Towards a Stratified Learning Approach to Predict Future Citation Counts

Tanmoy Chakraborty

Google India PhD Fellow

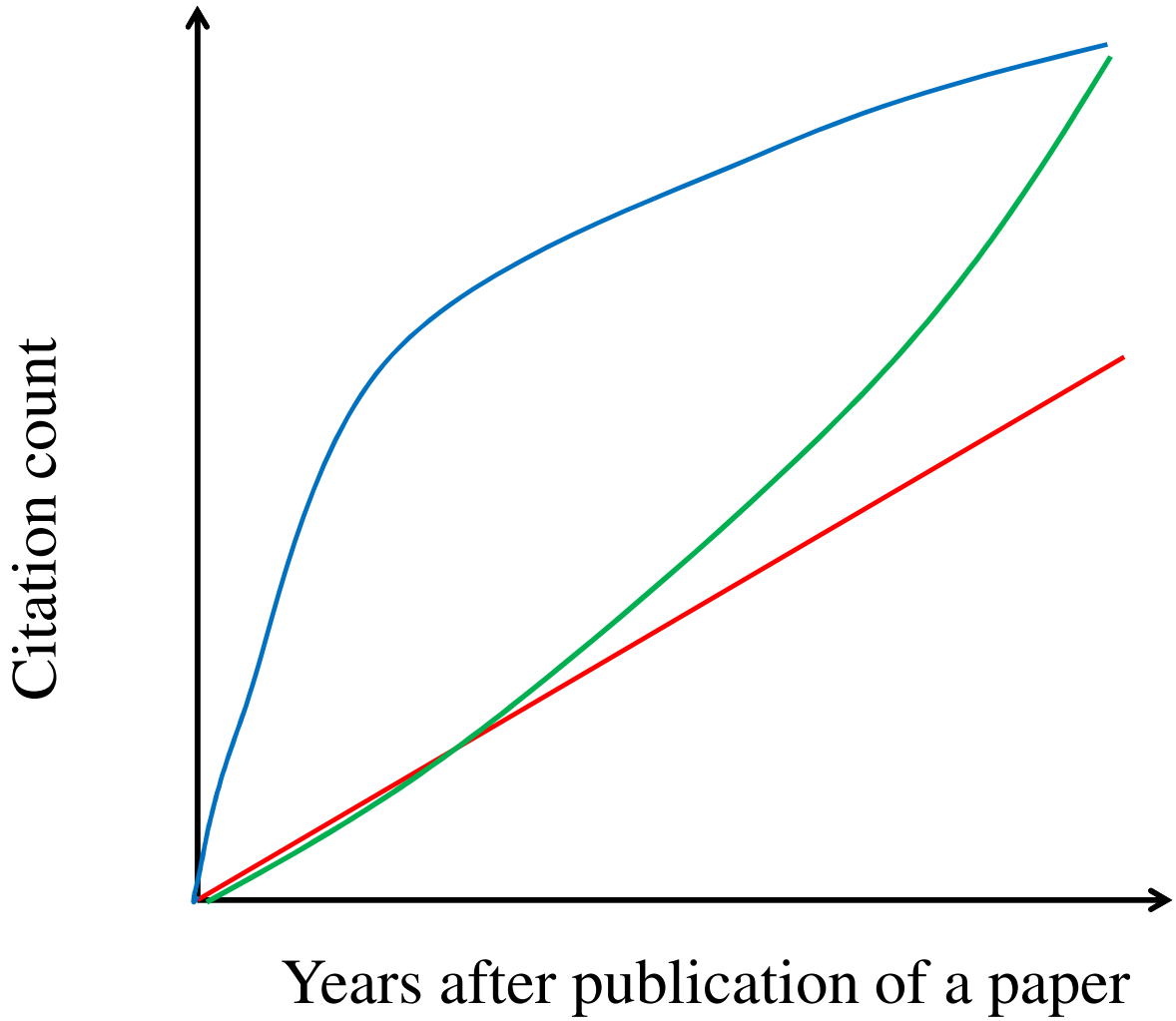
IIT Kharagpur, India

Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, Animesh Mukherjee

Dept. of CSE, IIT Kharagpur, India

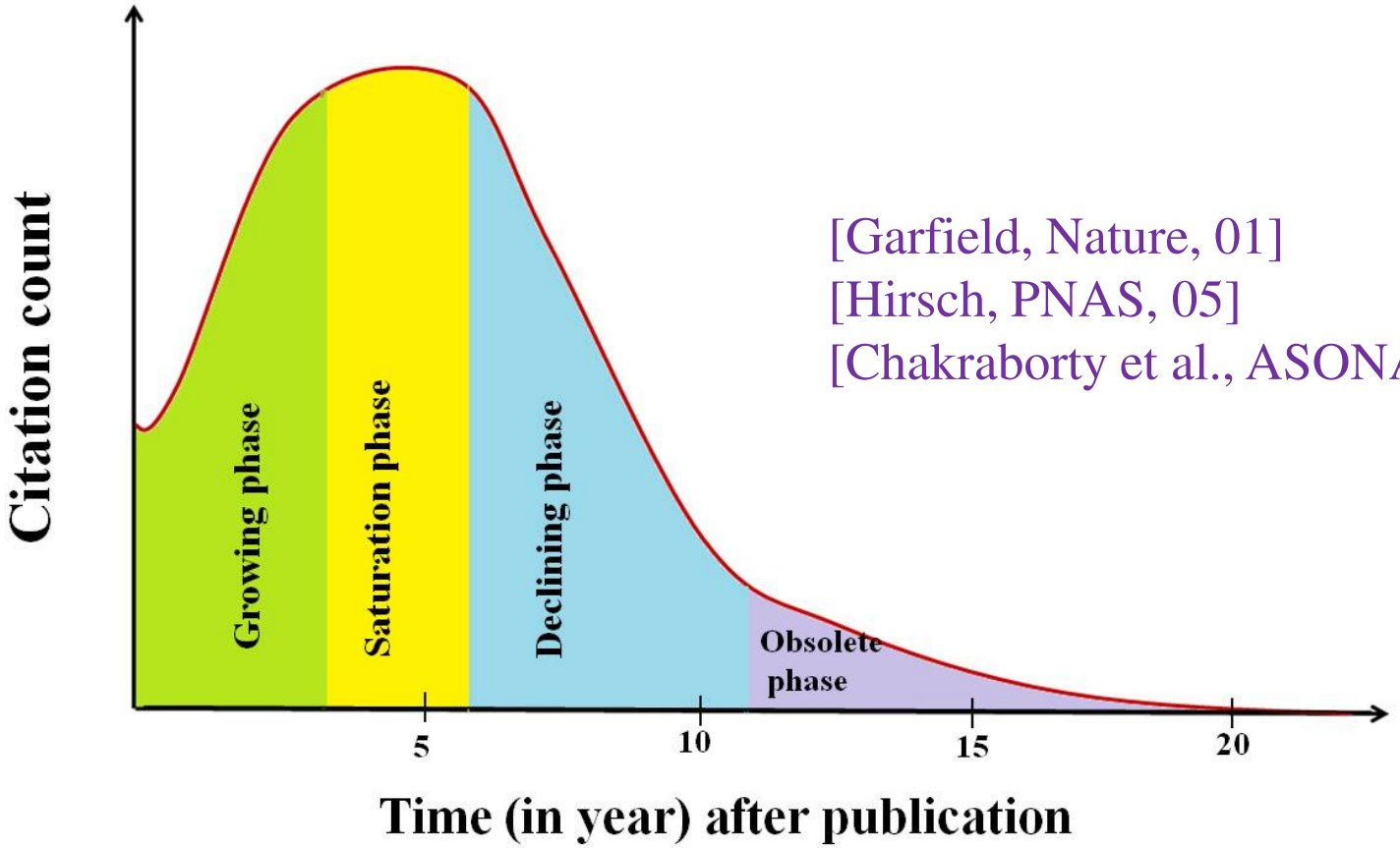
Digital Libraries, September 8-12, 2014

Citation Patten over the Year



Citation Profile of An Article

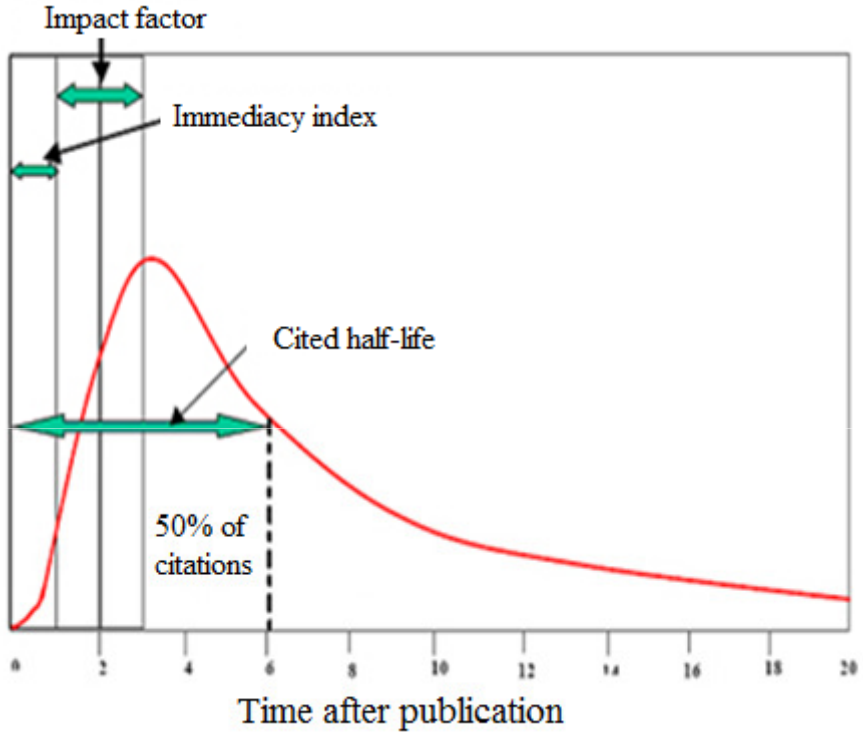
Common consensus about the growth of citation count of a paper over time after publication



[Garfield, Nature, 01]
[Hirsch, PNAS, 05]
[Chakraborty et al., ASONAM, 13]

Bibliometrics

- Journal Impact factor
- Immediacy factor
- Altmetric
- 5 years Impact factor



This observation was drawn from the analysis of a very limited set of publication data

[Kulkarni et al., PLoS ONE, 07] [Callaham et al., JAMA, 02]

Publication Universe

- Crawled entire Microsoft Academic Search
- Papers in Computer Science domain
- Basic preprocessing

Basic Statistics of papers from 1960-2010	Values
Number of valid entries	3,473,171
Number of authors	1,186,412
Number of unique venues	6,143
Avg. number of papers per author	5.18
Avg. number of authors per paper	2.49

Publication Universe (Contd…)

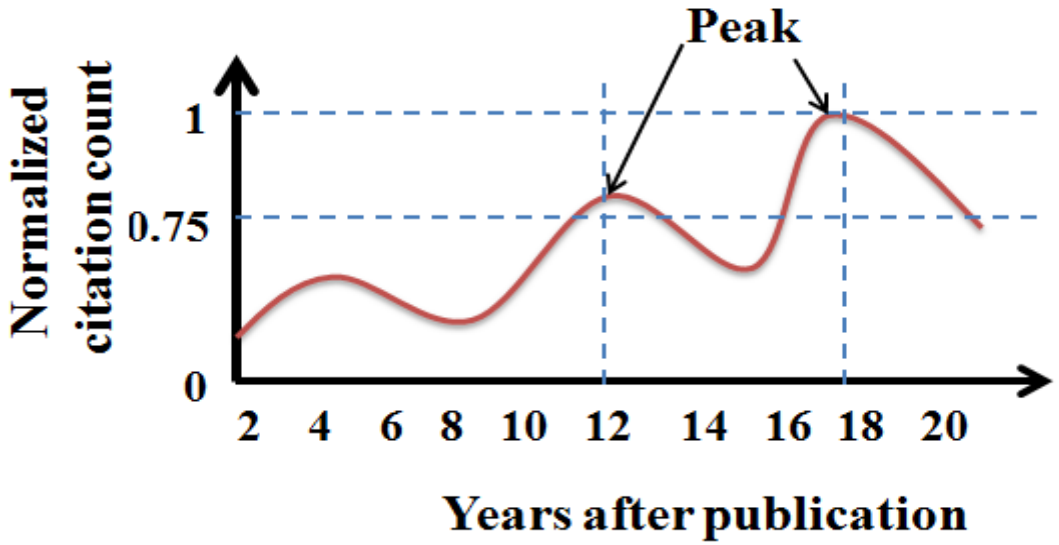
Available Metadata
Title
Unique ID
Named entity disambiguated authors' name
Year of publication
Named entity disambiguated publication venue
Related research field(s)
References
Keywords
Abstract

Available @ http://cnerg.org

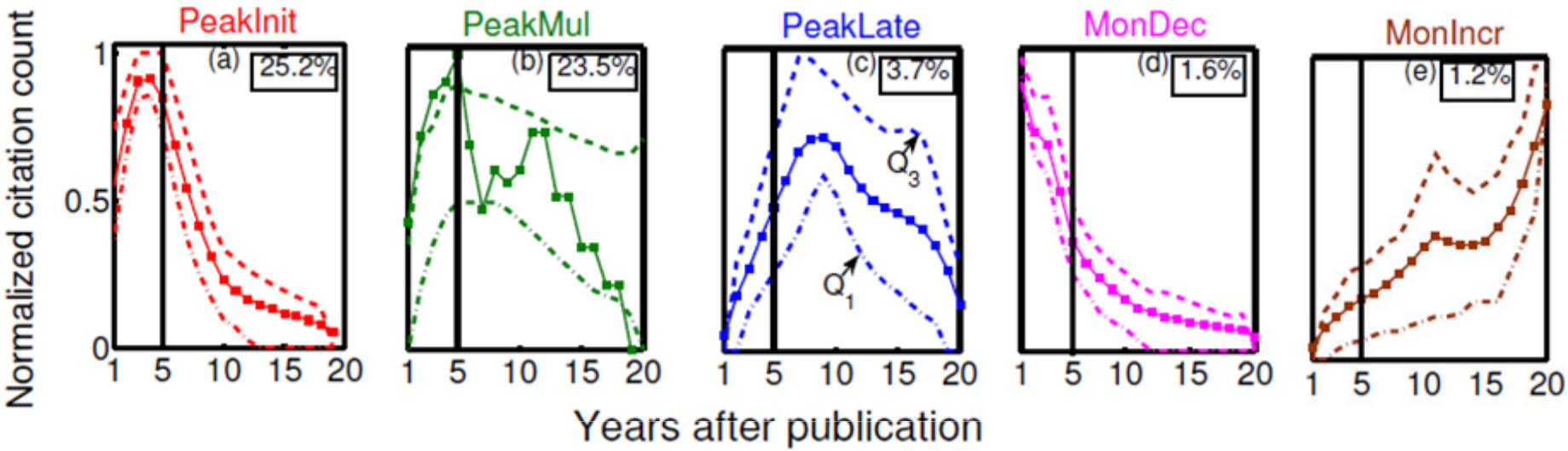
Citation Profile Analysis

An exhaustive analysis of the citation profiles

- Papers having at least 10 years history
- Scale the entries of the citation profile between 0-1
- Use peak-detection heuristics
  - Each peak should be at least 75% of the max peak
  - Two consecutive peaks should be separated at least 3 yrs



Six Universal Citation Profiles



Q_1 and Q_3 represent the first and third quartiles of the data points respectively.

Another category: **'Oth'** => having less than one citation (on avg) per year

Application:
Future Citation Count Prediction

Problem Definition

Citation counts:
Given the set of scientific articles D, the citation counts ($C_T(.)$) of an article $d \in D$ is defined as:

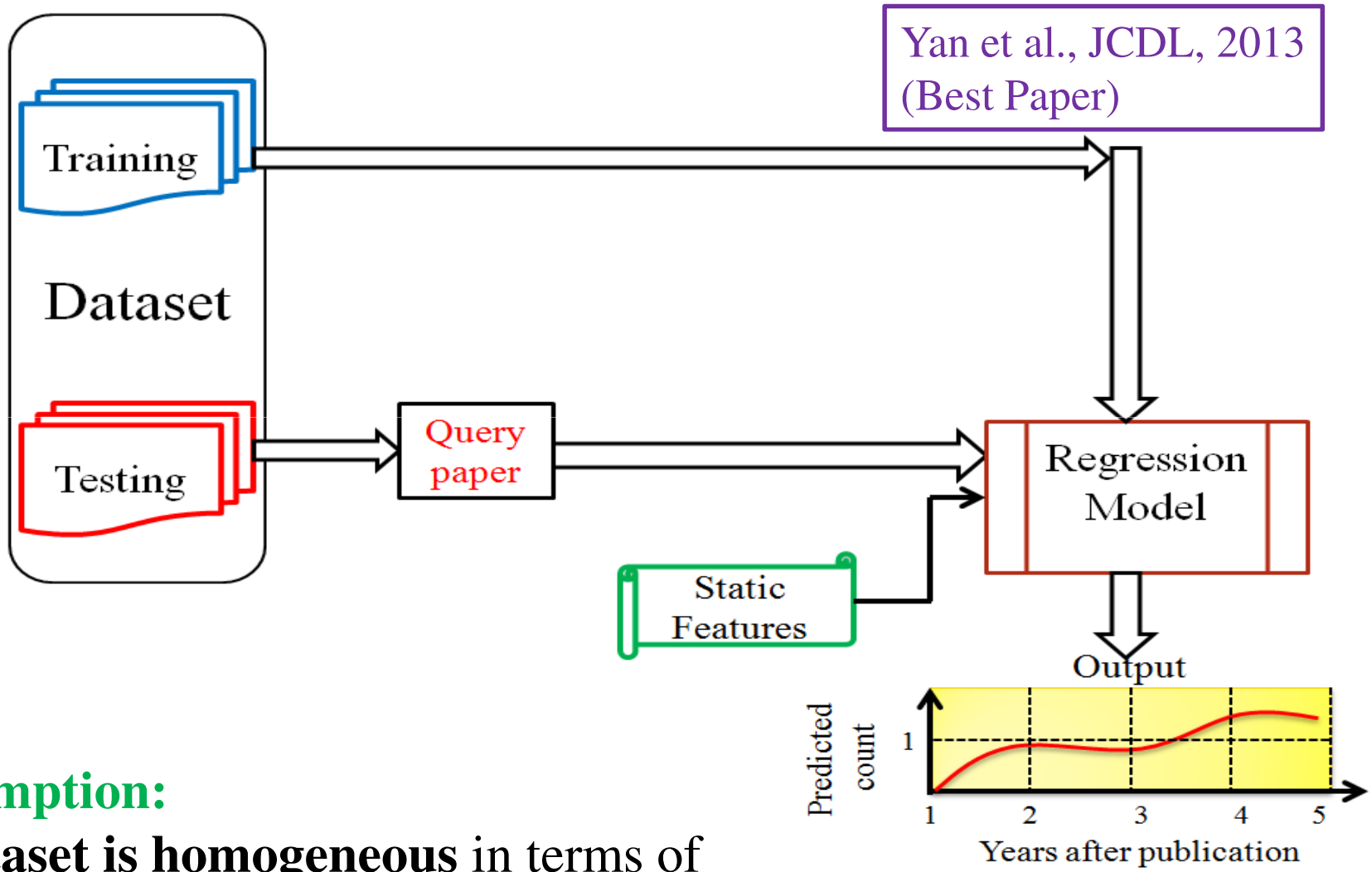
$$citing(d) = \{d' \in D : d' \ cites \ d\}$$
$$C_T(d) = |citing(d)|$$

Learning task: Given a set of features $F = \{f_1, f_2, ..., f_n\}$, our goal is to learn a predictive function ψ to predict the citation counts of an article d after a give time period Δt of its publication. Formally, this can be written as:

$$\psi(d|F, \Delta t) \rightarrow C_T(d|\Delta t)$$

we consider $\Delta t \in \{1, 5\}$

Traditional Framework



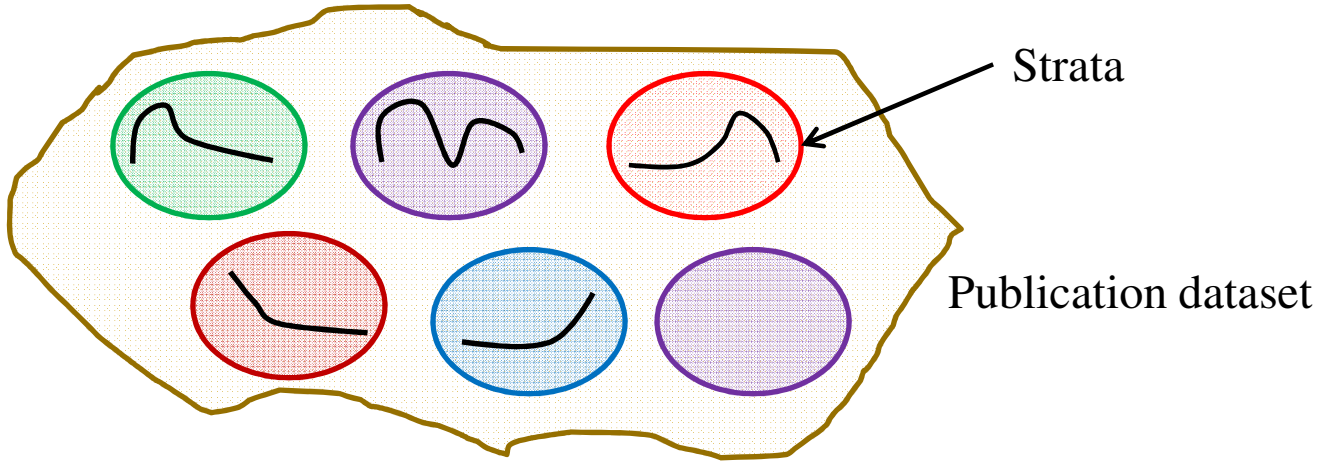
Assumption:

Dataset is homogeneous in terms of citation profile

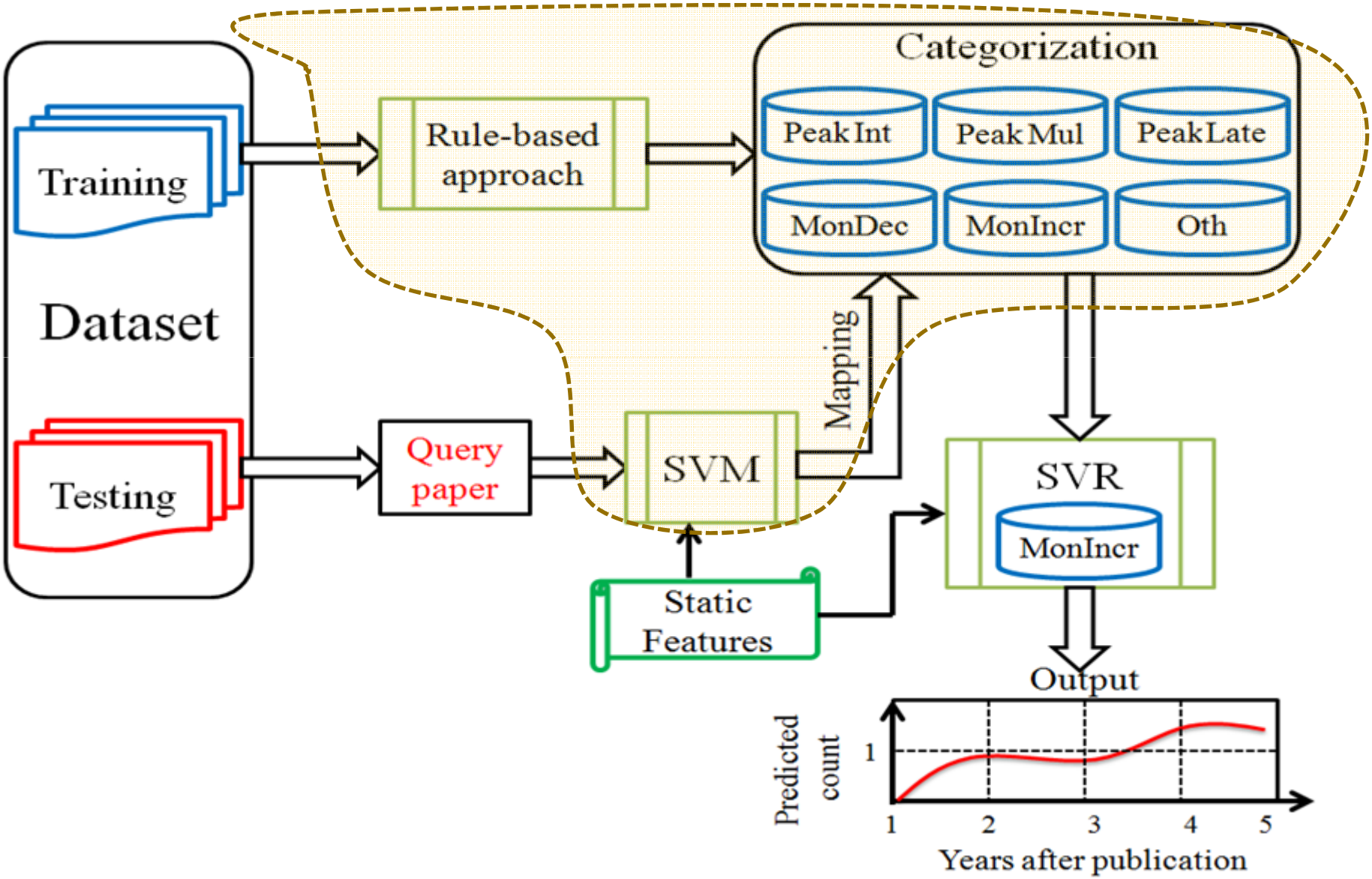
Stratified Learning

- **Stratification** is the process of dividing members of the population into **homogeneous subgroups (strata)** before sampling.

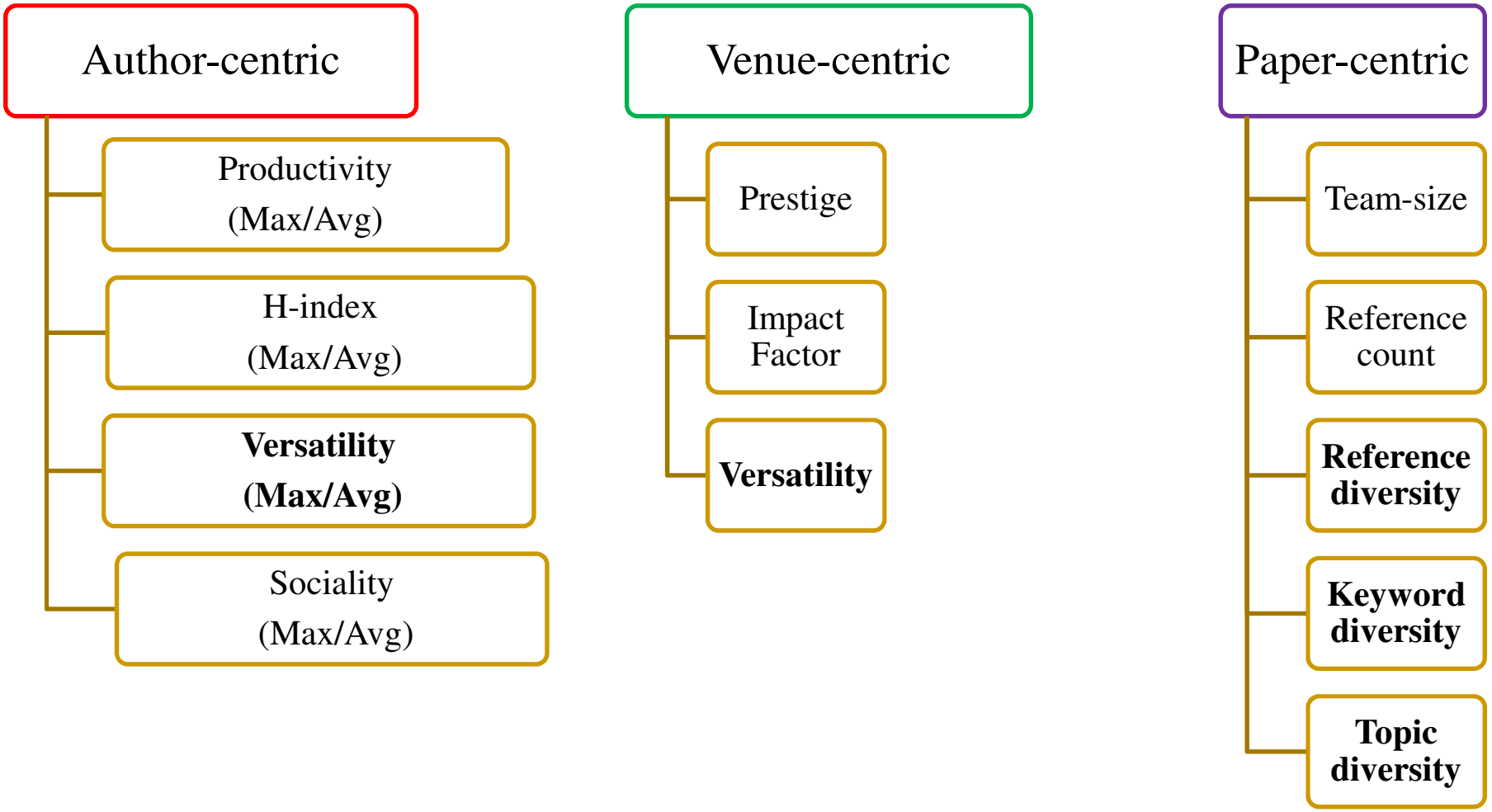
- The **strata** should be mutually exclusive
  - Every element in the population must be assigned to only one stratum



Our Framework: 2-stage Model



Static Features



Performance Evaluation

(i) Coefficient of determination (R^2)

The more, the better

(ii) Mean squared error (θ)

The less, the better

(iii) Pearson correlation coefficient (ρ)

The more, the better

Performance of SVM

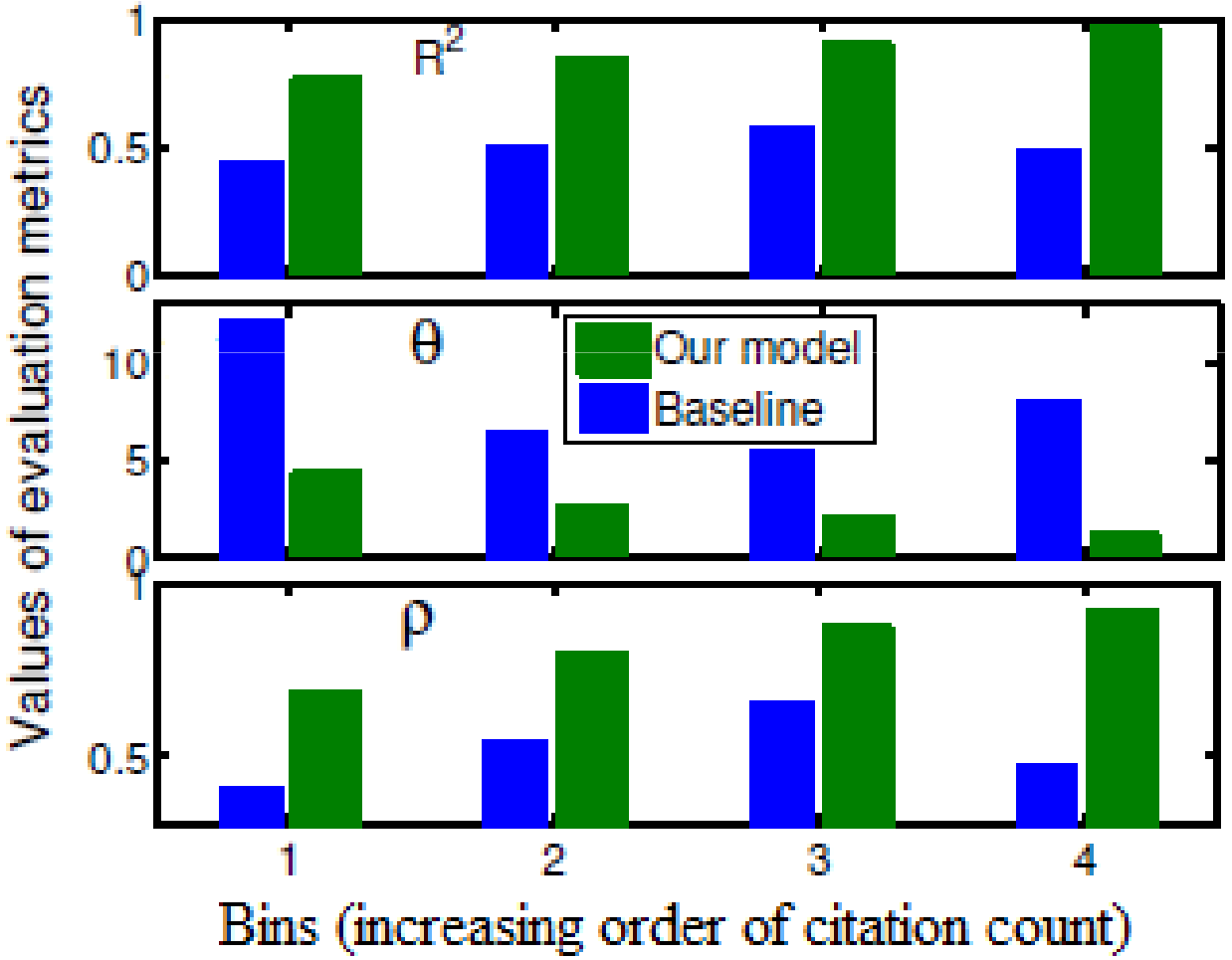
Confusion Matrix

	PeakInit	PeakMul	PeakLate	MonDec	MonIncr	Oth	Accuracy
PeakInit	**9550**	70	20	20	0	2419	0.79
PeakMul	29	**15261**	2500	3	0	3000	0.73
PeakLate	7	718	**4842**	2	489	518	0.73
MonDec	398	444	157	**2247**	0	453	0.61
MonIncr	2	403	0	0	**2789**	0	0.87
Oth	55	5142	5	2	0	**154188**	0.96
Overall accuracy							0.78

Performance Evaluation

	Baseline		
	R^2	θ	ρ
Δt=1	0.57	5.06	0.61
Δt=2	0.55	7.10	0.59
Δt=3	0.52	8.78	0.65
Δt=4	0.50	10.06	0.75
Δt=5	0.45	13.06	0.42

Performance in Different Citation Regions



Feature Analysis

Features		Decrease in overall accuracy of SVM
Author-centric	AvgProAuth	**0.21**
	MaxProAuth	0.05
	AvgHindex	0.06
	MaxHindex	0.04
	AvgAuthVer	0.12
	MaxAuthVer	0.18
	AvgNOCA	0.10
	MaxNOCA	0.16
Venue-centric	VenPres	0.12
	VenIF	0.13
	VenVer	**0.18**
Paper-centric	Team	**0.11**
	RefCount	0.01
	RDI	0.07
	KDI	0.03
	Topic	0.10

More About the Model

- Robustness of the categorization
  - **Merging of similar categories** (such as PeakInit and MonDec) deteriorates the performance

- Impact of early citation information
  - Inclusion of **first year's citations** of a paper enhances the performance

Take Away

- **Publication universe** is **heterogeneous** in terms of citation profile
- **Stratified Learning**, a generic approach in machine learning helps enhancing a citation count prediction model
- **Author-centric features** are the most distinguishing ones
- **Adding first year's citation count** as a feature can improve the prediction accuracy

Future Plan

- Deeper analysis of the categorization
- Inclusion of **content information** as a feature in the model
- **A new growth-model** to mimic this categorization

Thank you

http://cnerg.org
http://cse.iitkgp.ernet.in/~tanmoyc