Factors Leading to Success or Abandonment of Open Source Commons: An Empirical Analysis of Sourceforge.net Projects

Charles M. Schweik
Robert English
Sandra Haire

National Center for Digital Government;
Center for Public Policy and Administration;
Department of Natural Resources Conservation;
University of Massachusetts, Amherst, USA

Broad Research Question and Motivation

- What factors lead to ***collaborative*** success or abandonment in open source commons?

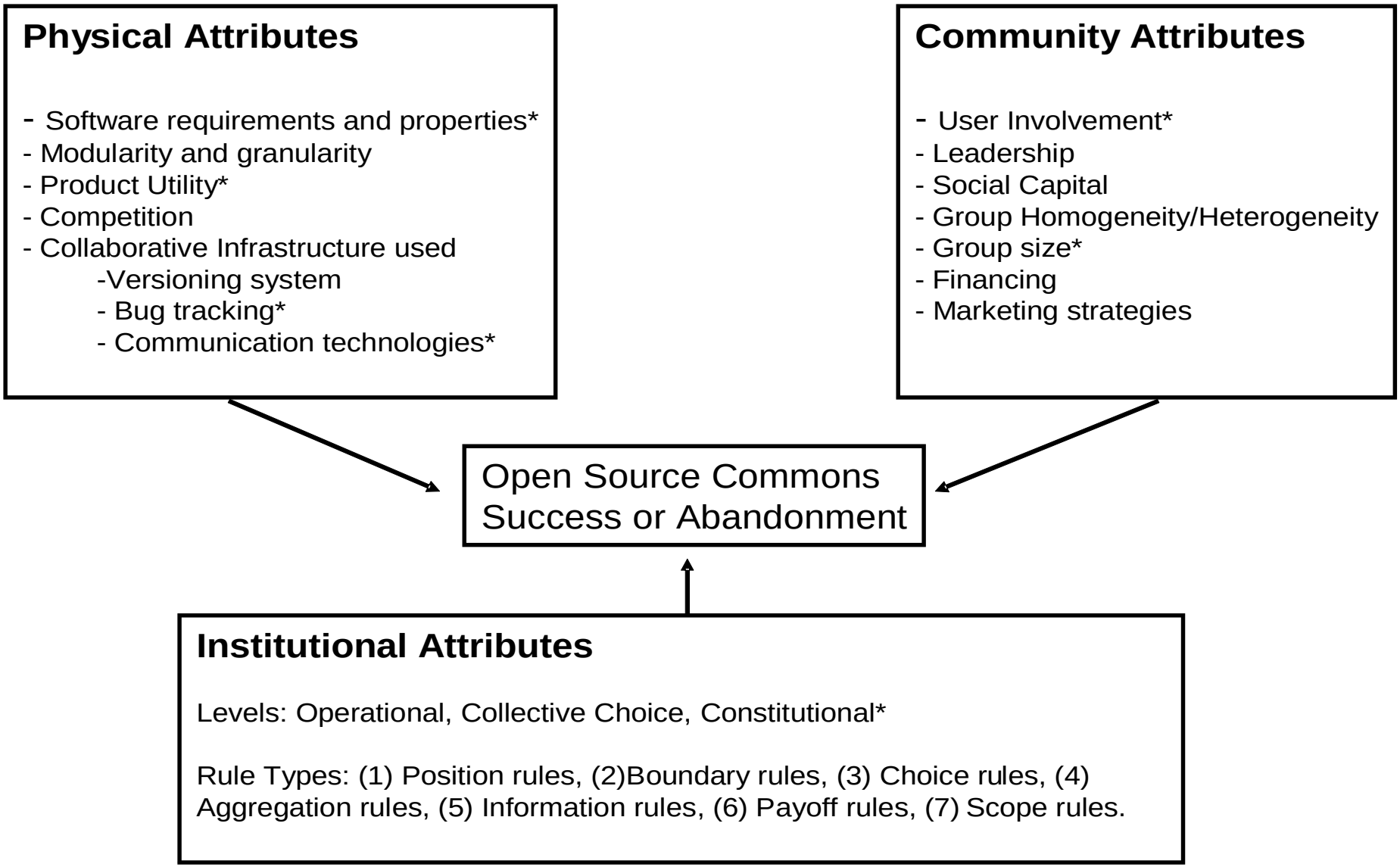
- Motivation: To inform future open source development efforts -- but also to inform "open content" collaborations outside of software.

The Paper - Two Parts

1. Literature review – identify factors thought to influence open source collaborations
    - Traditional IT and Software Engineering
    - Distributed Work and Virtual Teams
    - Collective Action and “Commons” literature

2. Exploratory Data Analysis of Sourceforge.net projects

Part 1. Literature Review



Note: "*" denotes concepts that we could operationalize using the Sourceforge.net dataset

Physical Attributes

- Software requirements and **properties***
- Modularity and granularity
- **Product Utility***
- Competition
- **Collaborative Infrastructure used**
    - -Versioning system
    - **Bug tracking***
    - **Communication technologies***

Community Attributes

- **User Involvement***
- Leadership
- Social Capital
- Group Homogeneity/Heterogeneity
- **Group size***
- Financing
- Marketing strategies

Institutional Attributes

Levels: Operational, Collective Choice, **Constitutional***

Rule Types: (1) Position rules, (2)Boundary rules, (3) Choice rules, (4) Aggregation rules, (5) Information rules, (6) Payoff rules, (7) Scope rules.

Are we missing any key variables?

Part 2. Exploratory Data Analysis of Sourceforge.net

- Sourceforge.net

  -- a kind of “remote sensor” of open source

  -- 107,747 projects in our dataset

  -- Two stages: Initiation and Growth

Defining Success and Abandonment

Success, Initiation Stage	First release of code
Abandonment, Initiation Stage	No release and no apparent activity
Success, Growth Stage	At least 3 meaningful releases and a few users
Abandonment, Growth Stage	Appears abandoned before producing 3 releases

Defining Success and Abandonment

Success, Initiation Stage	First release of code
Abandonment, Initiation Stage	No release and no apparent activity
Success, Growth Stage	At least 3 meaningful releases and a few users
Abandonment, Growth Stage	Appears abandoned before producing 3 releases

- Validation process
- Details in English and Schweik (2007), *Upgrade* paper

http://www.upgrade-cepis.com/issues/2007/6/upg8-6English_Schweik_v2.pdf

SF Numeric Independent Variables

VARIABLE	DESCRIPTION	ASSOCIATED THEORETICAL CONCEPT (P-Physical, C-Community, I-Institutional)
Developers	Total # of developers on project	Group size
Tracker reports	Total # of bug reports, feature requests, patches, and support requests	Collaborative infrastructure
Page visits	Total # of views of project web pages	Product Utility; Group Size
Forum posts	Total # of forum posts from 10/6/2005 – 8/2/2006	Collaborative infrastructure; Group size
Downloads	Total # of downloads	Product Utility; Group Size

SF Categorical Independent Variables

VARIABLE	DESCRIPTION	ASSOCIATED THEORETICAL CONCEPT (P-Physical, C-Community, I-Institutional)
Intended Audience	Type of person targeted:(1) business; (2) end users; (3) computer professionals; (4) government/ nonprofit; and (5) other	User Involvement (C)
Operating System	Operating system(s) the software will run on	Product Utility; Critical Infrastructure (P)
Programming Language	Types of languages used	Product Utility (preferred technologies) (P)

Categorical variables – possible values: 0 – not chosen; 1 – chosen; 2-no subcategory chosen either

SF Categorical Independent Variables

VARIABLE	DESCRIPTION	ASSOCIATED THEORETICAL CONCEPT (P-Physical, C-Community, I-Institutional)
User Interface	How software interfaces with user (e.g., command line, GUI, etc.)	Product Utility (preferred technologies) (P)
Database Environment	Database technology used (if applicable)	Product Utility (preferred technologies) (P)
Project Topic	SF classification of project topics (19 subcategories such as education, games, security, printing, etc.)	Product Utility (critical infrastructure) (P)
Project License	Type of open source license used	Constitutional rules (I)

Classification Trees

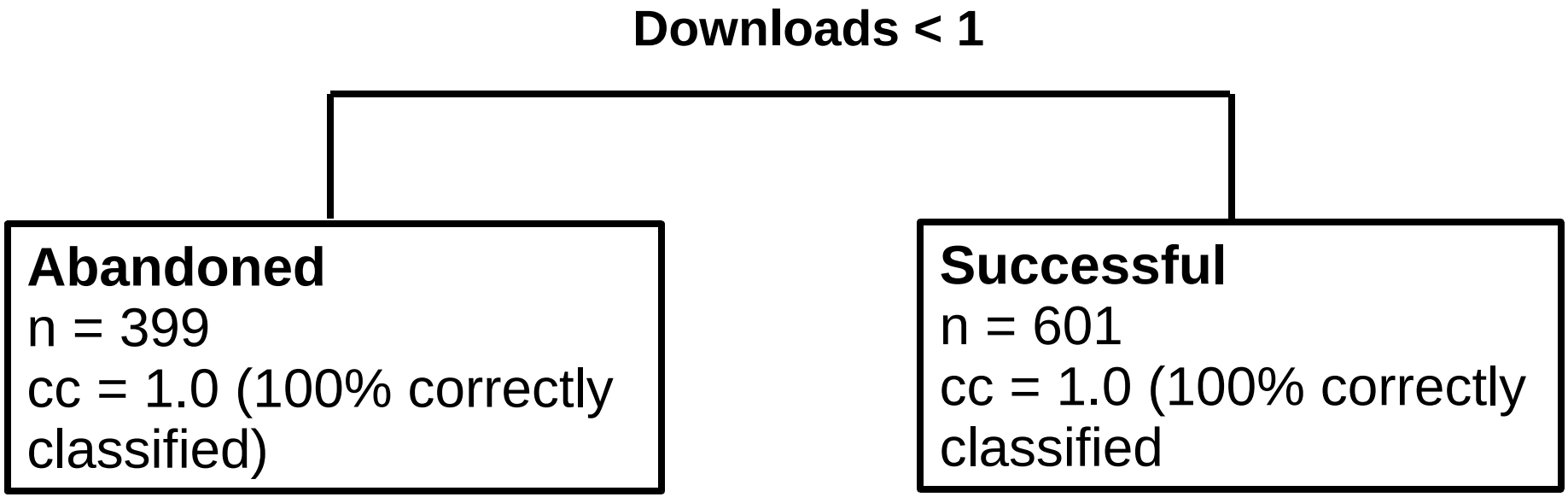
- Exploratory, nonparametric approach
- Which variables help to discriminate success and abandonment in the Growth Stage?
- Problem – TOO much data!
- Random samples of 1000 projects

A Simple Classification Tree Example

Success in the ***Initiation Stage*** is defined as :

- A first release
- There IS some software to download...

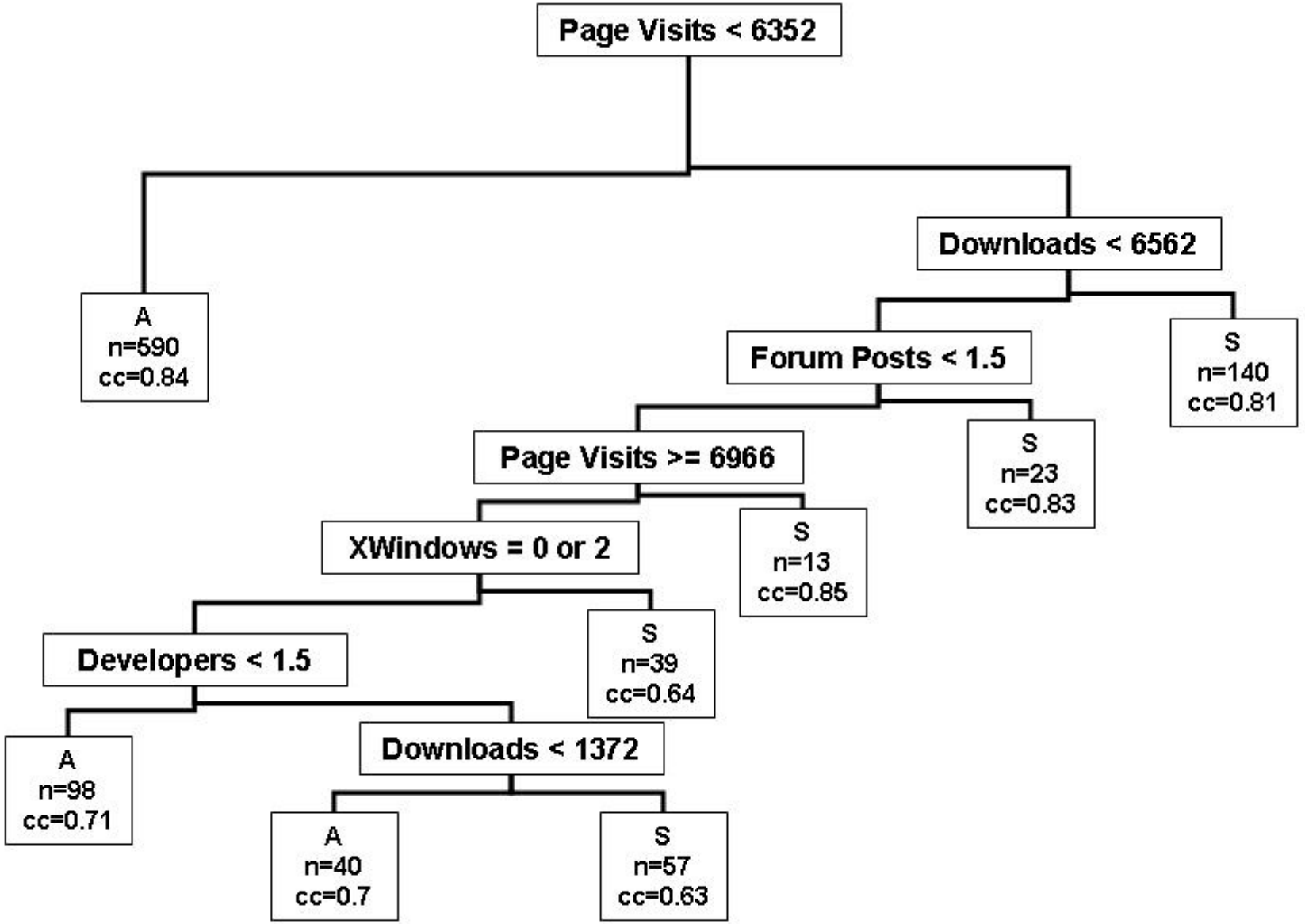
A Simple Classification Tree – Initiation Stage Example



High classification accuracy is because:

- Projects with downloads are successful by definition (have releases);
- Projects without releases (and therefor 0 downloads) after 1 year are abandoned by definition
- Could have a project with a release and 0 downloads -- but VERY rare

Example of Growth Stage Results (n=1000)



“RANDOM FORESTS” OUTPUT

RANDOM FORESTS:

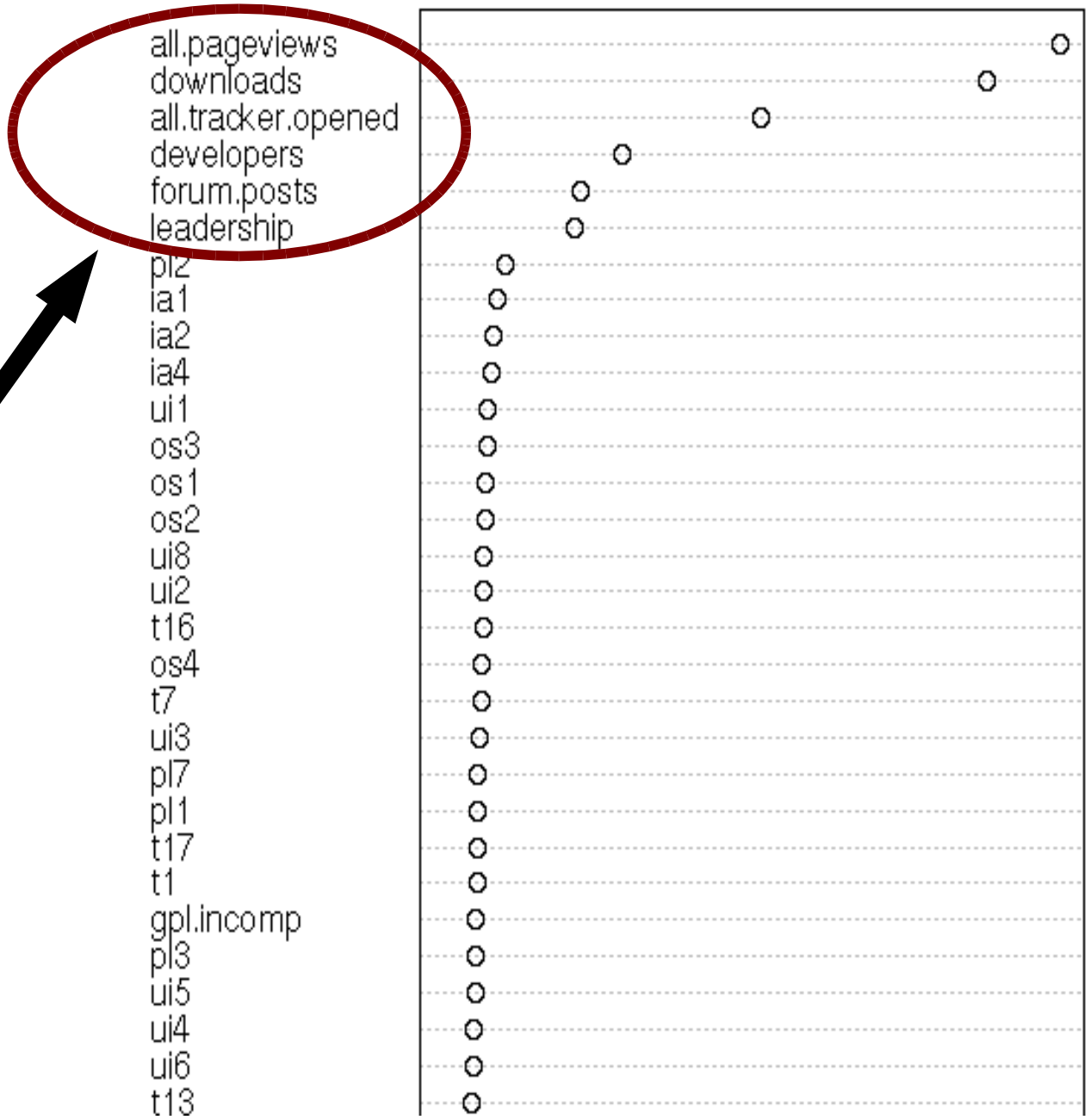
Fits many classification trees to the dataset

Combines predictions from all of these trees

RESULTS:

Numeric data more important than categorical

- Page views
- Downloads
- Tracker
- # of developers
- forum posts



Interpretation of results

- Successful Growth Stage projects:
  - Have large user community
    - Intuitive, but we could have found that majority of SG projects were small projects with few users
  - Use bug tracking and forums
    - A lack of tracker or forum activity marks a disinterested community
  - Widely distributed among categorical variables
    - People collaborating in all sorts of open source, not just “user-centric” (von Hippel) or key open source technologies (e.g., to support Linux)
    - Sign that open source is maturing

Next Steps

- Book Manuscript: “Collaborative Principles of Open Source Commons”
- OSGeo Foundation and affiliated projects as a case study
  - Help to identify attributes not in SF dataset
- Next steps
  - Online survey of open source projects
  - Classification tree with data on institutional and community variables

Thank you for attending our session!

Acknowledgments

Support for this work was provided by a grant from the U.S. National Science Foundation (NSFIIS 0447623).

Thanks to the FLOSSMole project at Syracuse University for some of the SF data used in this study

Trajectories of Open Source Projects