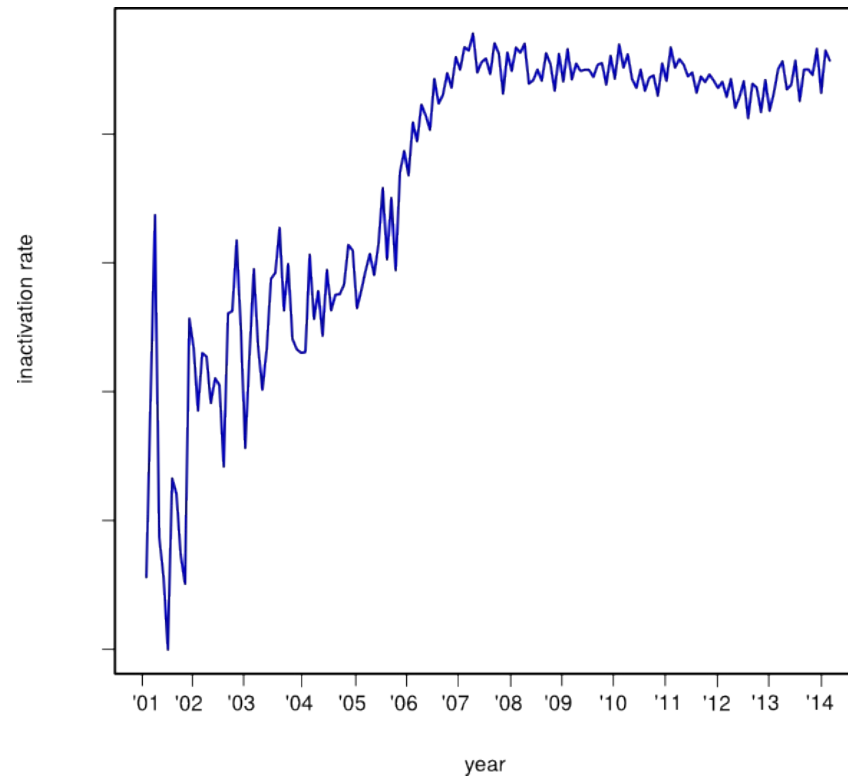
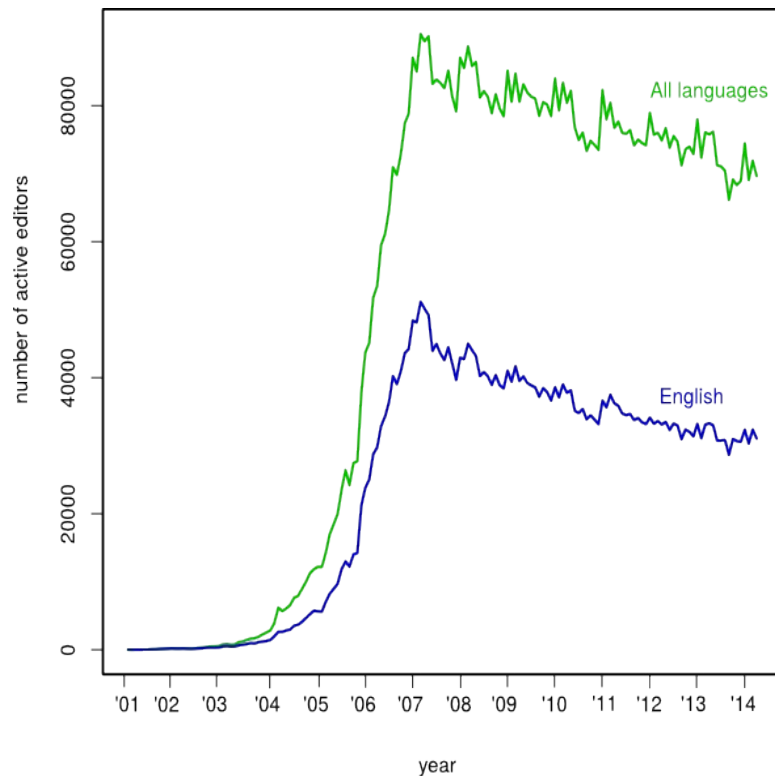
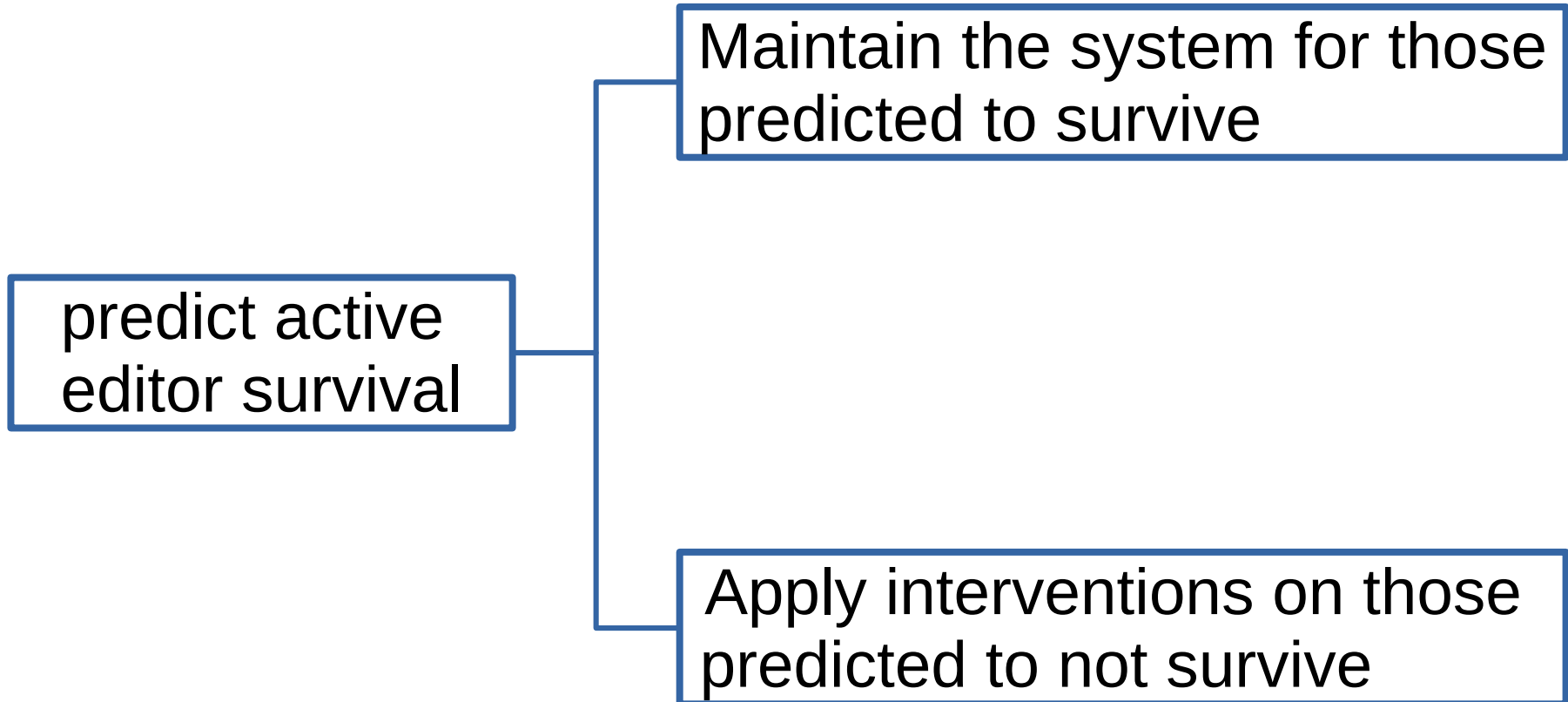


Active Editors

- An editor with 5+ edits in a main namespace
- 29% of those edited did 5+ edits (2013)
- Across wp projects, 87% of the monthly content was generated by active editors (March 2013)



Goal



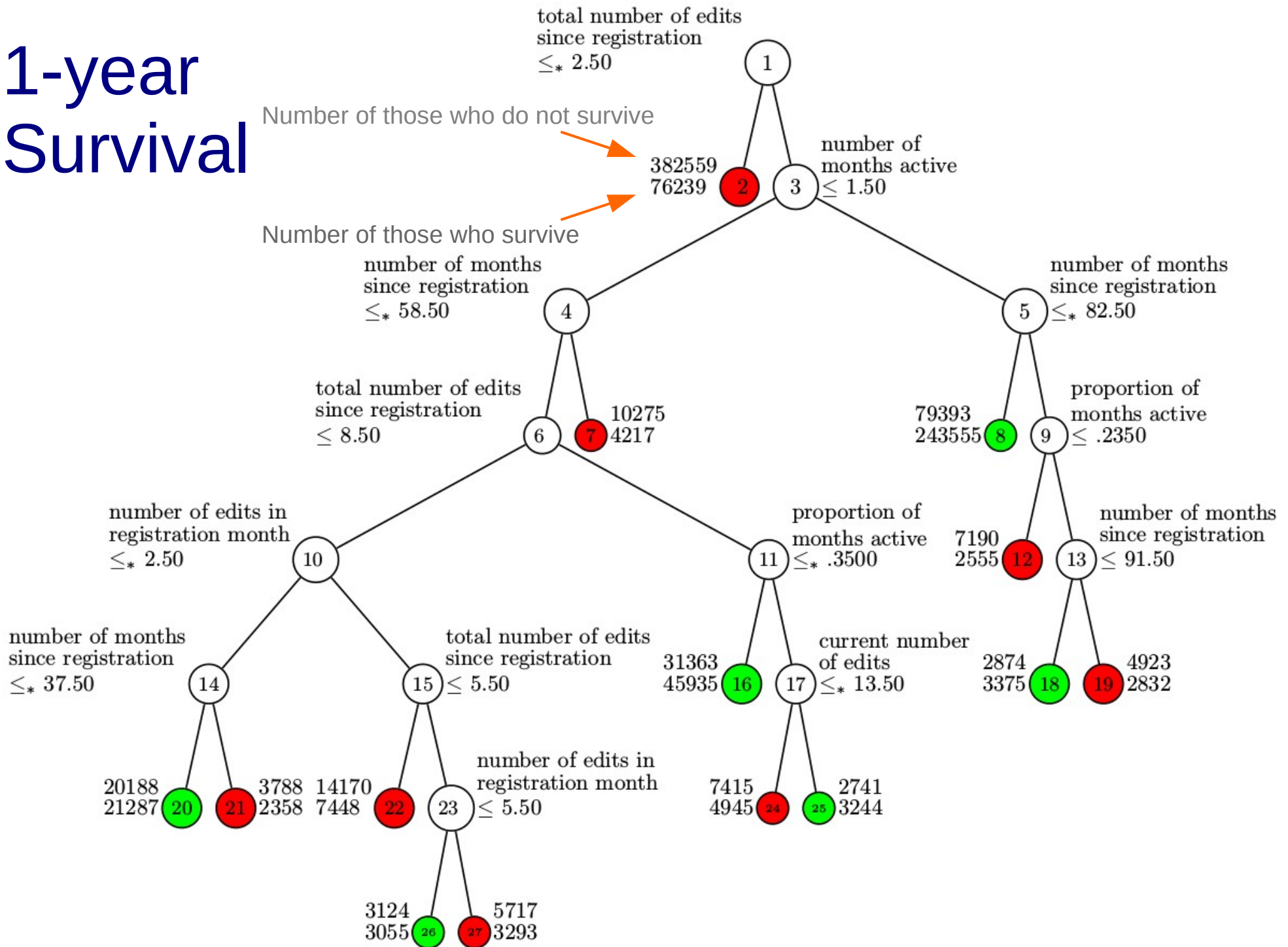
Data and Prediction Model

- (userId, edit date, number of edits, registration date)
- 12,500,726 editor records for enwp
- February 2001 to May 2014
- Sampled 10% of the data
 - 80% for training
 - 20% for testing
- Classification trees [GUIDE]

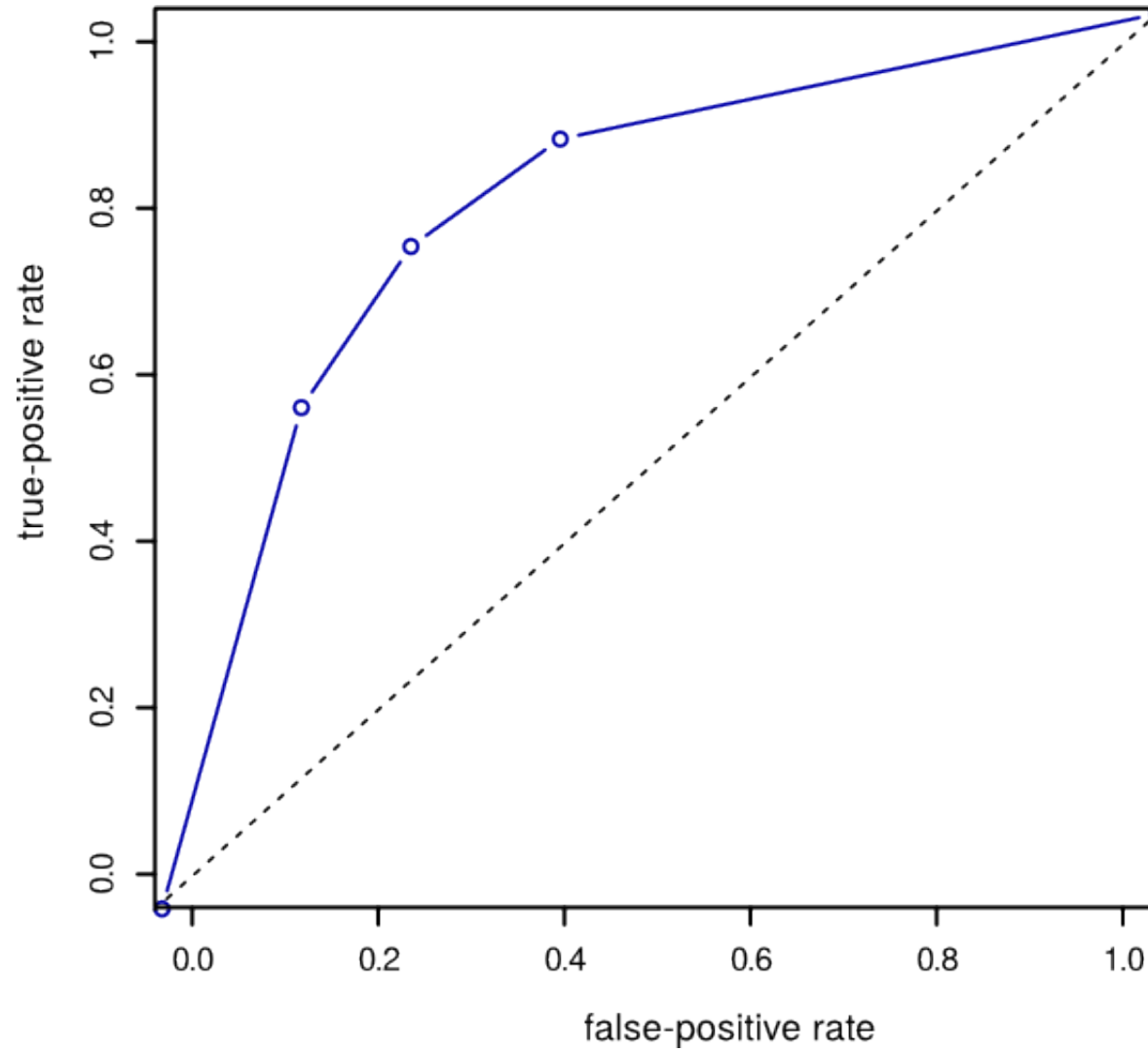
Variables for n -month Survival Model

- Outcome: 1, if an active editor in month m is active in month $m+n$; 0, otherwise.
- Independent Variables (11)
 - Number of edits (in the registration month, current month, since registration)
 - Commitment
 - maximum length of time in months during which the editor stayed active
 - number of months between current activity month and the last time the editor was active
 - proportion of months since registration in which the editor was active
 - Number of months during which the editor stayed active
 - Registration and vintage (registration day and month, number of months since registration)
 - First time editor

1-year Survival



ROC Curves for 1-year Survival



Variable Importance Ranking

Variables	30-day	1-year
total number of edits since registration	1	1
proportion of months active	2	2
number of months active	4	3
first time active editor	3	4
maximum continuity	5	5
number of months since registration	6	6
number of edits in registration month	7	7
number of edits in current month	8	8
number of months since last time active	9	9
registration day	10	10
registration month	11	11

Discussion

- Few independent variables, good results
- Next steps
 - Add more independent variables
 - Expand to other projects
 - Other prediction algorithms for performance comparison