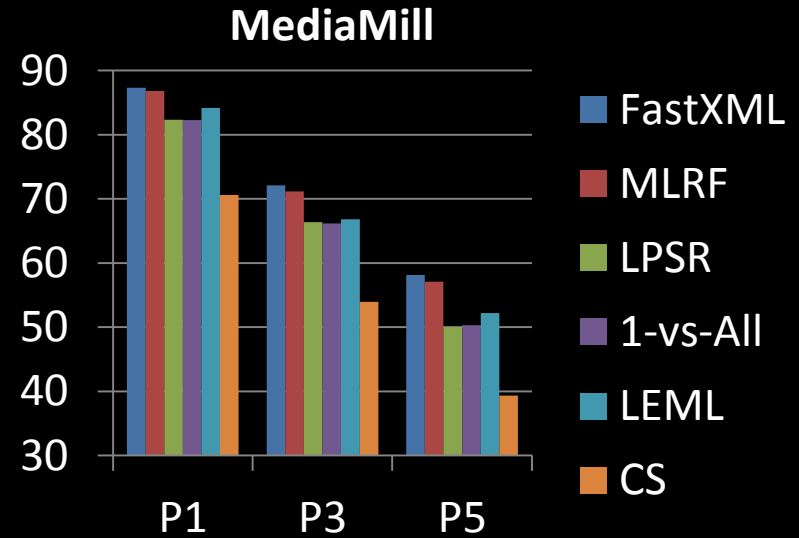
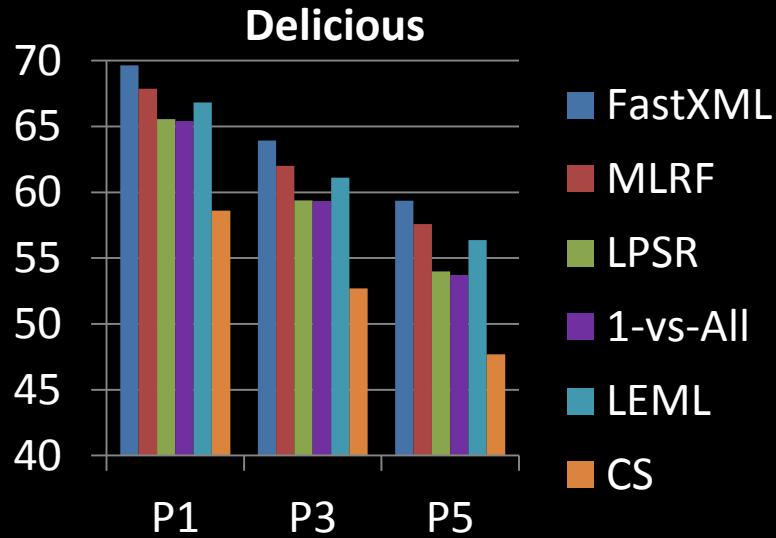
Small data sets

Data Set	# of Training Points	# of Test Points	# of Dimensions	# of Labels
Delicious	12,920	3,185	500	983
MediaMill	30,993	12,914	120	101
RCV1-X	781,265	23,149	47,236	2,456
BibTeX	4,880	2,515	1,836	159

Large data sets

Data Set	# of Training Points (M)	# of Test Points (M)	# of Dimensions (M)	# of Labels (M)
WikiLSHTC	1.89	0.47	1.62	0.33
Ads-430K	1.12	0.50	0.088	0.43
Ads-1M	3.92	1.56	0.16	1.08
Ads-9M	70.46	22.63	2.08	8.84

Results on Small Data Sets

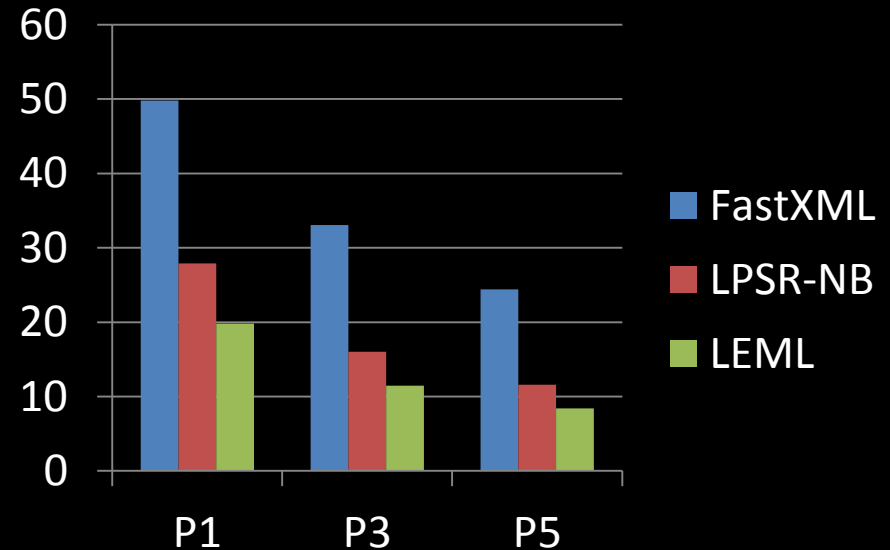


Large Data Sets - WikiLSHTC

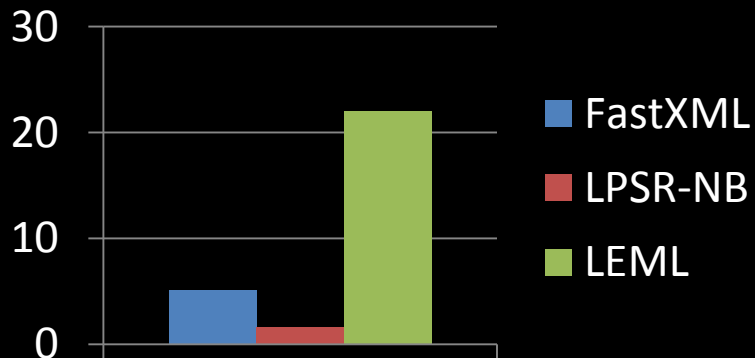
Dataset Statistics

Training Points	1,892,600
Features	1,617,899 (sparse)
Labels	325,056
Test Points	472,835

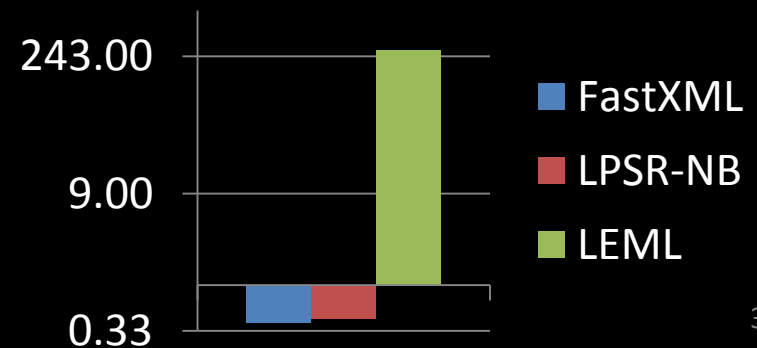
Precision at K



Training Time (hr)



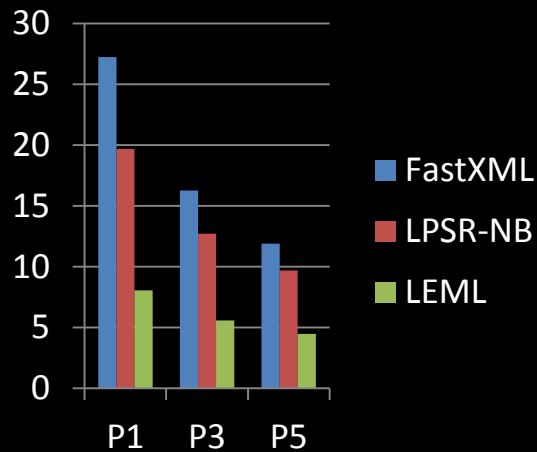
Test Time (millisec)



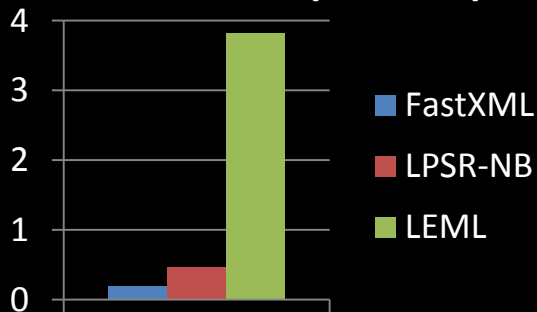
Large Data Sets - Ads

Ads-430K

Precision at K

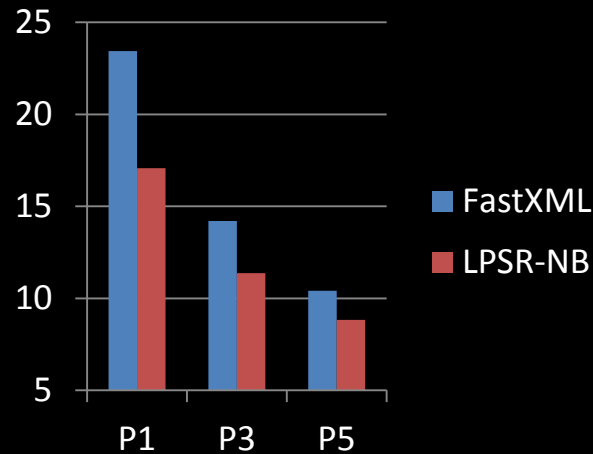


Test Time (millisec)

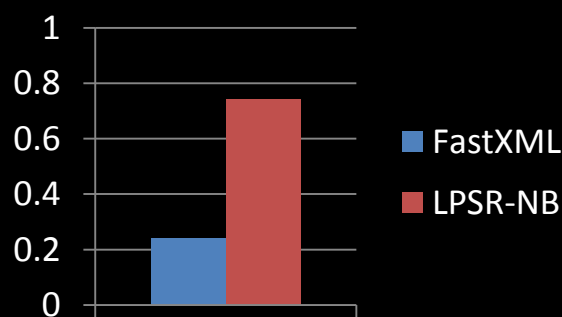


Ads-1M

Precision at K

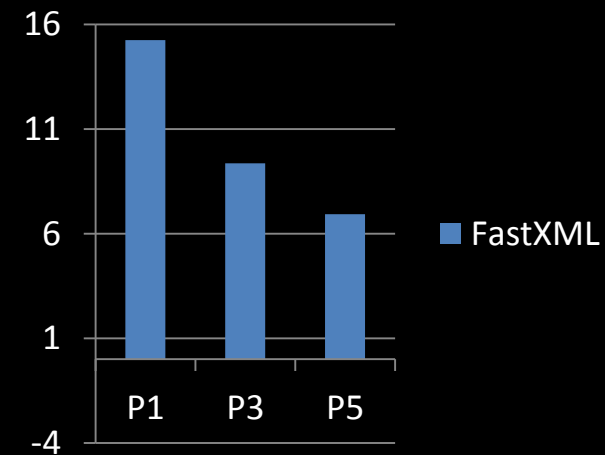


Test Time (millisec)

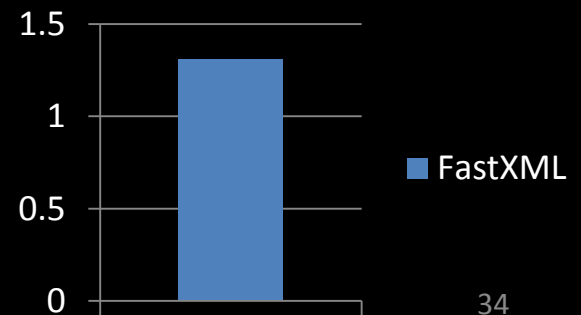


Ads-9M

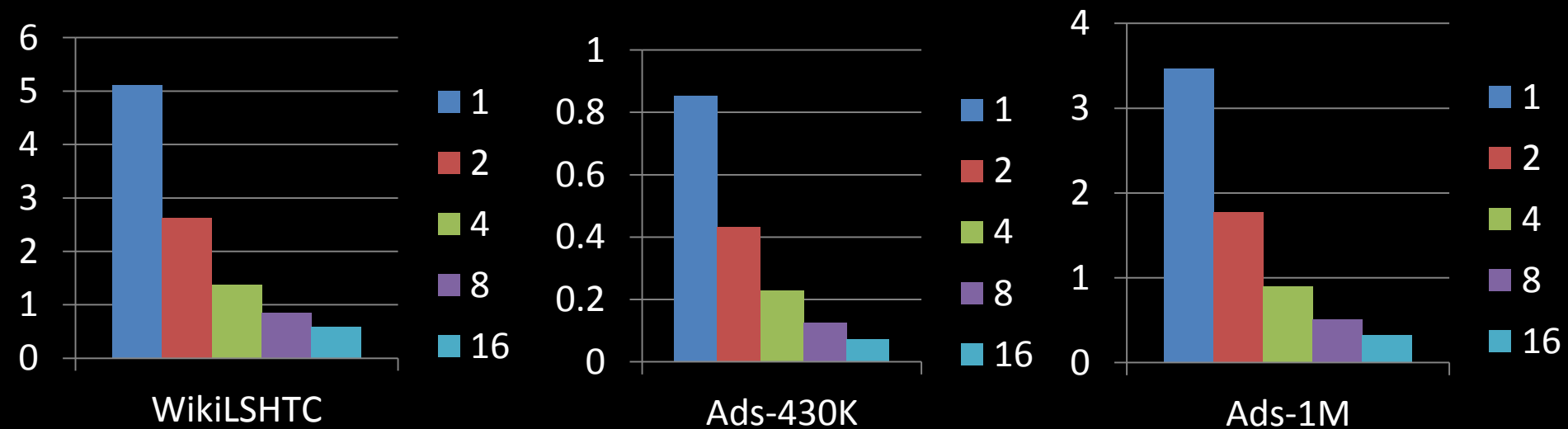
Precision at K



Test Time (millisec)



Training Times in Hours Versus Cores





Conclusions

- Extreme classification
 - Tackle applications with millions of labels
 - A new paradigm for recommendation
- FastXML
 - Significantly higher prediction accuracy
 - Can train on a single desktop
- Publications and code
 - WWW13, KDD14, NIPS15 paperps
 - Code and data available from my website



Unbiased Performance Evaluation

Himanshu Jain (IIT Delhi)

Yashoteja Prabhu (IIT Delhi)

Manik Varma (Microsoft Research)

Traditional Loss/Gain Functions

- Hamming loss
- Subset 0/1 loss
- Precision
- Recall
- F-score
- Jaccard distance



Washington



Lincoln



Kennedy



Jefferson



Roosevelt

1

history	history	history	history	history
politics	politics	politics	politics	politics
people	people	people	people	people
usa	usa	usa	usa	usa
america	america	america	america	america

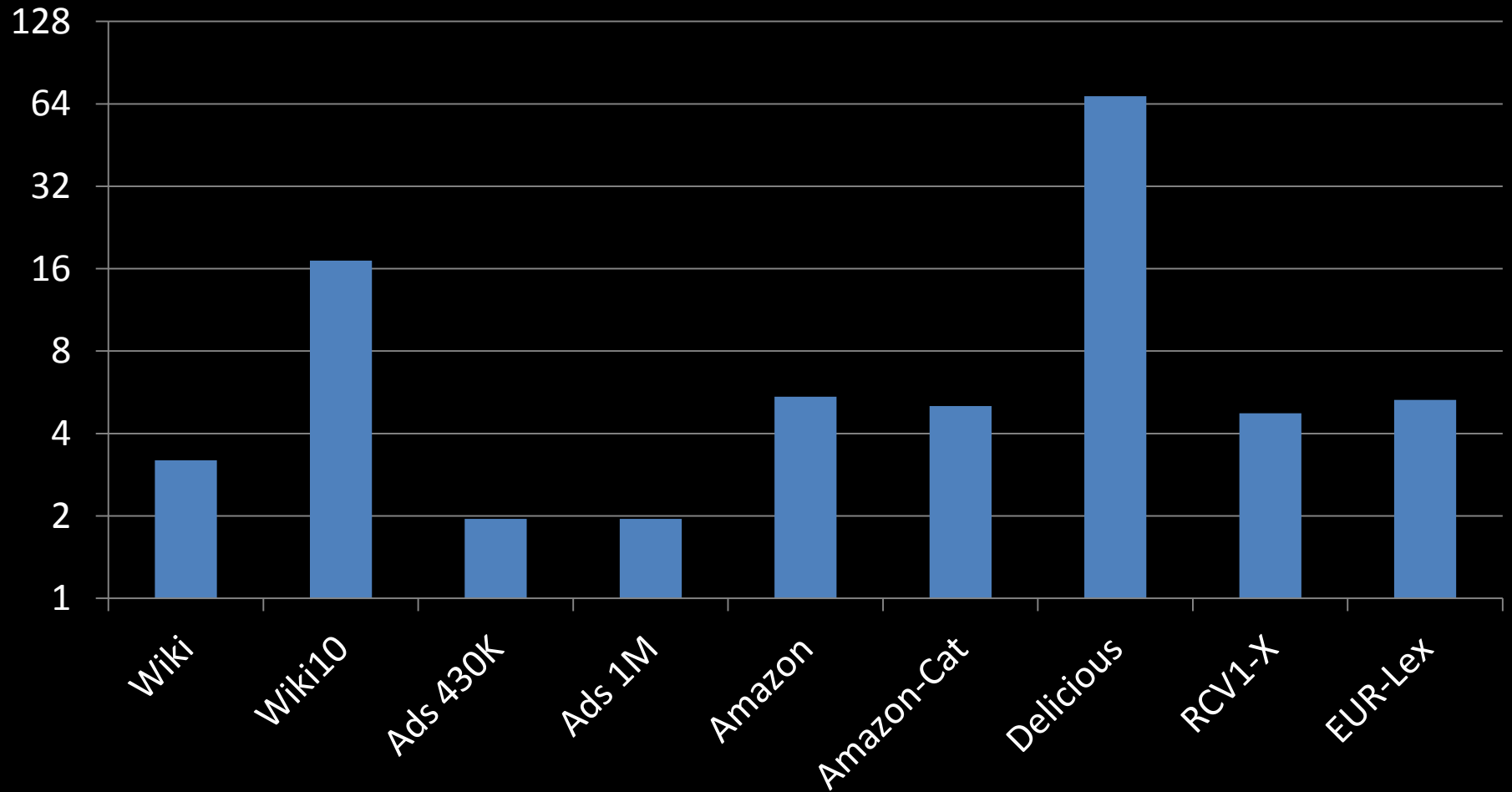
2

history	usa	leader	people	us citizen
politics	america	writer	usa	19 th century born
war	politician	american	thinker	us history
-	-	-	philosopher	-
-	-	-	-	-

3

usa	usa	president	president	usa
first president	president	cuban missile crisis	founding fathers of the us	president
founding fathers of the us	emancipation proclamation	project apollo	declaration of independence	attack on pearl harbour
american revolutionary war	assassinated	assassinated	acquisition of louisiana	great depression
whiskey rebellion	abolition of slavery	-	american revolutionary war	-

Average # of Positive Labels per Point



- +ve labels are more important than -ve ones

Missing Labels

Jeannette Wing - Wikipedia, the free encyclopedia - Windows Internet Explorer

http://en.wikipedia.org/wiki Jeannette Wing - Wikipedia...

NIPS FutureConferences Suggested Sites Web Slice Gallery

the conference will be held on 9th and 10th January 2015 at Jadavpur University
Registration for the conference is now open

Jeannette Wing

From Wikipedia, the free encyclopedia

Jeannette Marie Wing is Corporate Vice President of [Microsoft Research](#) with oversight of its core research laboratories around the world and Microsoft Research Connections.^{[2][3]} Prior to 2013, she was the President's Professor of [Computer Science](#) at [Carnegie Mellon University](#), Pittsburgh, Pennsylvania, United States. She also served as assistant director for Computer and Information Science and Engineering at the [NSF](#) from 2007 to 2010.^{[4][5][6][7][8][9][10][11][12][13]}

Jeannette Wing



[Contents](#) [\[hide\]](#)

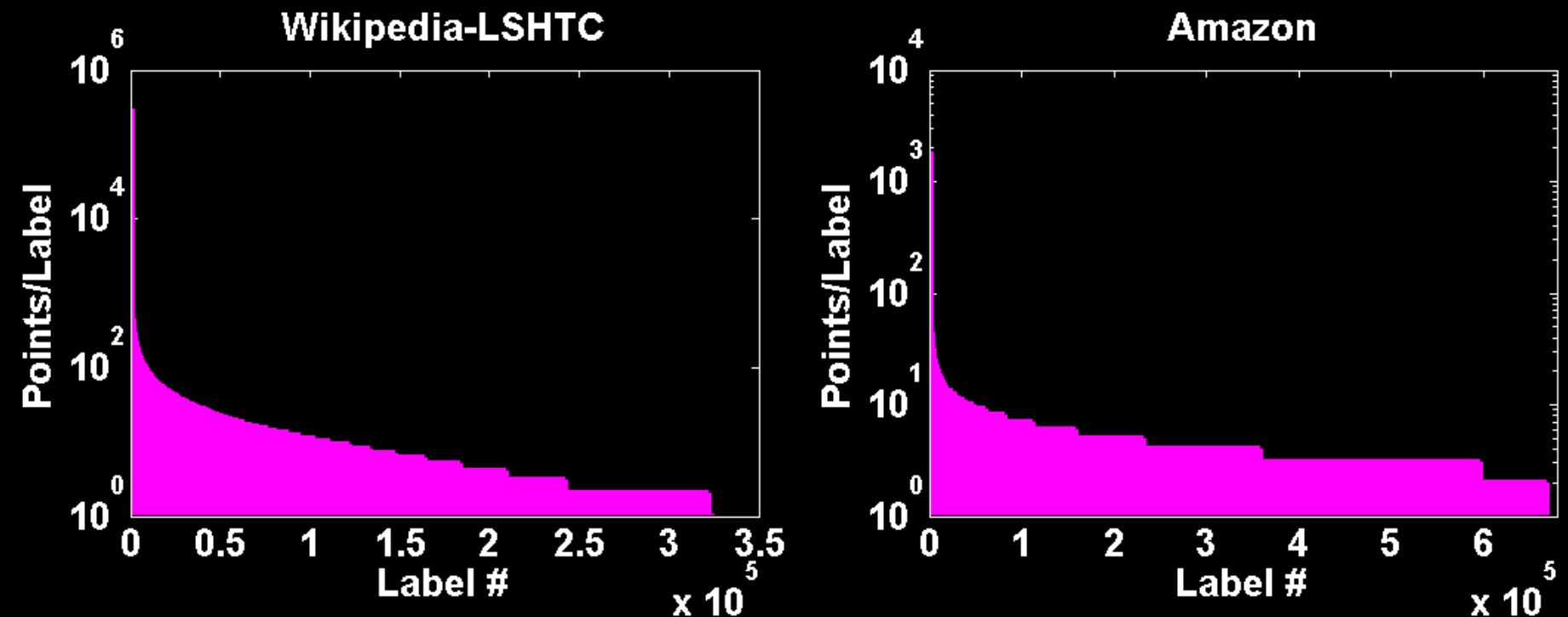
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here

Labels: Living people, American computer scientists, Formal methods people, Carnegie Mellon University faculty, Massachusetts Institute of Technology alumni, Academic journal editors, Women in technology, Women computer scientists.

Tail Labels



- # of relevant labels > # of prediction slots
- Not all positive labels are equally important

Extreme Loss/Gain Functions

- Accuracy – handle biased ground truth
- Rareness / Novelty
- Diversity
- Explainability



Research Problems

- Applications
- Obtaining good quality training data
- Log time and space training and prediction
- Obtaining discriminative features at scale
- Extreme loss functions
- Performance evaluation
- Dealing with tail labels and label correlations
- Dealing with missing and noisy labels
- Explore/exploit for tail labels
- Statistical guarantees
- Fine-grained classification



Acknowledgements

Rahul Agrawal

Kush Bhatia

Shilpa G.

Archit Gupta

Himanshu Jain

Prateek Jain

Abhishek Kadian

Purushottam Kar

Abhirup Nath

Ambuj Tewari

C. Yeshwanth

Ranking and Recommendation

- Traditional approaches – content based methods

$$h : (X, Y) \rightarrow \{\text{✗}, \text{✓}\}$$

$h(\text{👤}, \text{🍊})$

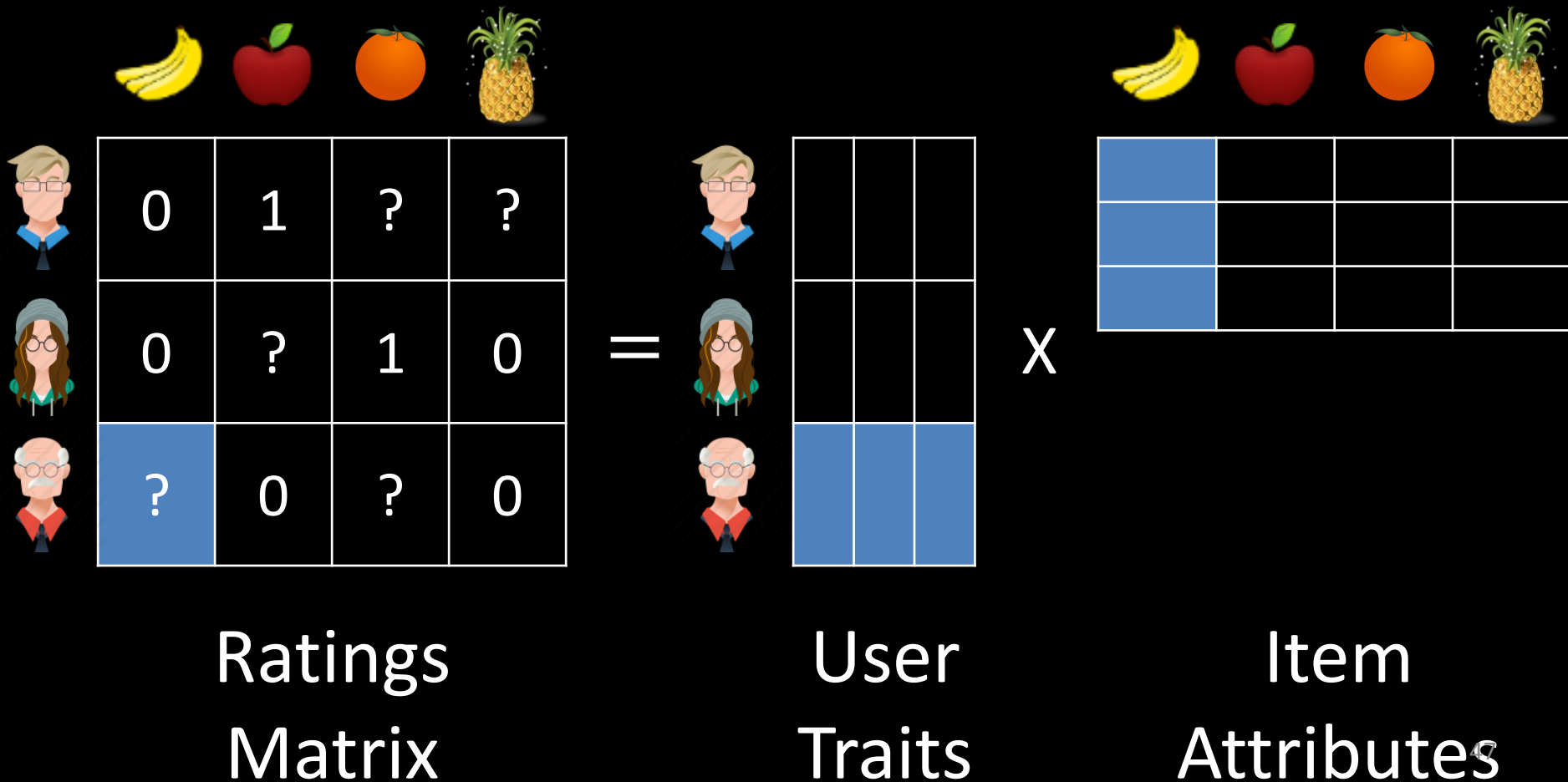


$h(\text{👤}, \text{🍏})$



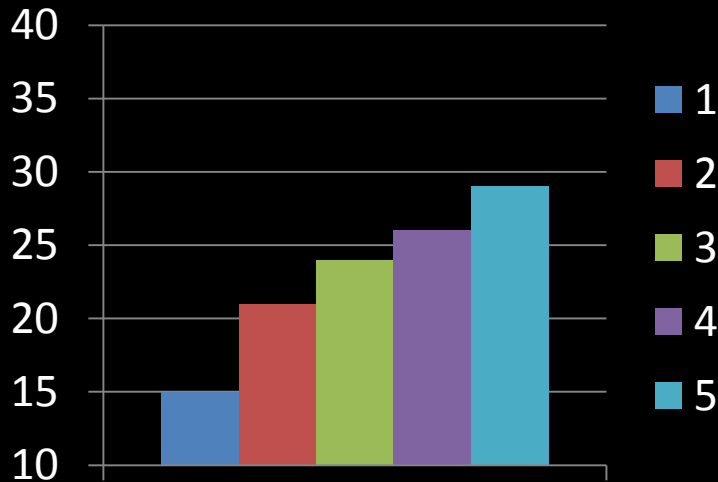
Ranking and Recommendation

- Traditional approaches – matrix factorization

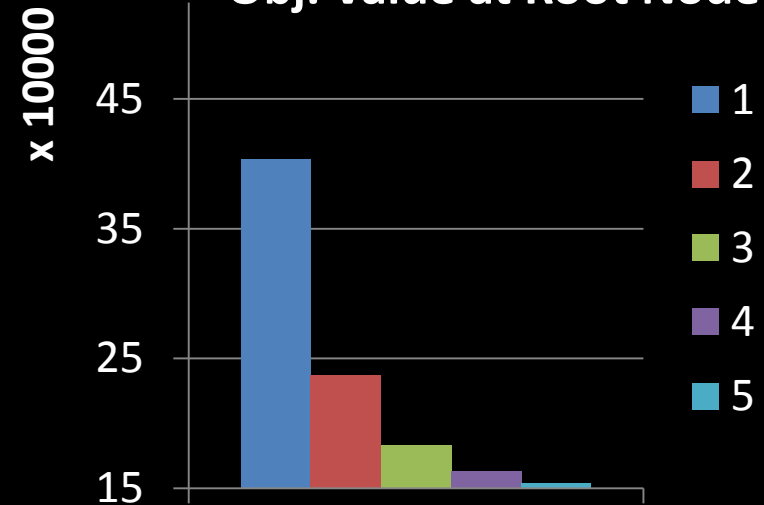


Multiple Iterations - Ads-430K

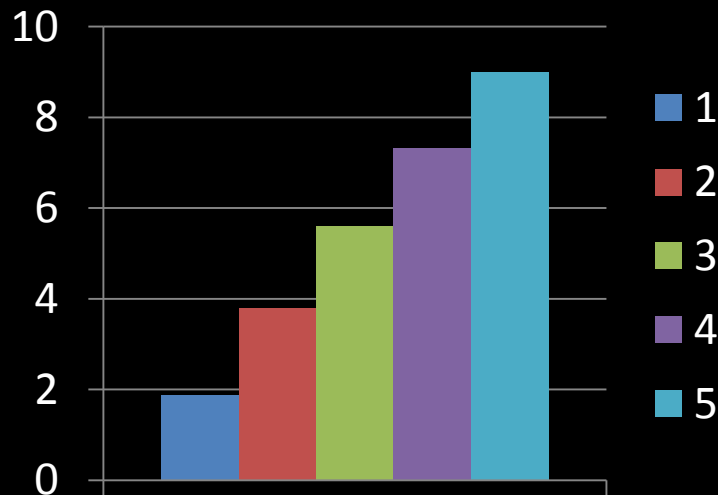
w update Iterations



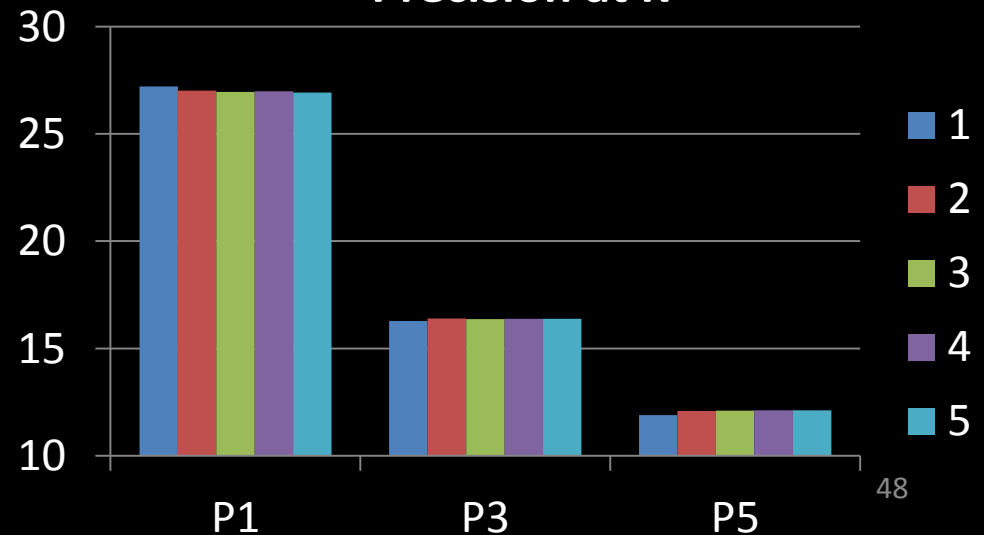
Obj. Value at Root Node



Training Time (hr)

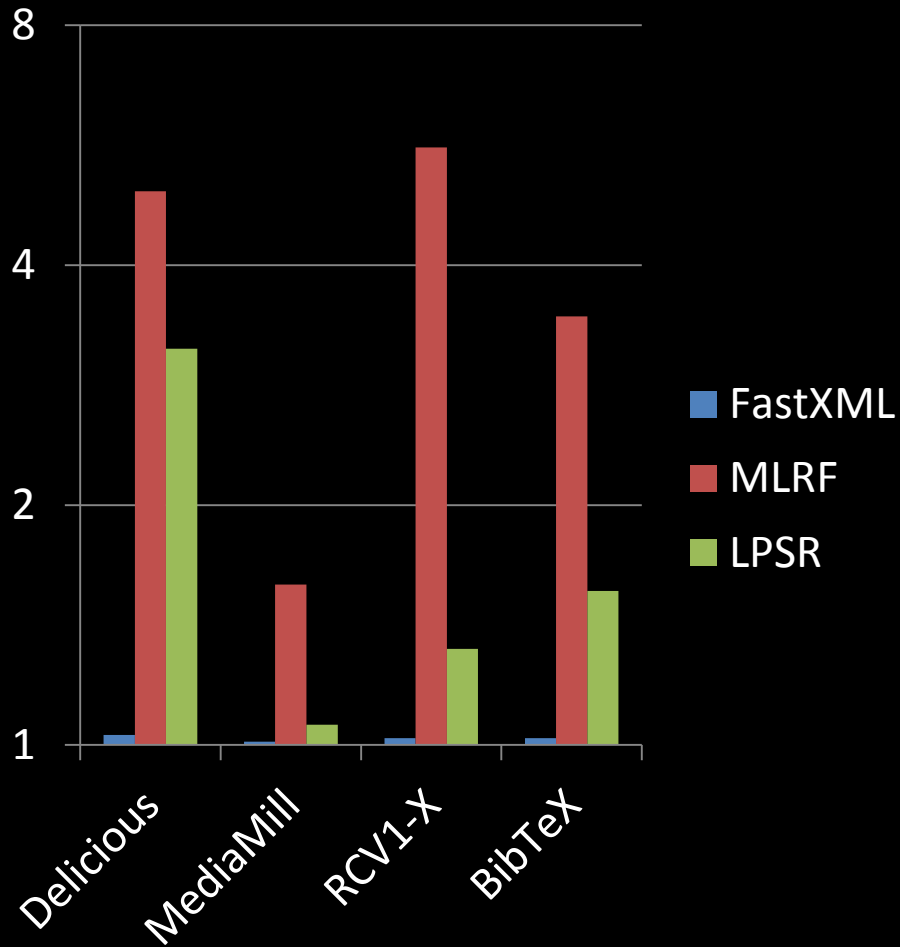


Precision at K

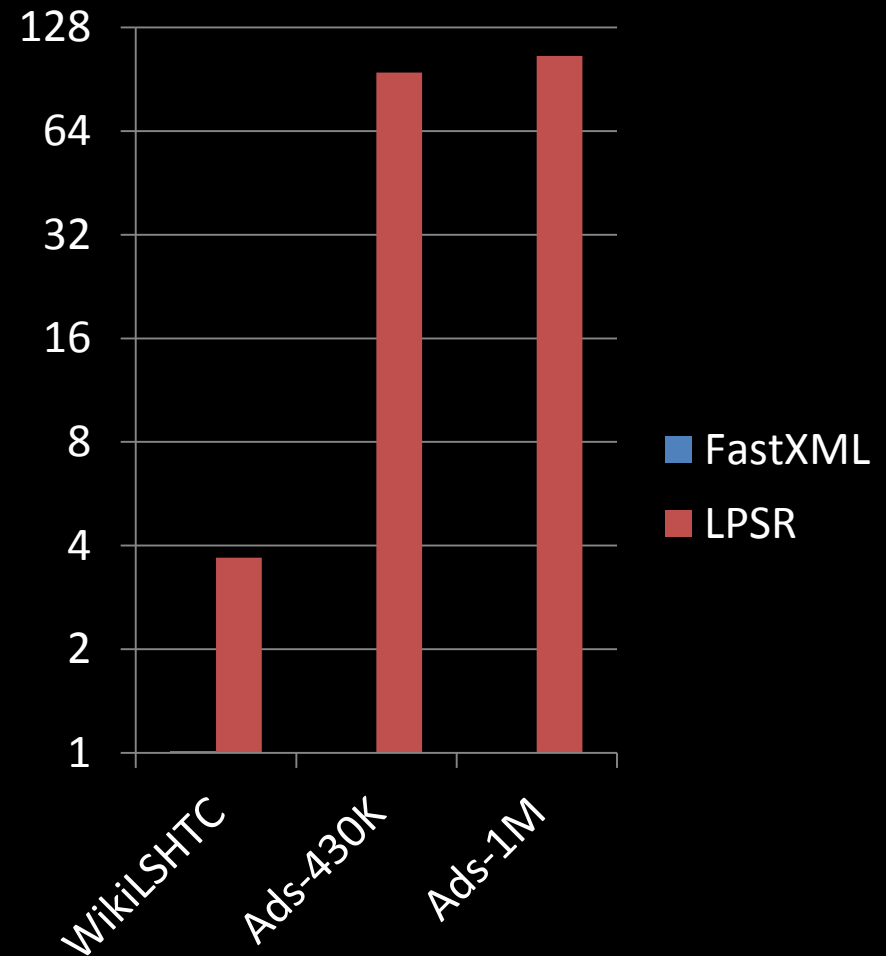


Tree Imbalance

Small Data Sets

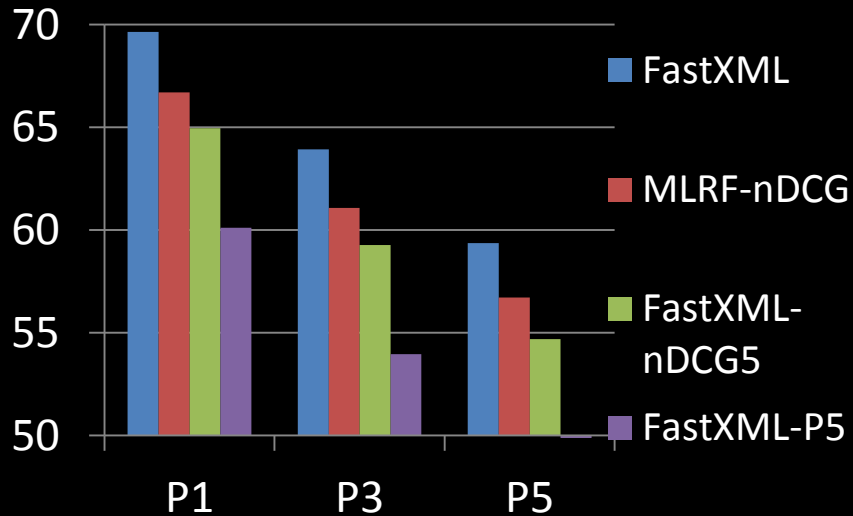


Large Data Sets

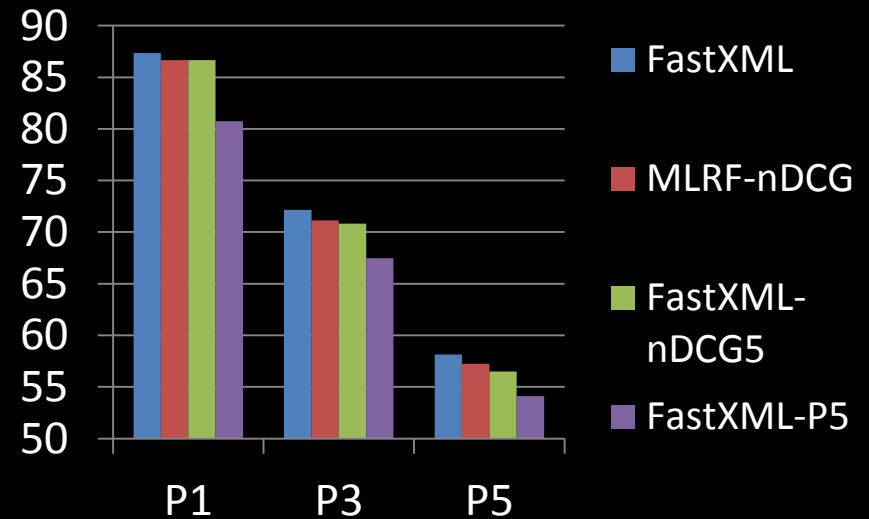


Variants of FastXML - Small Data Sets

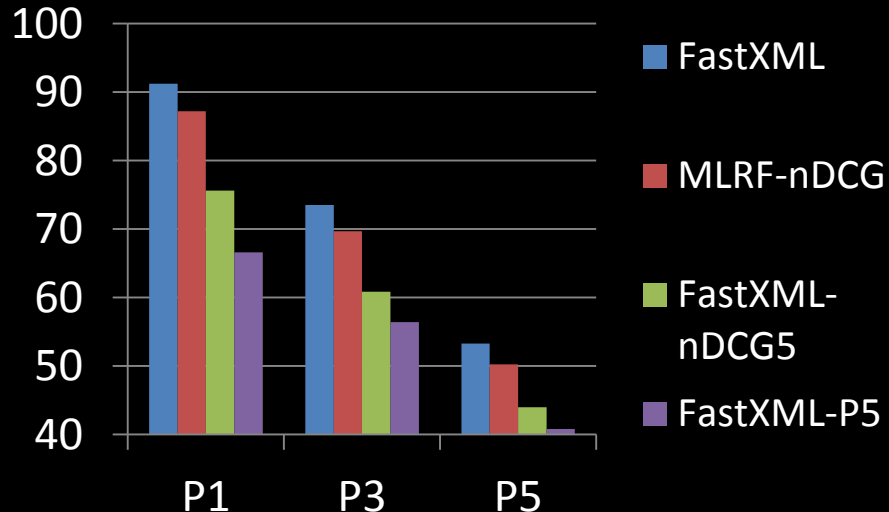
Delicious



MediaMill



RCV1-X

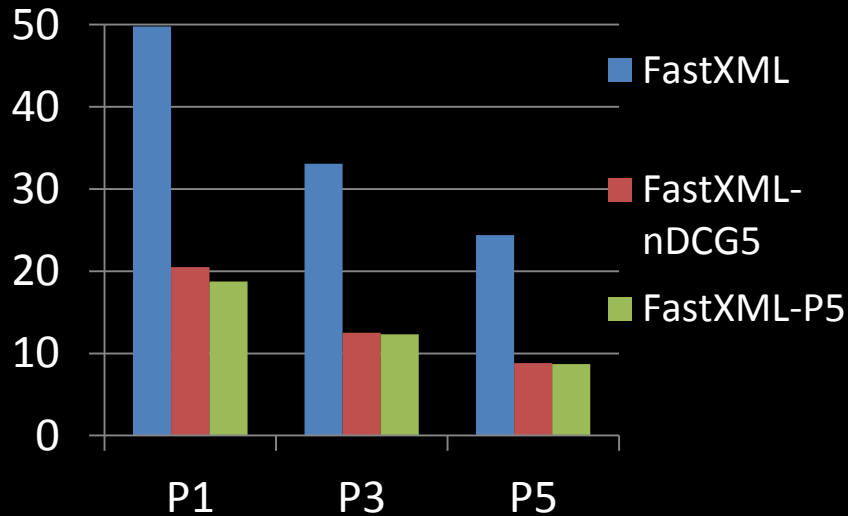


BibTeX

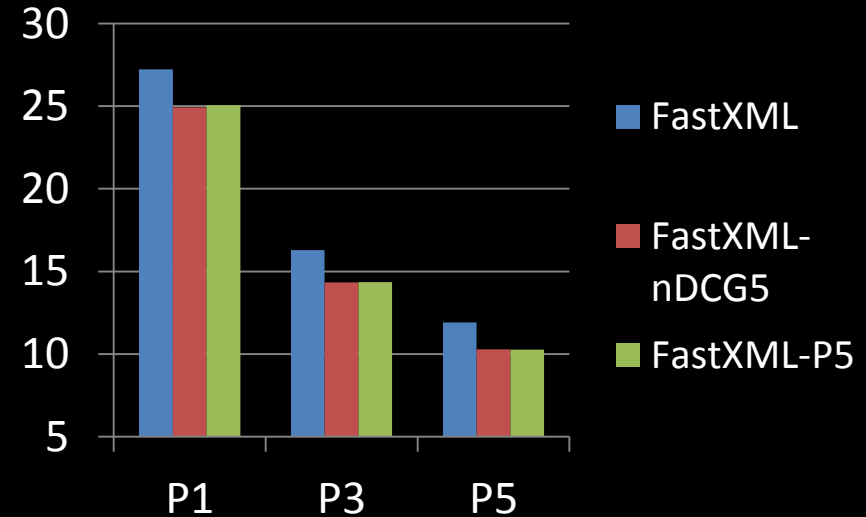


Variants of FastXML - Large Data Sets

WikiLSHTC



Ads-430K



Ads-1M



Random Tree Selection

