

# The Child is Father of the Man: Foresee the Success at the Early Stage

Presenter: [Liangyue Li \(ASU\)](#)

Joint work by



[Liangyue Li](#)  
ASU



[Hanghang Tong](#)  
ASU

# High-impact Scientific Work



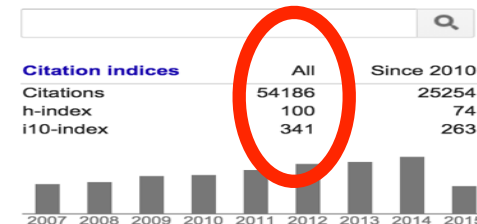
Christos Faloutsos

CMU  
Data Mining, Graph Mining, Databases  
Verified email at cs.cmu.edu - Homepage



Title	1-20	Cited by	Year
<a href="#">On power-law relationships of the internet topology</a> M Faloutsos, P Faloutsos, C Faloutsos ACM SIGCOMM computer communication review 29 (4), 251-262	5357	1999	
<a href="#">QBIC project: querying images by content, using color, texture, and shape</a> CW Niblack, R Barber, W Equitz, MD Flickner, EH Glasman, D Petkovic, ... IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology, 173-187	2447	1993	
<a href="#">Efficient similarity search in sequence databases</a> R Agrawal, C Faloutsos, A Swami Foundations of Data Organization and Algorithms, 69-84	2066	1993	
<a href="#">Efficient and effective querying by image content</a> C Faloutsos, R Barber, M Flickner, J Hafner, W Niblack, D Petkovic, ... Journal of intelligent information systems 3 (3-4), 231-262	1794	1994	
<a href="#">Fast subsequence matching in time-series databases</a> C Faloutsos, M Ranganathan, Y Manolopoulos ACM SIGMOD Record 23 (2), 419-429	1787	1994	

Google Scholar



Co-authors View all...

Jure Leskovec  
Hanghang Tong  
Michalis Faloutsos  
U Kang  
Jia-Yu Pan  
Jimeng Sun  
Agma J. M. Traina  
Leman Akoglu

## Important implications of high-impact scientific work:

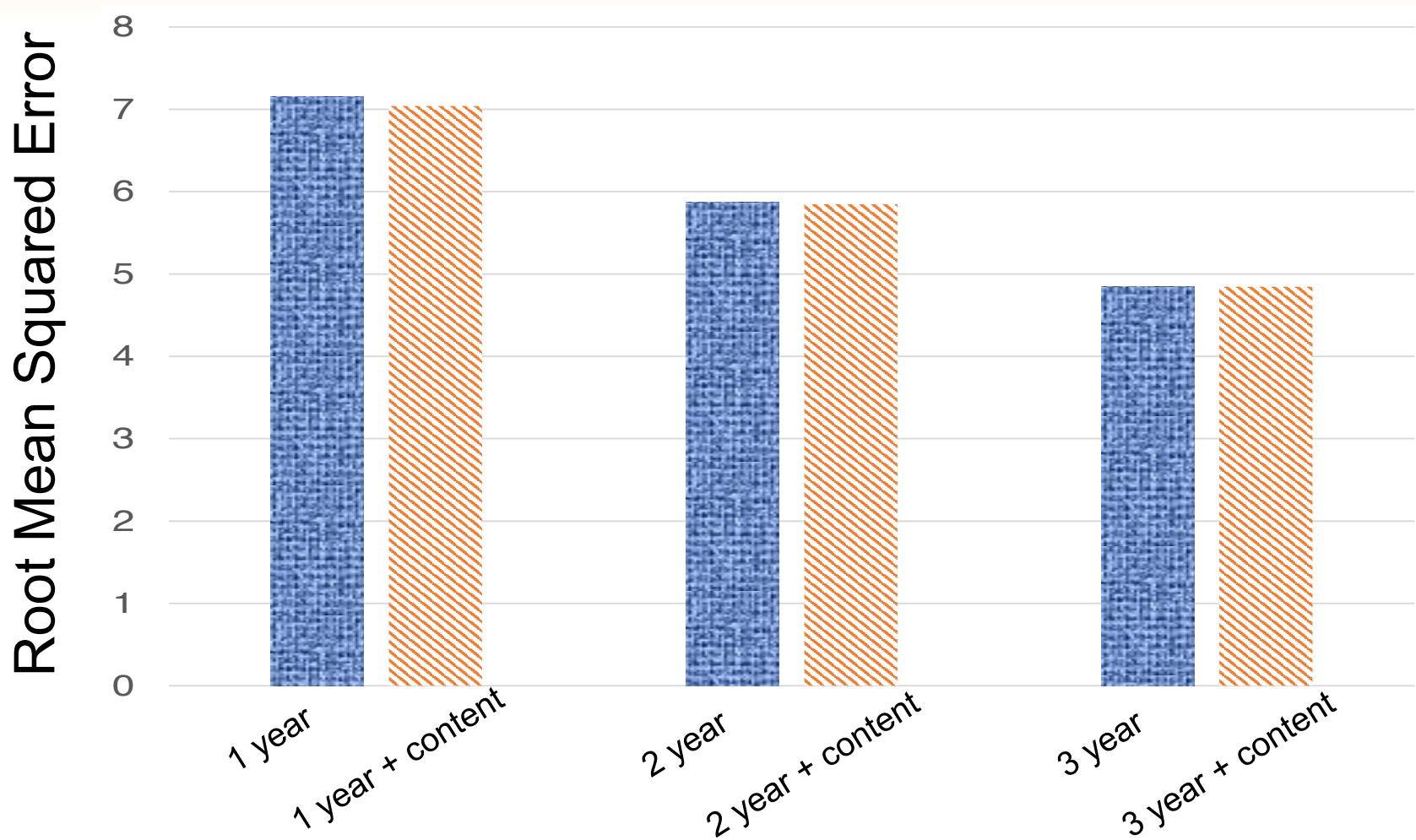
- personal career development
- recruitment search
- jurisdiction of research resources

**Question:** how to forecast the long-term impact at the early stage?

# Challenges

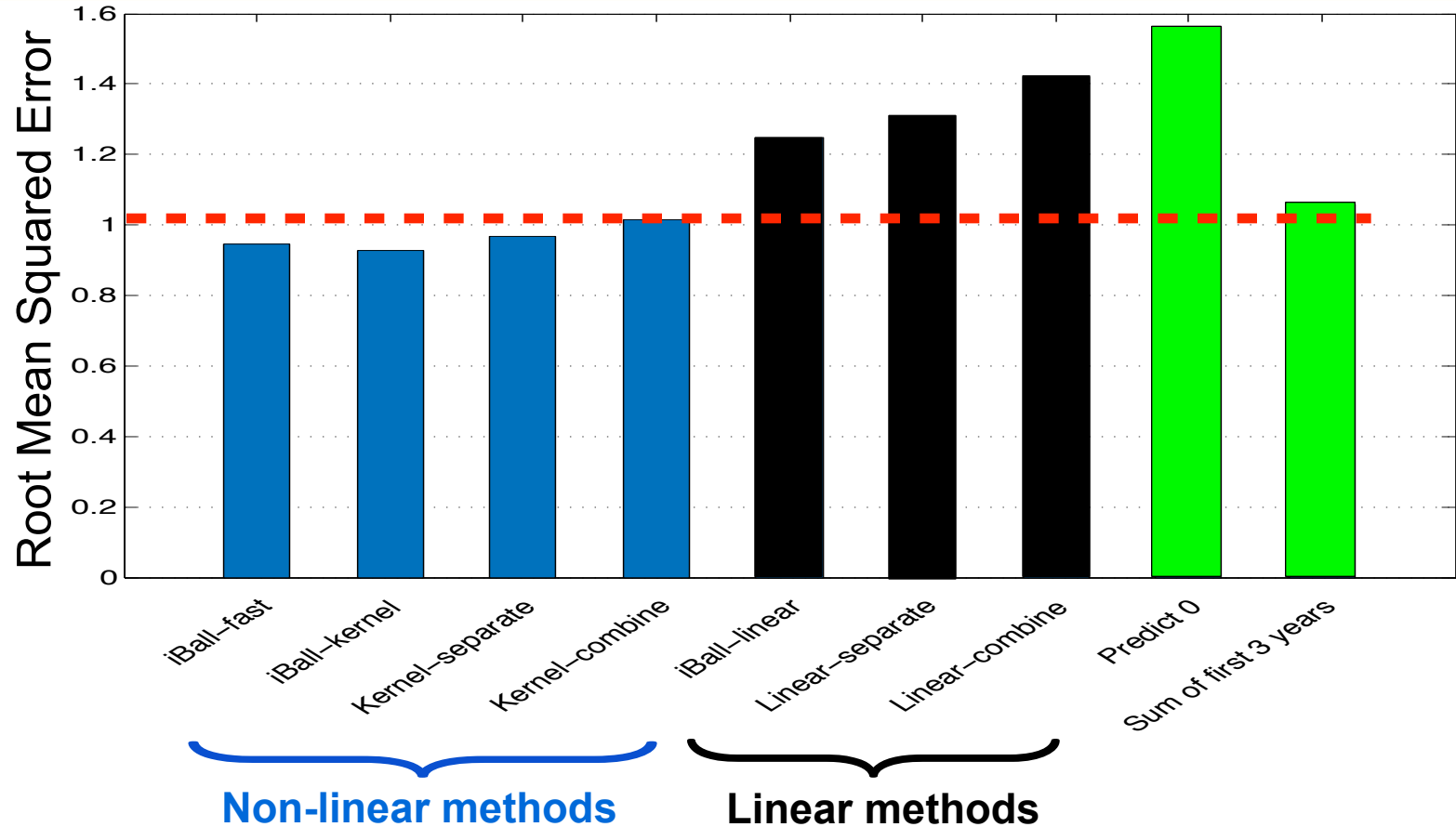
- C1: Scholarly feature design
- C2: Non-linearity
- C3: Domain heterogeneity
- C4: Dynamics

# C1: Scholarly Feature Design



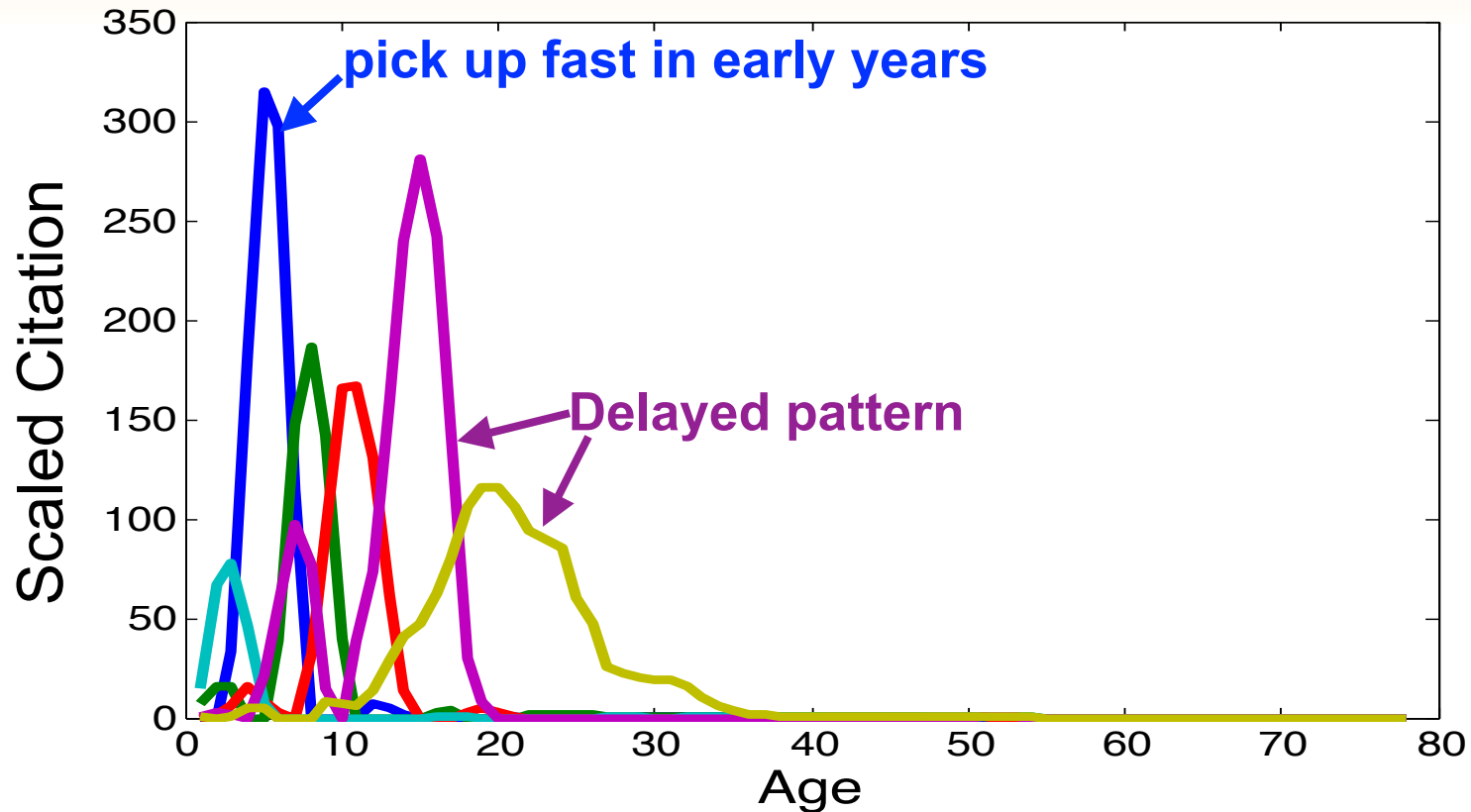
**Obs:** Adding content features brings little improvement

# C2: Non-linearity



**Obs:** Non-linear methods outperform linear ones

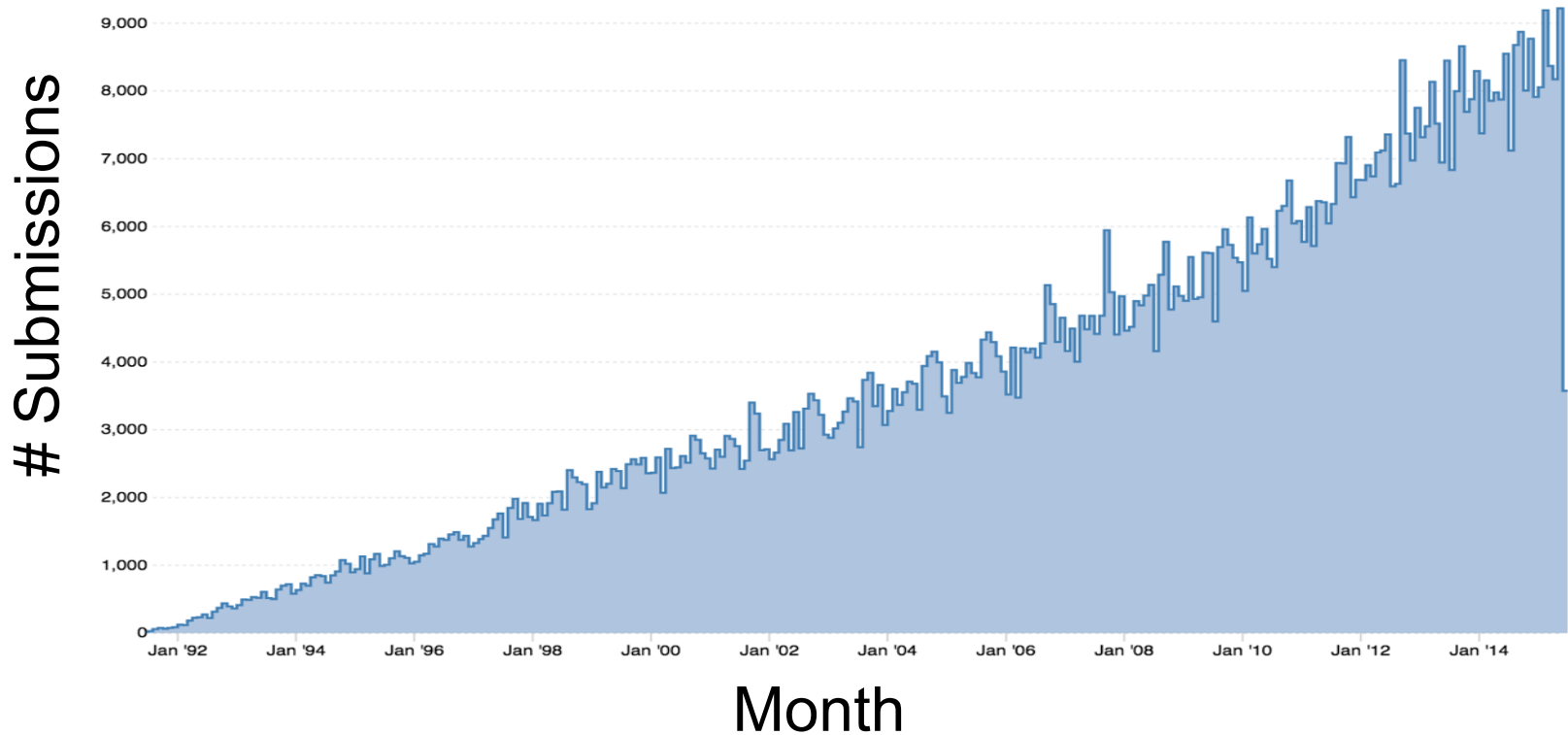
# C3: Domain heterogeneity



**Obs:** Impact of scientific work from different domains behaves differently

# C4: Dynamics

## arXiv monthly submission rates



**Question:** How to quickly update the predictive model?

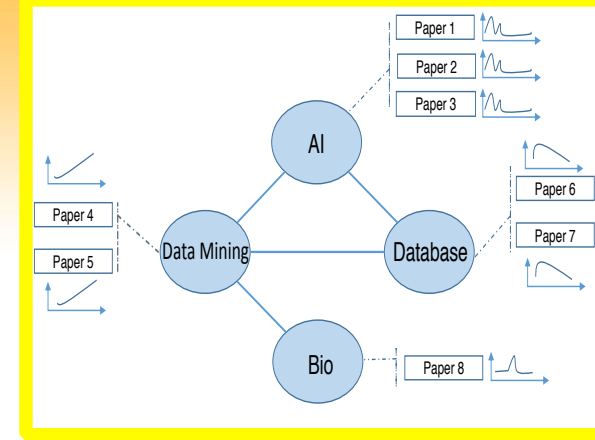
# Roadmap

- Motivations
- **Proposed Solutions: iBall**
- Experimental Results
- Conclusions



# iBall — Formulations

## ■ Optimization Formulation



### Within-Domain Model

$$\min_{\mathbf{w}^{(i)}, i=1, \dots, n_d} \sum_{i=1}^{n_d} \mathcal{L}[f(\mathbf{X}^{(i)}, \mathbf{w}^{(i)}), \mathbf{Y}^{(i)}] + \lambda \sum_{i=1}^{n_d} \Omega(\mathbf{w}^{(i)})$$

$$+ \theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \mathbf{A}_{ij} g(\mathbf{w}^{(i)}, \mathbf{w}^{(j)})$$

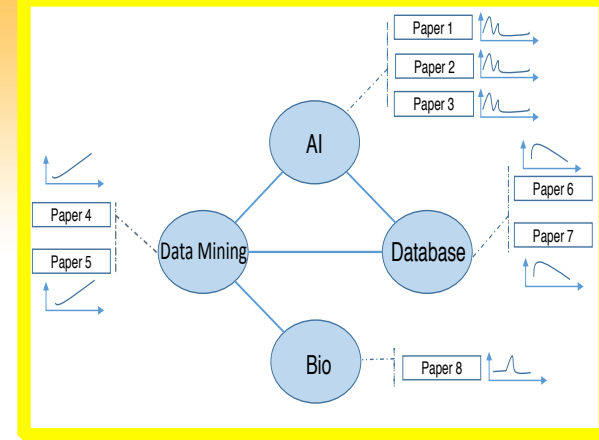
**Cross-Domain Consistency**

## ■ Remarks

- **Within-Domain Model**: regression/classification, linear/non-linear
- **Cross-Domain Consistency**: similar domains have similar models

**Question: how to instantiate such consistency?**

# iBall — linear formulation



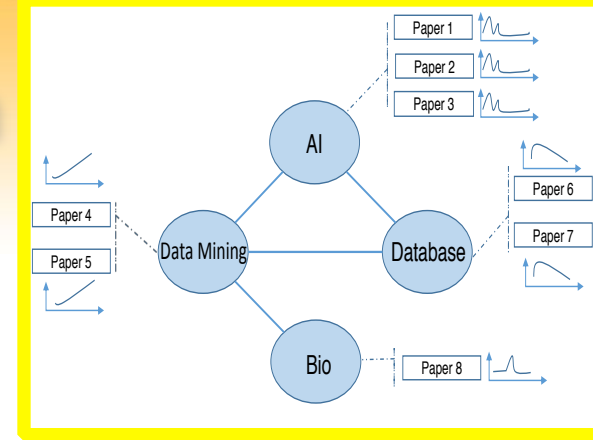
**Details:**

$$\min_{\mathbf{w}^{(i)}, i=1, \dots, n_d} \sum_{i=1}^{n_d} \|\mathbf{X}^{(i)} \mathbf{w}^{(i)} - \mathbf{Y}^{(i)}\|_2^2 + \lambda \sum_{i=1}^{n_d} \|\mathbf{w}^{(i)}\|_2^2 + \theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \mathbf{A}_{ij} \|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\|_2^2$$

**Intuitions:** similar domain (large  $\mathbf{A}_{ij}$ )

→ same feature has similar impact (small  $\|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\|_2^2$ )

# iBall — non-linear formulation



**Details:**

$$\min_{\mathbf{w}^{(i)}, i=1, \dots, n_d} \sum_{i=1}^{n_d} \|\mathbf{K}^{(i)} \mathbf{w}^{(i)} - \mathbf{Y}^{(i)}\|_2^2 + \lambda \sum_{i=1}^{n_d} \mathbf{w}^{(i)'} \mathbf{K}^{(i)} \mathbf{w}^{(i)} + \theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \mathbf{A}_{ij} \|\mathbf{K}^{(i)} \mathbf{w}^{(i)} - \mathbf{K}^{(j)} \mathbf{w}^{(j)}\|_2^2$$

Predicted output  
(domain  $i \rightarrow$  domain  $i$ )

Predicted output  
(domain  $j \rightarrow$  domain  $i$ )

**Intuitions:** similar domain (large  $\mathbf{A}_{ij}$ )

→ similar predicted outputs (small  $\|\mathbf{K}^{(i)} \mathbf{w}^{(i)} - \mathbf{K}^{(j)} \mathbf{w}^{(j)}\|_2^2$ )

# iBall — Closed-form Solutions

- Closed-form Solution

$$\mathbf{w} = \mathbf{S}^{-1} \mathbf{Y}$$

➔ iBall — linear:

$$\mathbf{w} = [\mathbf{w}^{(1)}; \dots; \mathbf{w}^{(n_d)}] \quad \mathbf{Y} = [\mathbf{X}^{(1)'} \mathbf{Y}^{(1)}; \dots; \mathbf{X}^{(n_d)'} \mathbf{Y}^{(n_d)}]$$

$$\mathbf{S} = \begin{bmatrix} \dots & \dots & \dots \\ \dots & \mathbf{X}^{(i)'} \mathbf{X}^{(i)} + (\theta \sum_{j=1}^{n_d} \mathbf{A}_{ij} + \lambda) \mathbf{I} & -\theta \mathbf{A}_{ij} \mathbf{I} \\ \dots & \dots & \dots \end{bmatrix} \begin{array}{l} \text{i-th block column} \\ \text{j-th block column} \\ \text{i-th block row} \\ \text{row} \end{array}$$

**Time Complexity:**  $O((dn_d)^3)$  *d: feature dim  $n_d$ : # of domains*  
 *$dn_d$  is in the order of 10 or 100*

# iBall — Closed-form Solutions

- Closed-form Solution

$$\mathbf{w} = \mathbf{S}^{-1} \mathbf{Y}$$

➔ iBall — non-linear:

$$\mathbf{w} = [\mathbf{w}^{(1)}; \dots; \mathbf{w}^{(n_d)}] \quad \mathbf{Y} = [\mathbf{Y}^{(1)}; \dots; \mathbf{Y}^{(n_d)}]$$

i-th block column                      j-th block column

$$\mathbf{S} = \begin{bmatrix} \dots & \dots & \dots \\ \dots & (1 + \theta \sum_{j=1}^{n_d} \mathbf{A}_{ij}) \mathbf{K}^{(i)} + \lambda \mathbf{I} & -\theta \mathbf{A}_{ij} \mathbf{K}^{(ij)} \\ \dots & \dots & \dots \end{bmatrix}$$

i-th block row

**Time Complexity:**  $O(n^3)$      $n$  : total # of training samples  
 $n$  is in the order of millions

# iBall — Scale-up with Dynamic Update

- **Key idea #1:** Approx  $\mathbf{S}$  by low-rank approx
- **Details:**

$$\mathbf{S}_{t+1} \approx \mathbf{U}_{t+1} \mathbf{\Lambda}_{t+1} \mathbf{U}'_{t+1} \quad \longrightarrow \quad \mathbf{w}_{t+1} = \mathbf{S}_{t+1}^{-1} \mathbf{Y}_{t+1}$$

(Overall:  $O(n^2 r)$ )

$$= \mathbf{U}_{t+1} \mathbf{\Lambda}_{t+1}^{-1} \mathbf{U}'_{t+1} \mathbf{Y}_{t+1}$$

(Overall:  $O(nr)$ )

- **Complexity:**  $O(n^3) \rightarrow O(n^2 r + nr)$
- **Benefit:** avoid matrix inverse

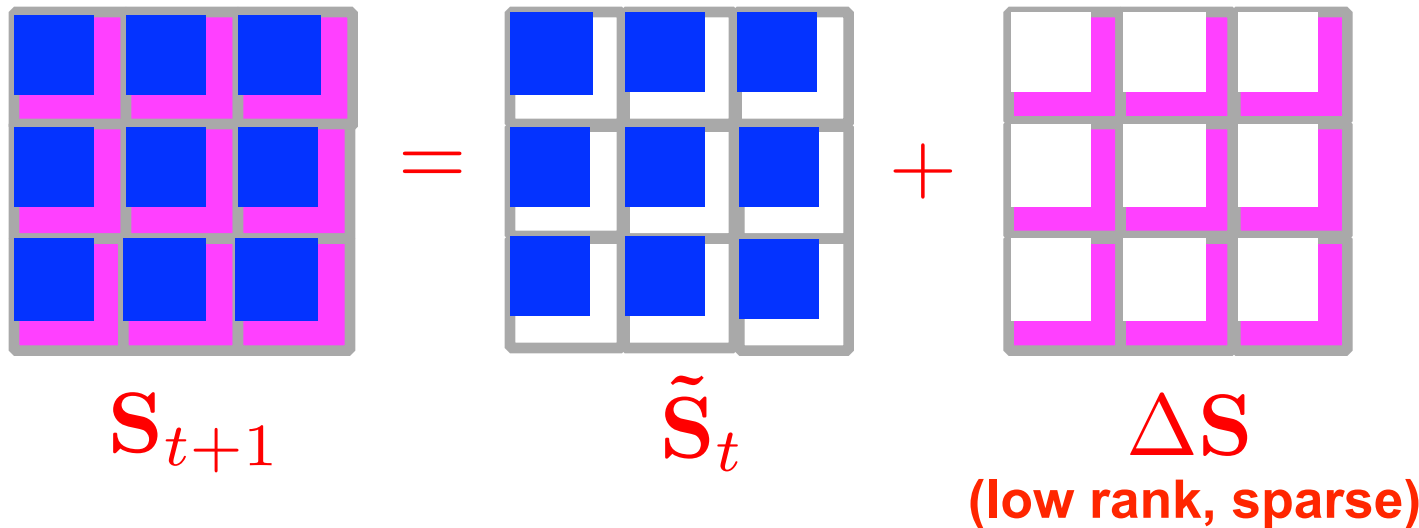
**Question:** how to avoid re-computing low-rank approx at each time step?

# iBall — Scale-up with Dynamic Update

- **Key idea #2:** Incrementally update the low rank structure of  $\mathbf{S}$

- **Details:**

white: zeros  
blue: old at  $t$   
pink: new at  $t+1$



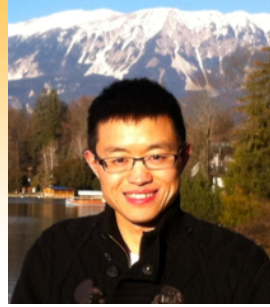
- **Complexity:**  $O(n^2 r) \rightarrow O((n+m)(r^2 + r'^2)), r \ll n$
- **Benefit:** avoid re-computing low-rank approx

# Roadmap

- Motivations
- Proposed Solutions: iBall
- **Experimental Results**
- Conclusions



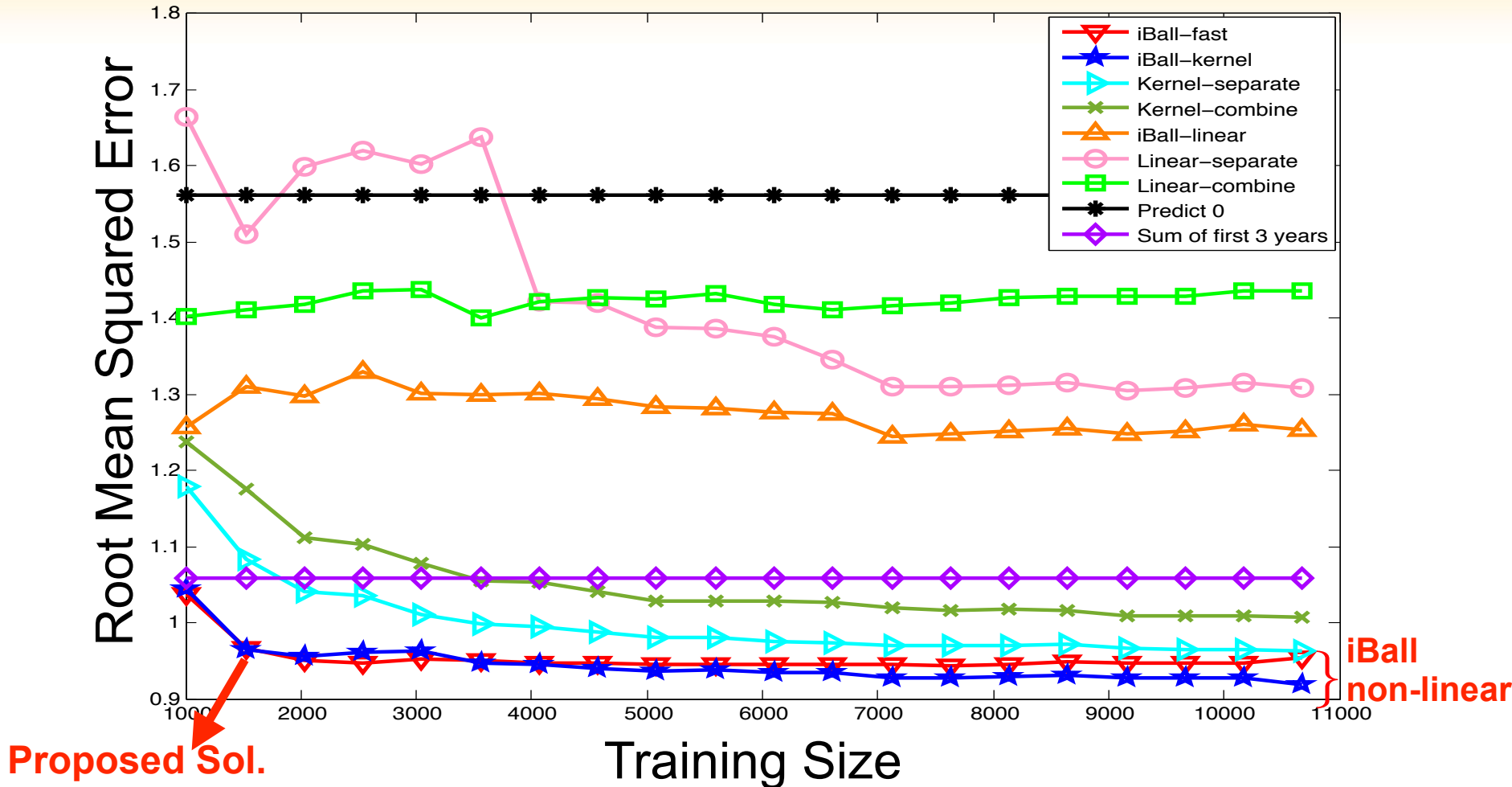
# Experiment Setup



- **Datasets:** AMiner<sup>1</sup> (2,243,976 papers, 1,274,360 authors, 8,882 venues)
- **Evaluation Metric:** Root Mean Squared Error (RMSE)
- **Evaluation Objects:**
  - ➔ Effectiveness
  - ➔ Efficiency

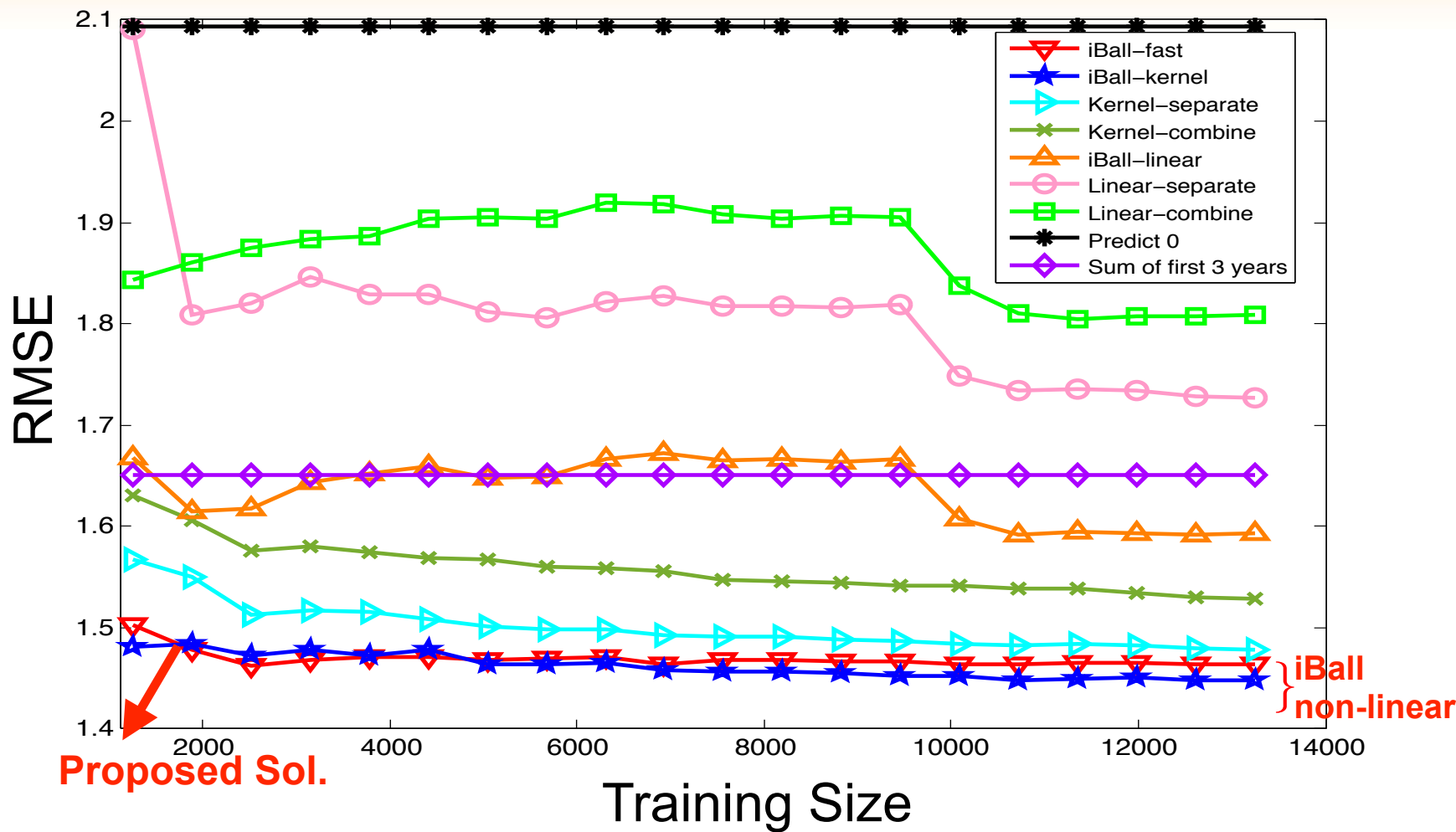
<sup>1</sup> <https://aminer.org/billboard/citation>

# Paper Citation Prediction Performance



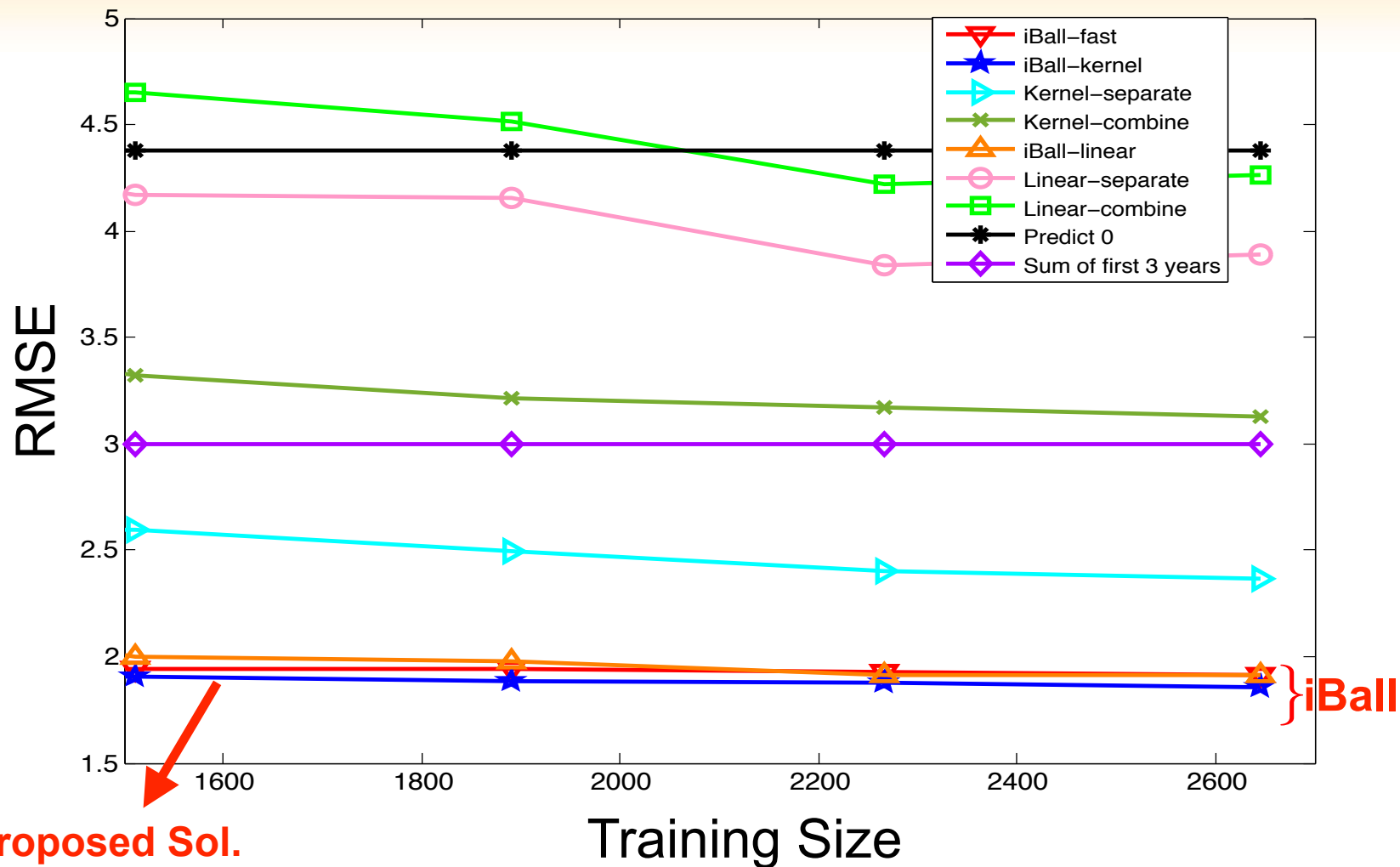
**Obs:** iBall family joint models better than separate versions

# Author Citation Prediction Performance



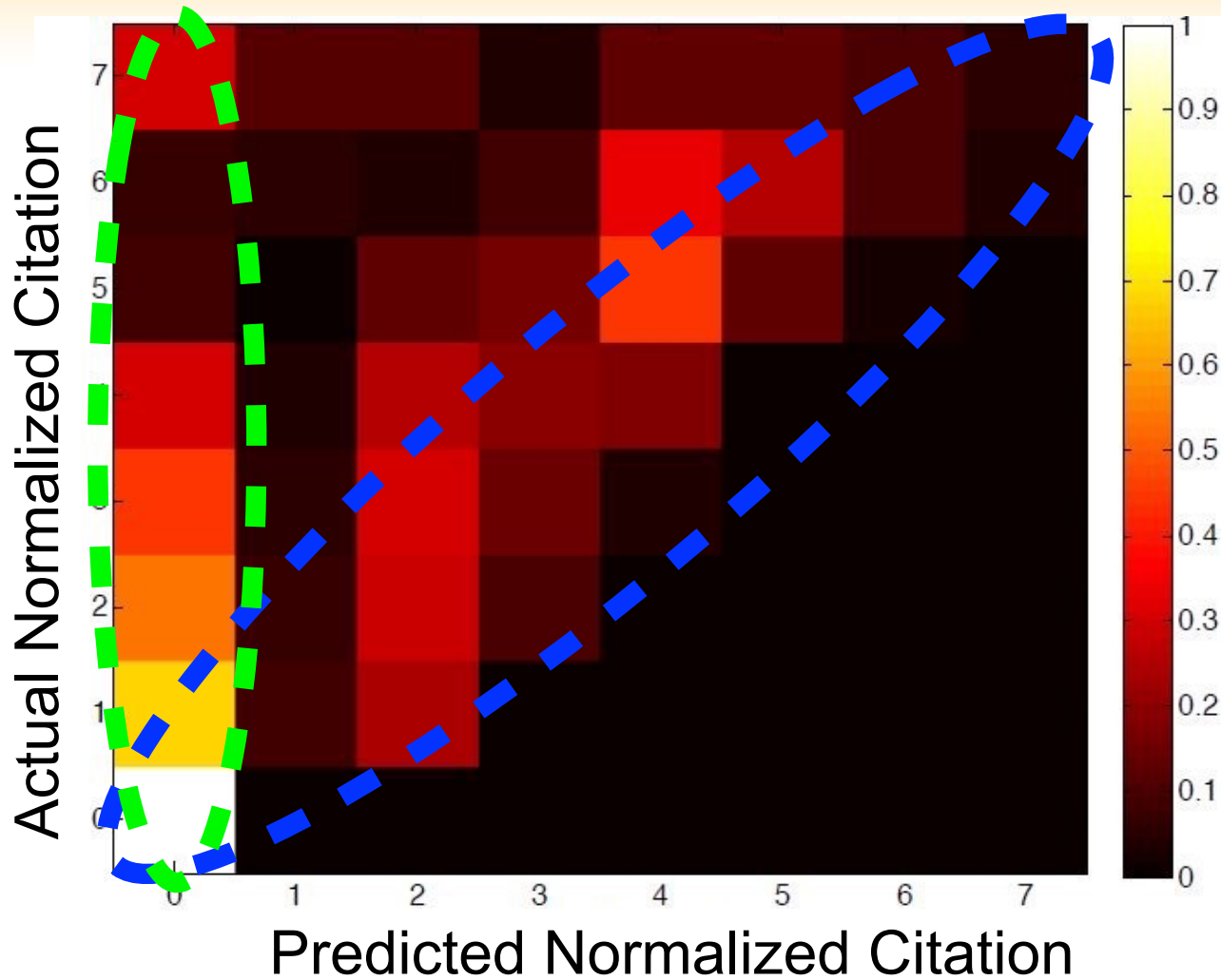
**Obs:** iBall family joint models better than separate versions

# Venue Citation Prediction Performance



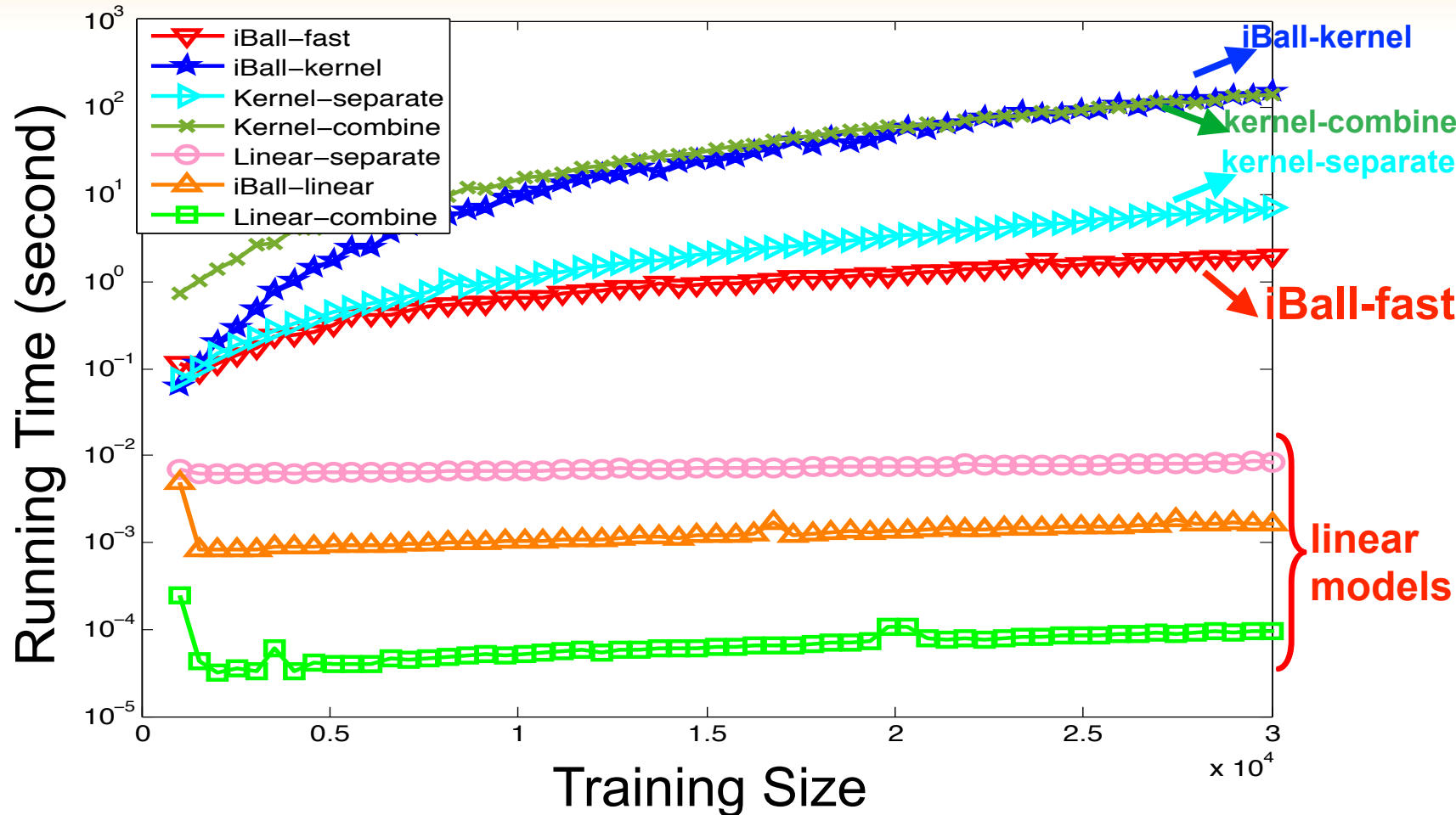
**Obs:** iBall family joint models better than separate versions

# Error Analysis



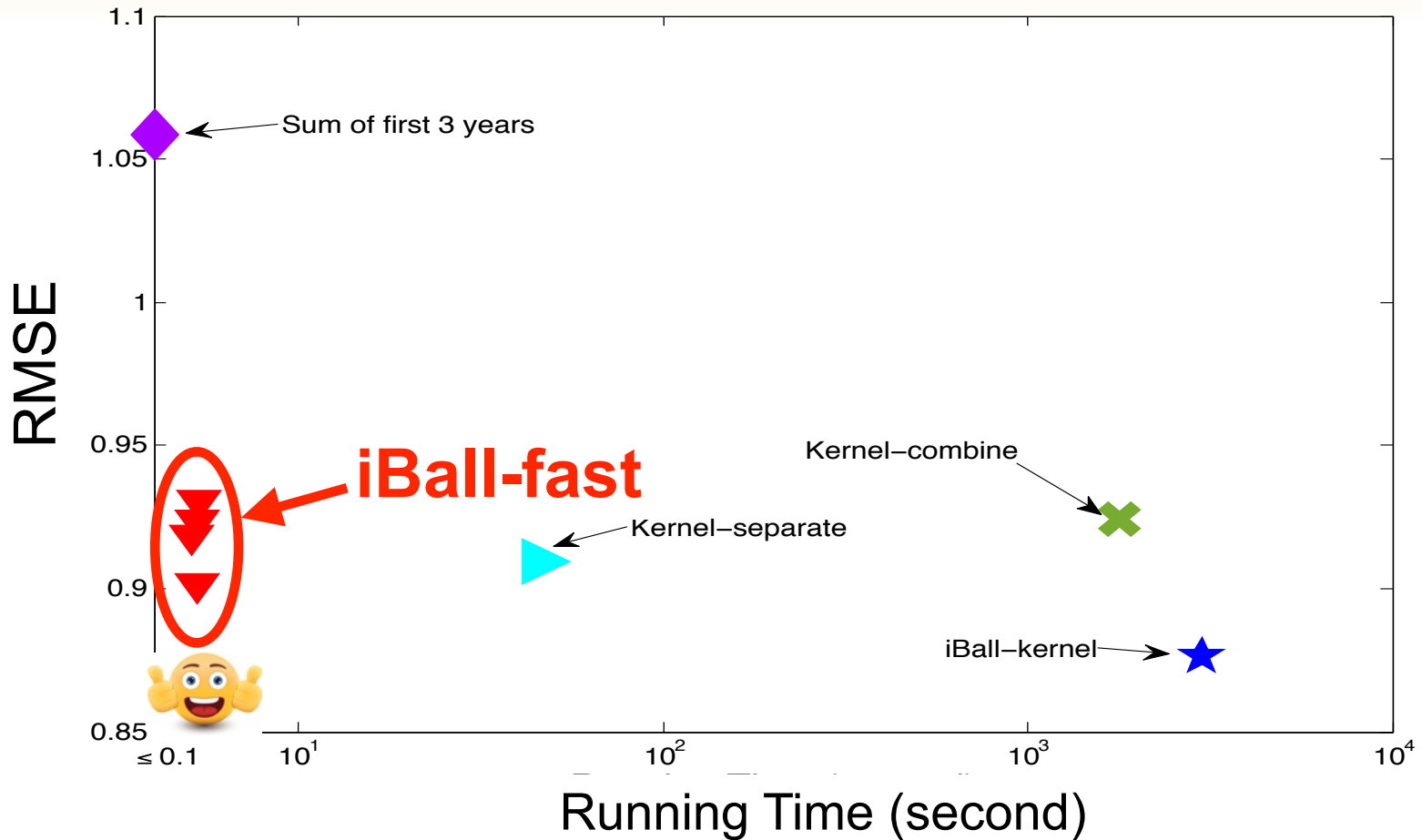
**Obs:** bright region at  $x = y$

# Running Time Comparison



**Obs:** iBall-fast outperforms other non-linear methods

# Quality vs. Speed



**Obs:** iBall-fast: good trade-off between quality and speed

# Roadmap

- Motivations
- Proposed Solutions: iBall
- Experimental Results
- **Conclusions**



# Conclusions

- **Goals:** predict long-term impact of scholarly entities
- **Solutions:** joint predictive model (**iBall**)

<i>Challenges</i>	<b>C1</b> <i>feature design</i>	<b>C2</b> <i>non-linearity</i>	<b>C3</b> <i>domain-heterogeneity</i>	<b>C4</b> <i>dynamics</i>
<i>Tactics</i>	<i>first 3 years' citation</i>	<i>kernel trick</i>	<i>domain consistency</i>	<i>low-rank approximation</i>

- **Results:**
  - iBall joint models better than separate versions
  - iBall-fast updates efficiently and accurately
- **More in paper:**
  - **correctness** and **error bound** analysis
  - **significance** and **sensitivity** tests