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

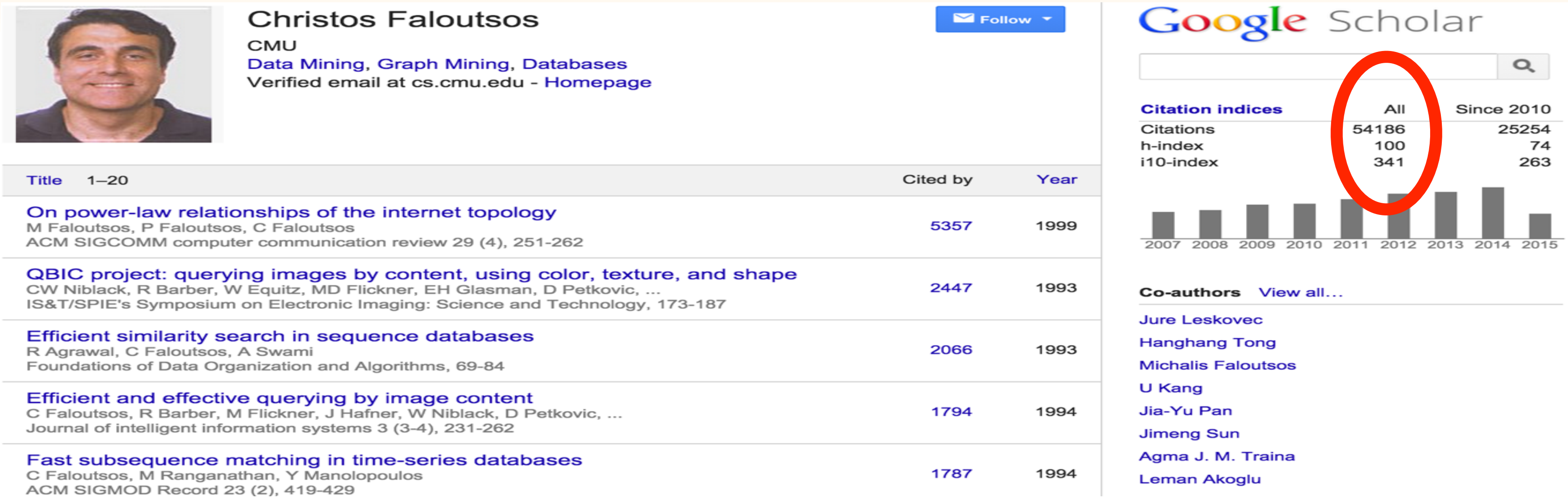
Presenter: **Liangyue Li (ASU)**

Joint work by

Liangyue Li
ASU

Hanghang Tong
ASU

High-impact Scientific Work



Important implications of high-impact scientific work:

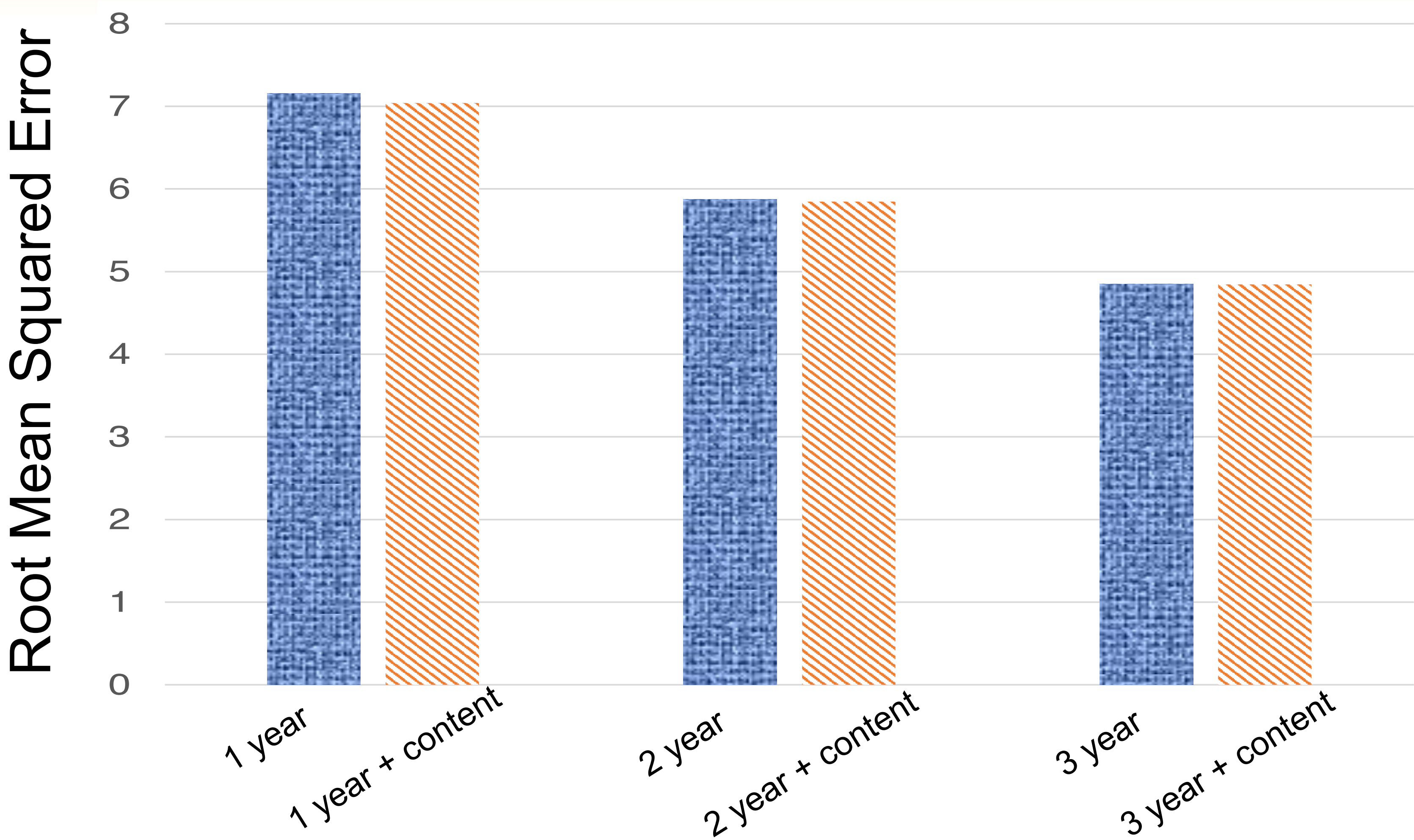
- personal career development
- recruitment search
- jurisdiction of research resources

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

Challenges

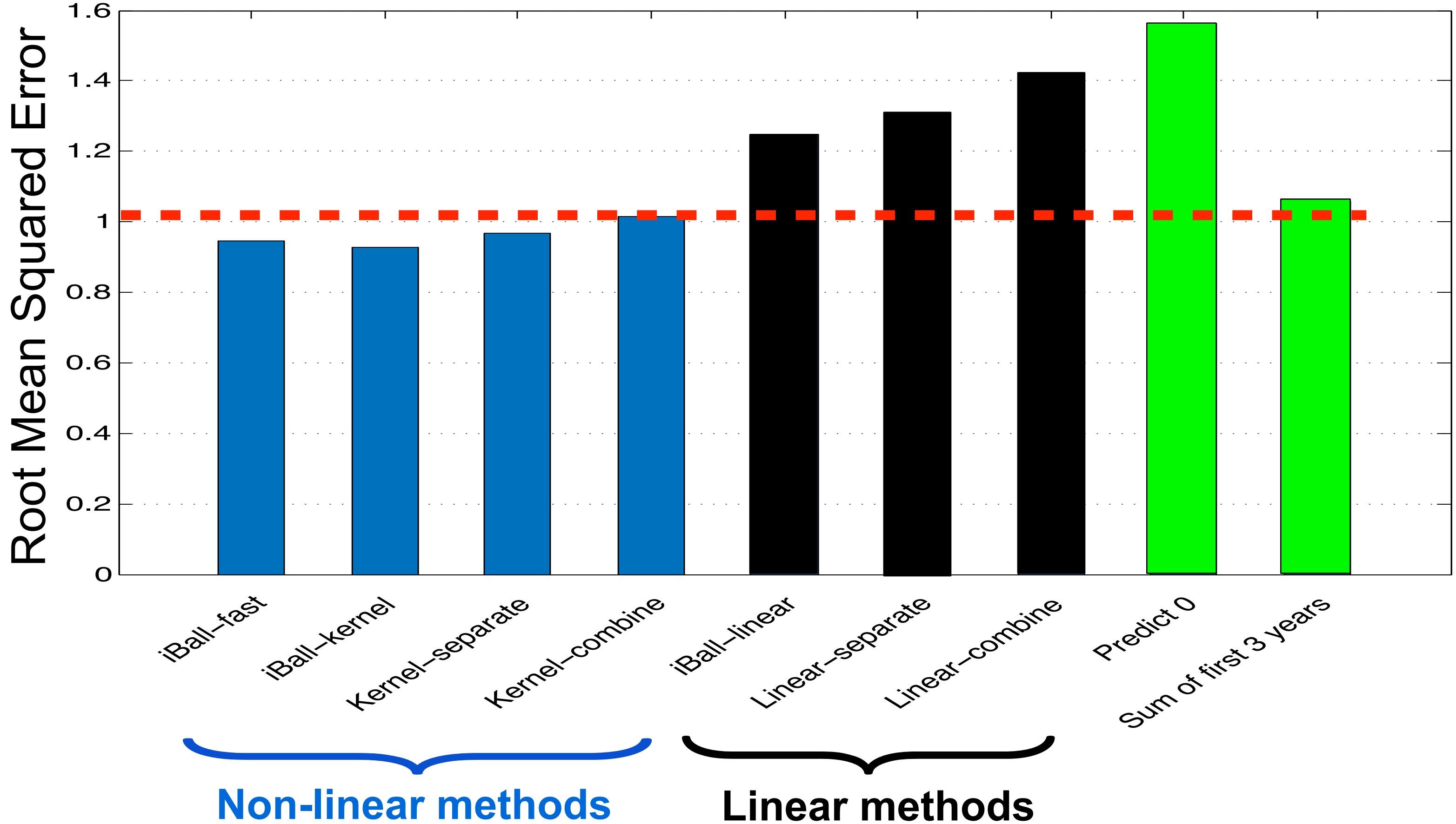
- C1: Scholarly feature design
- C2: Non-linearity
- C3: Domain heterogeneity
- C4: Dynamics

C1: Scholarly Feature Design



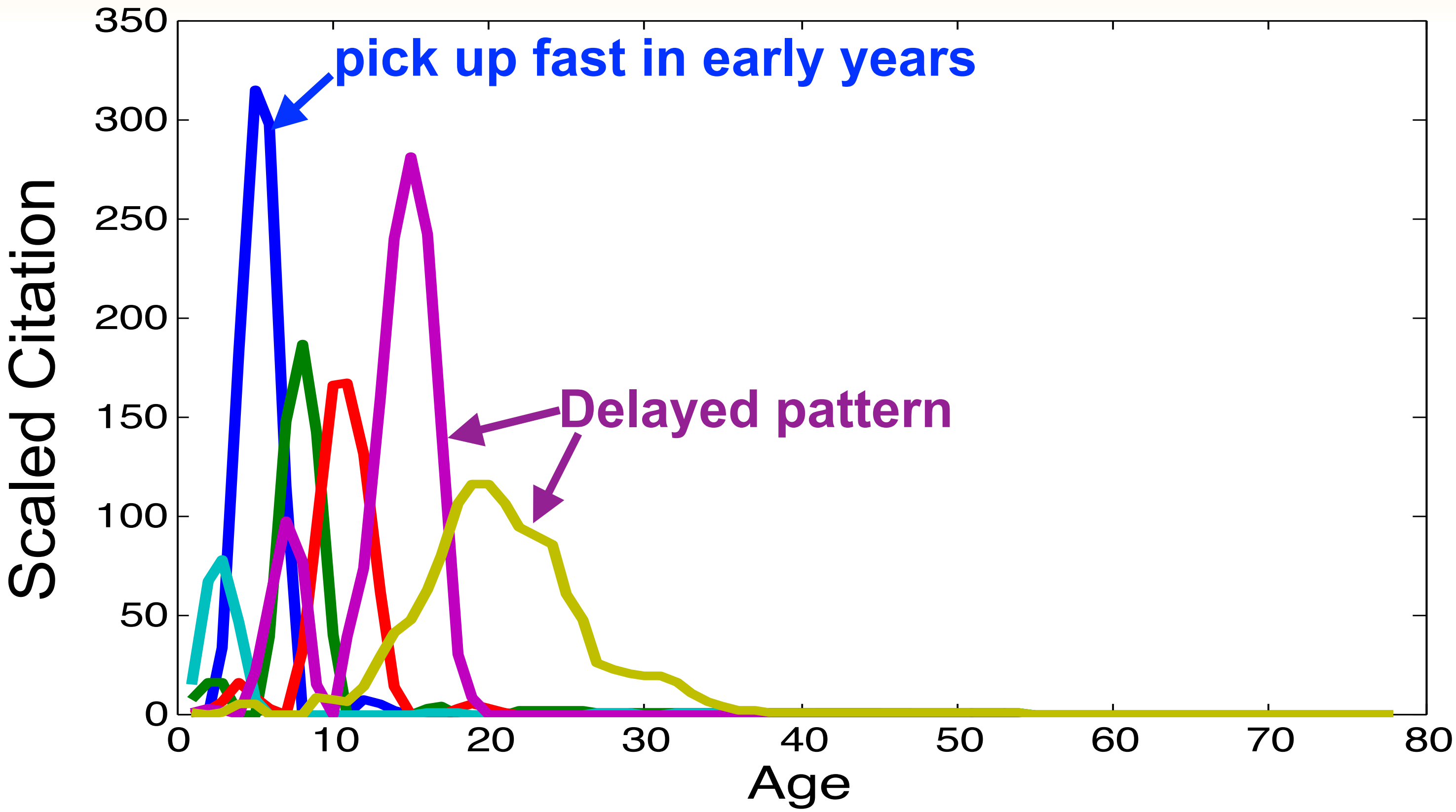
Obs: Adding content features brings little improvement

C2: Non-linearity



Obs: Non-linear methods outperform linear ones

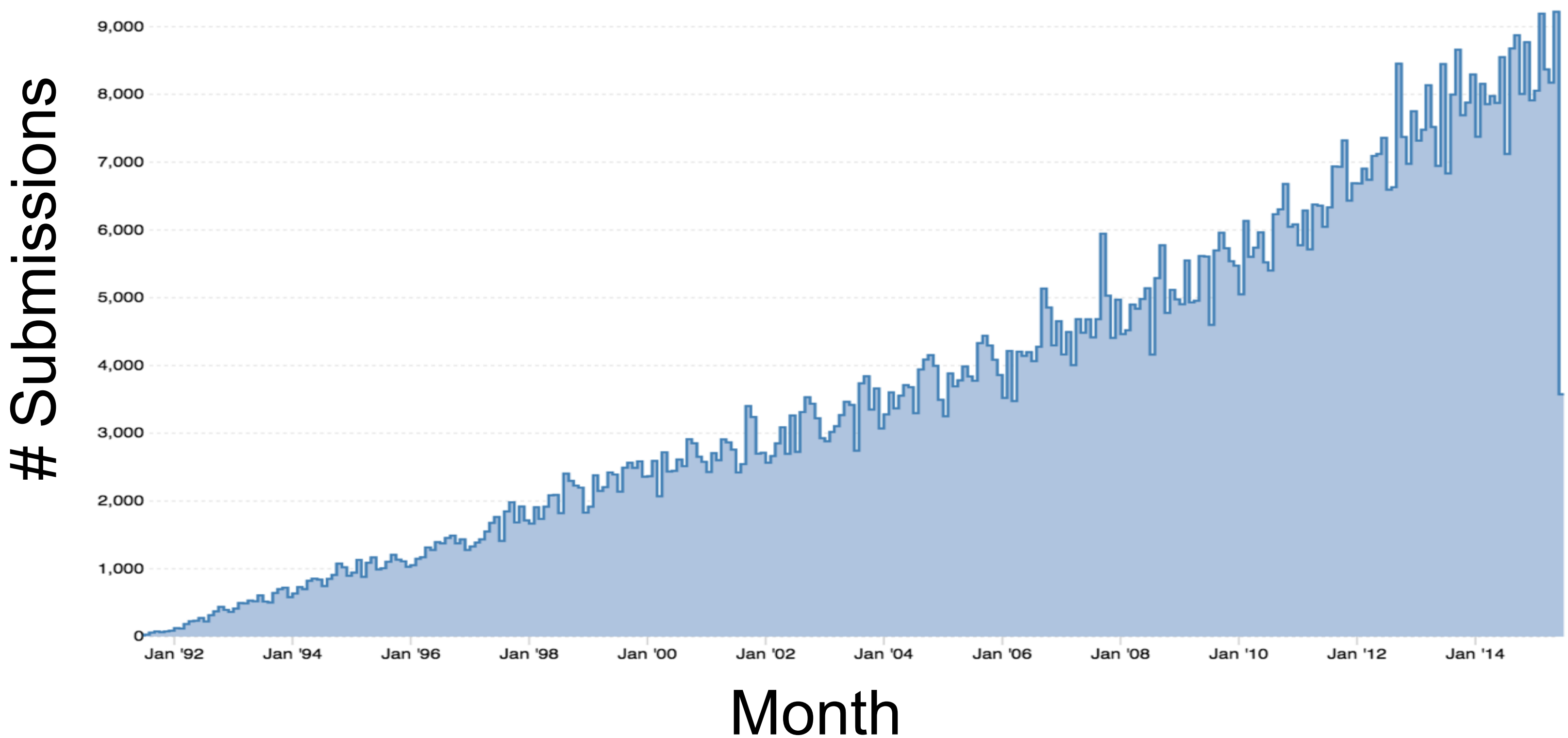
C3: Domain heterogeneity



Obs: Impact of scientific work from different domains behaves differently

C4: Dynamics

arXiv monthly submission rates

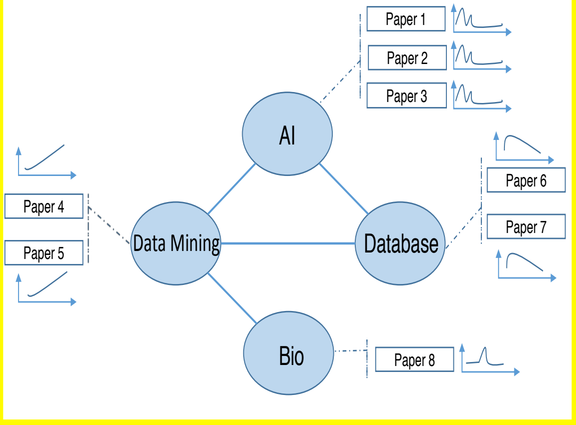


Question: How to quickly update the predictive model?

Roadmap

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

iBall — Formulations



- Optimization Formulation

Within-Domain Model

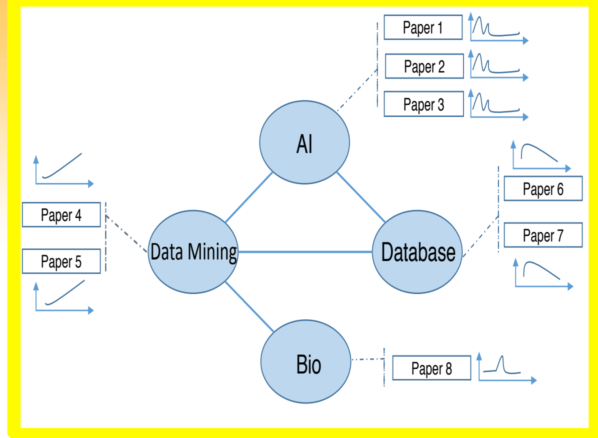
$$\min_{\mathbf{w}^{(i)}, i=1,\ldots,n_d} \sum_{i=1}^{n_d} \boxed{\mathcal{L}[f(\mathbf{X}^{(i)}, \mathbf{w}^{(i)}), \mathbf{Y}^{(i)}] + \lambda \sum_{i=1}^{n_d} \Omega(\mathbf{w}^{(i)})}$$
$$+ \boxed{\theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \mathbf{A}_{\mathbf{ij}} g(\mathbf{w}^{(\mathbf{i})}, \mathbf{w}^{(\mathbf{j})})}$$

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



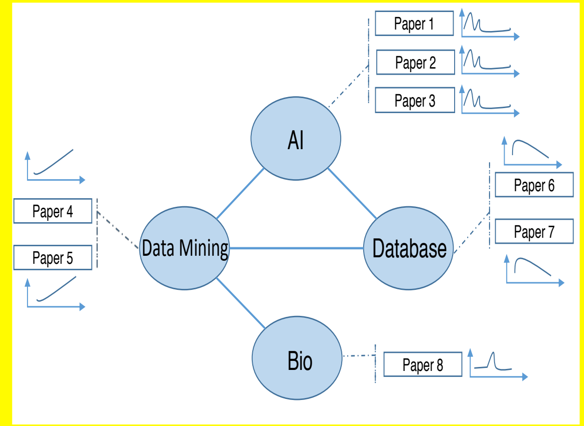
Details:

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

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:

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

Predicted output (domain *i* → domain *i*)

Predicted output (domain *j* →domain *i*)

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

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

iBall — Closed-form Solutions

- Closed-form Solution

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

➡ iBall — linear:

$$\mathbf{w} = [\mathbf{w}^{(\mathbf{1})}; \ldots ; \mathbf{w}^{(\mathbf{n_d})}] \qquad \mathbf{Y} = [\mathbf{X}^{(\mathbf{1})\prime}\mathbf{Y}^{(\mathbf{1})}; \ldots ; \mathbf{X}^{(\mathbf{n_d})\prime}\mathbf{Y}^{(\mathbf{n_d})}]$$

$$\mathbf{S} = \begin{array}{cc} \begin{array}{ccc} & \text{i-th block column} & \text{j-th block column} \end{array} & \\ \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \mathbf{X}^{(\mathbf{i})\prime}\mathbf{X}^{(\mathbf{i})} + (\theta \sum_{j=1}^{n_d} \mathbf{A_{ij}} + \lambda)\mathbf{I} & -\theta\mathbf{A_{ij}}\mathbf{I} \\ \cdots & \cdots & \cdots \end{bmatrix} & \begin{array}{l} \text{i-th block} \\ \text{row} \end{array} \end{array}$$

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:

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

$$\mathbf{S} = \begin{array}{cc} \begin{array}{ccc} & \text{i-th block column} & \text{j-th block column} \end{array} & \\ \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & (1+\theta\sum_{j=1}^{n_d}\mathbf{A_{ij}})\mathbf{K^{(i)}}+\lambda\mathbf{I} & -\theta\mathbf{A_{ij}}\mathbf{K^{(ij)}} \\ \cdots & \cdots & \cdots \end{bmatrix} & \begin{array}{l} \text{i-th block} \\ \text{row} \end{array} \end{array}$$

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 **S** by low-rank approx
- **Details:**

$$\mathbf{S}_{t+1} \approx \mathbf{U}_{t+1}\mathbf{\Lambda}_{t+1}\mathbf{U}'_{t+1}$$

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

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{S}_{t+1}^{-1}\mathbf{Y}_{t+1} \\ &= \mathbf{U}_{t+1}\mathbf{\Lambda}_{t+1}^{-1}\mathbf{U}'_{t+1}\mathbf{Y}_{t+1} \end{aligned}$$

(Overall: $O(nr)$)

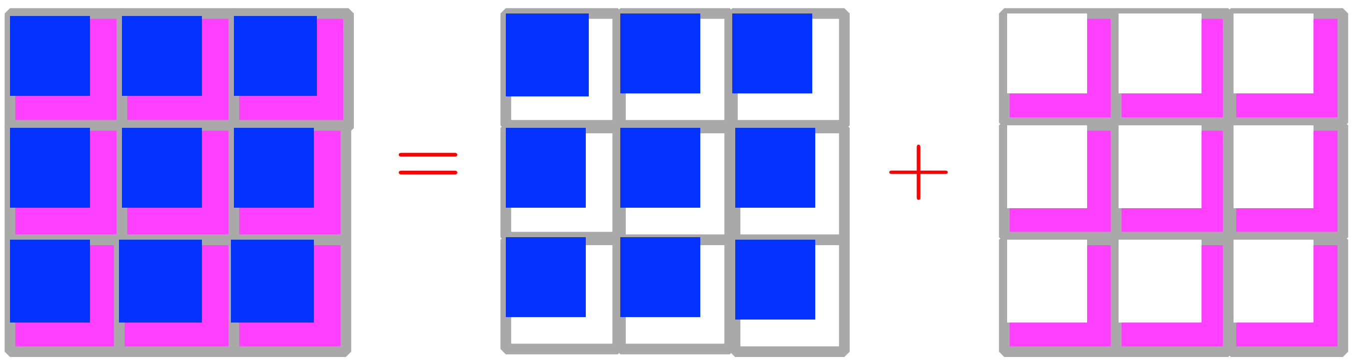
- **Complexity:** $O(n^3) \rightarrow O(n^2 r + nr)$
- **Benefit:** avoid matrix inverse

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 **S**
- **Details:**

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



$\mathbf{S}_{t+1}$ $=$ $\tilde{\mathbf{S}}_t$ $+$ $\Delta\mathbf{S}$

(low rank, sparse)

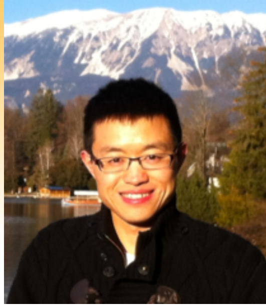
- **Complexity:** $O(n^2 r) \to O((n+m)(r^2 + r'^2)), r \ll n$
- **Benefit:** avoid re-computing low-rank approx

Liangyue Li, Hanghang Tong, Yanghua Xiao, Wei Fan. Cheetah: Fast Graph Kernel Tracking on Dynamic Graphs.(SDM), 2015.

Roadmap

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

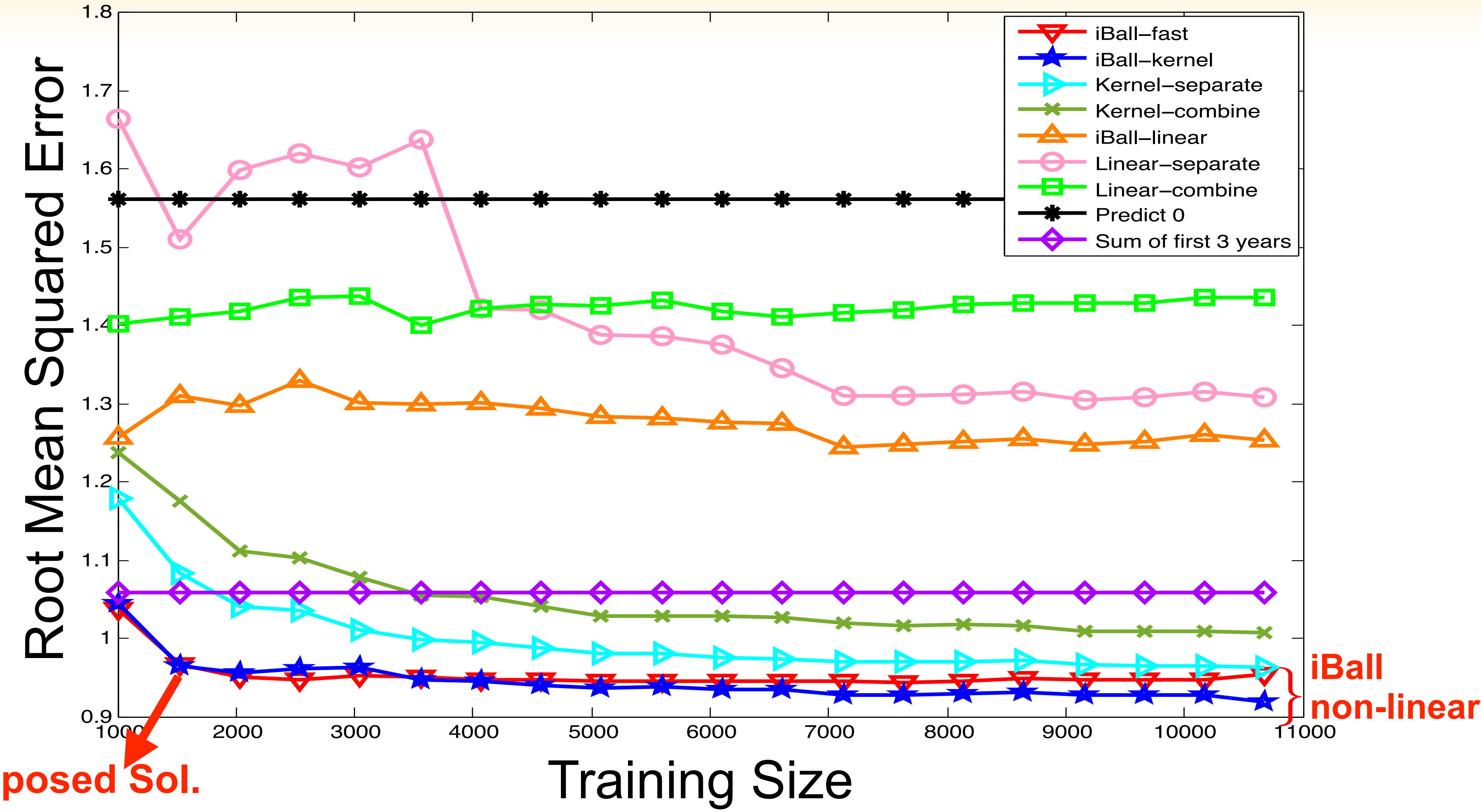
Experiment Setup



- Datasets: AMiner[1] (2,243,976 papers, 1,274,360 authors, 8,882 venues)
- Evaluation Metric: Root Mean Squared Error (RMSE)
- Evaluation Objects:
  - ➡Effectiveness
  - ➡Efficiency

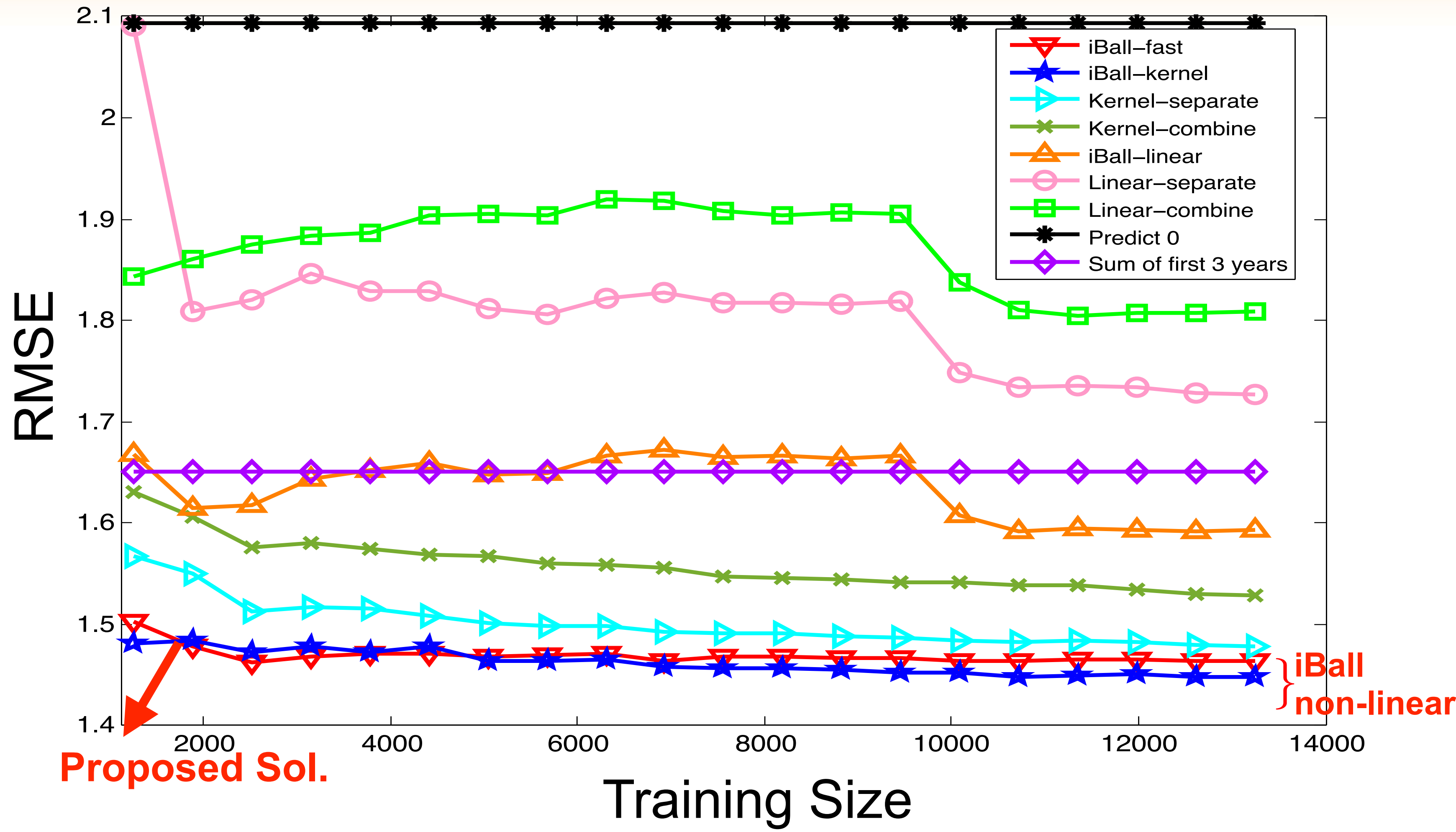
[1] https://aminer.org/billboard/citation

Paper Citation Prediction Performance



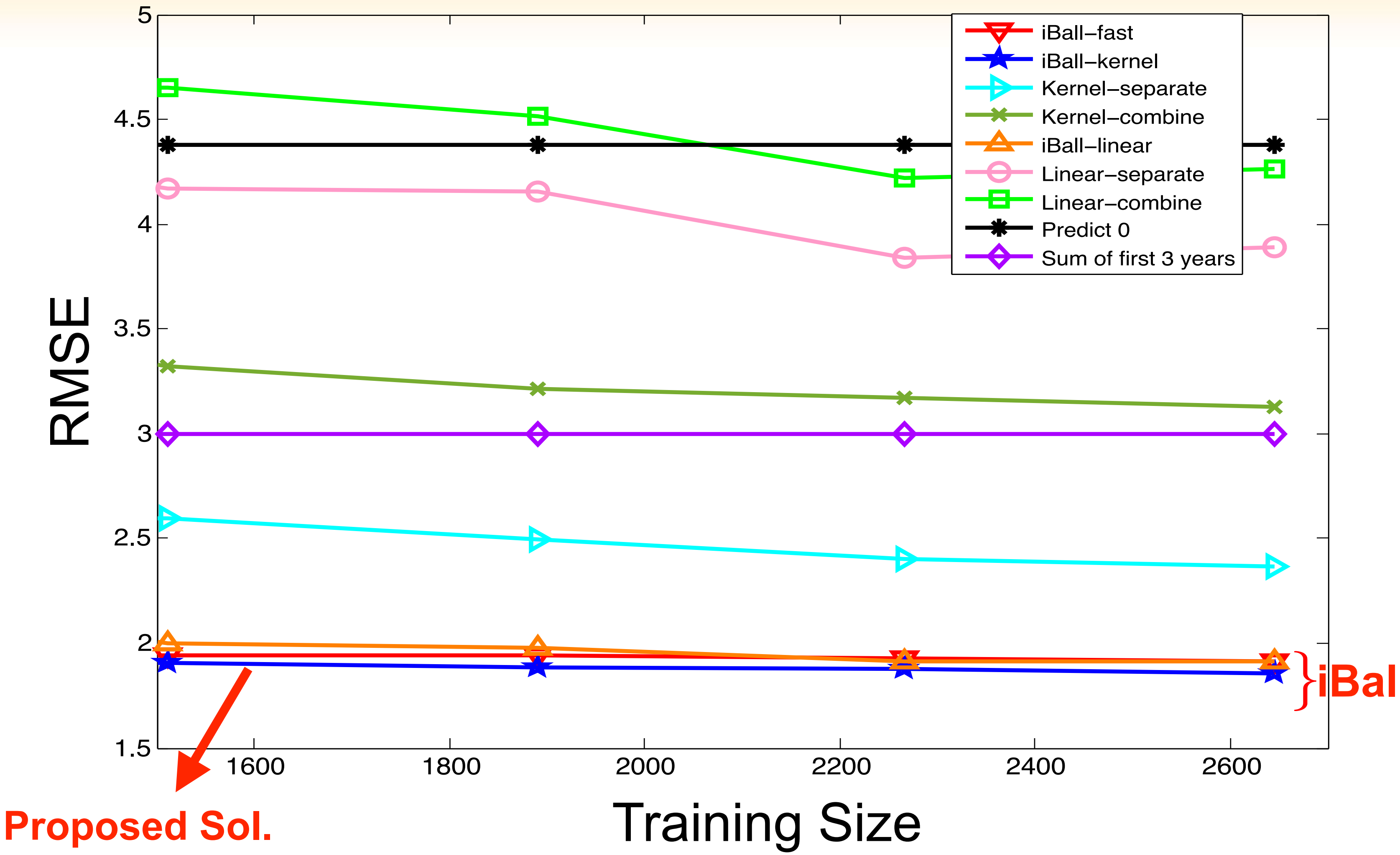
Obs: iBall family joint models better than separate versions

Author Citation Prediction Performance



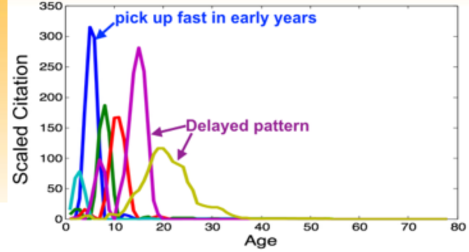
Obs: iBall family joint models better than separate versions

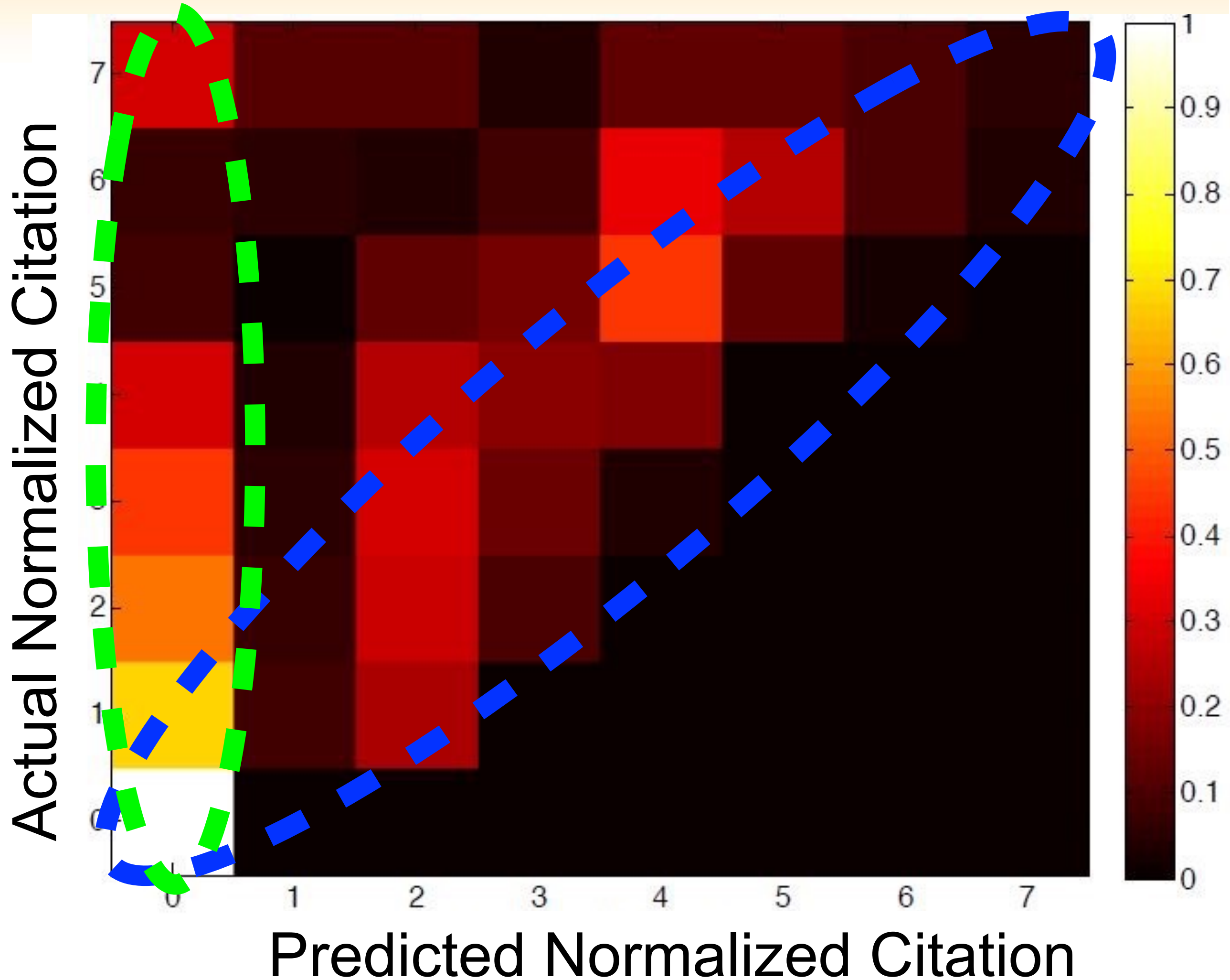
Venue Citation Prediction Performance



Obs: iBall family joint models better than separate versions

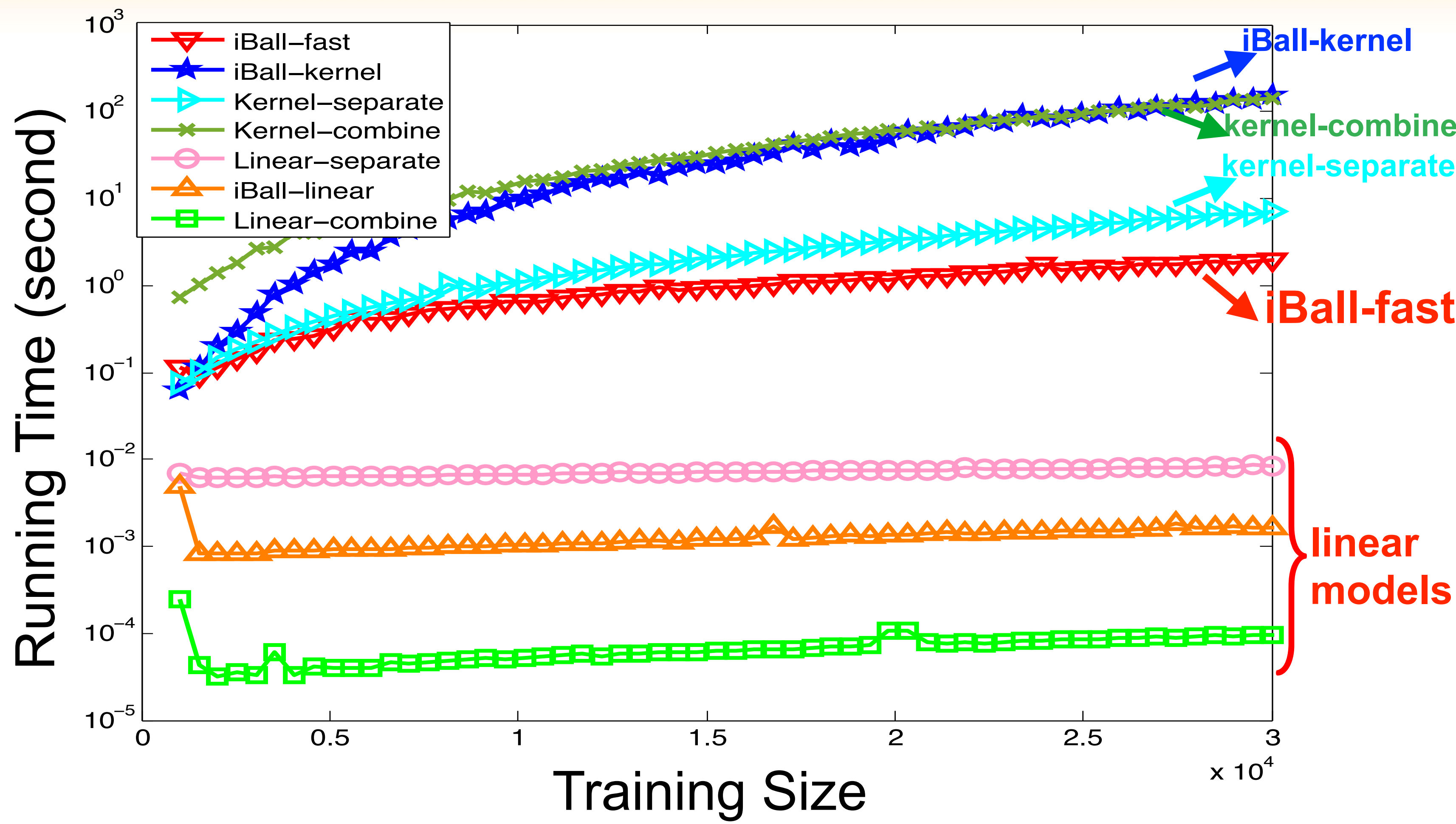
Error Analysis





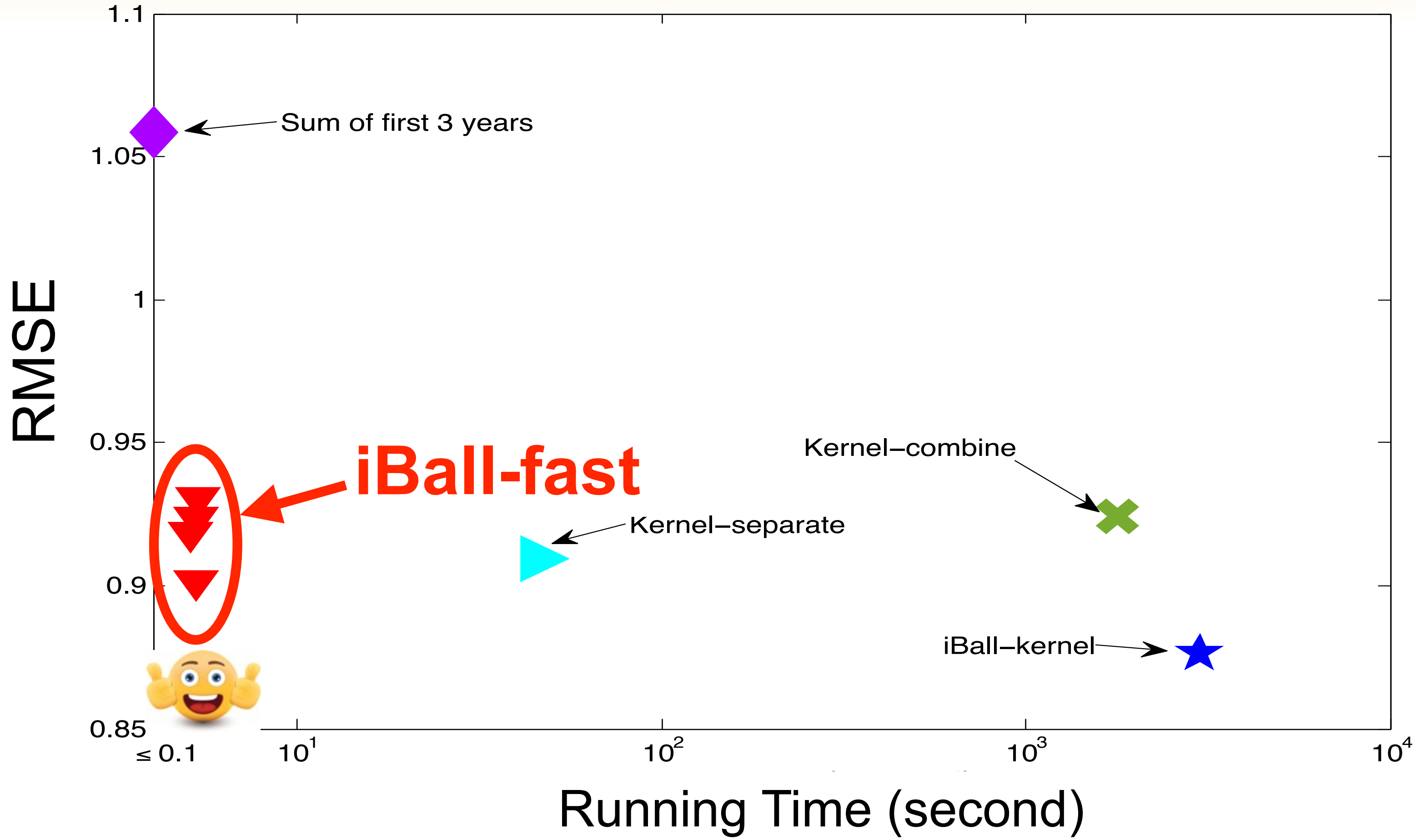
Obs: bright region at x = y

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**)

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

- **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