Automatically assembling a census of an academic field

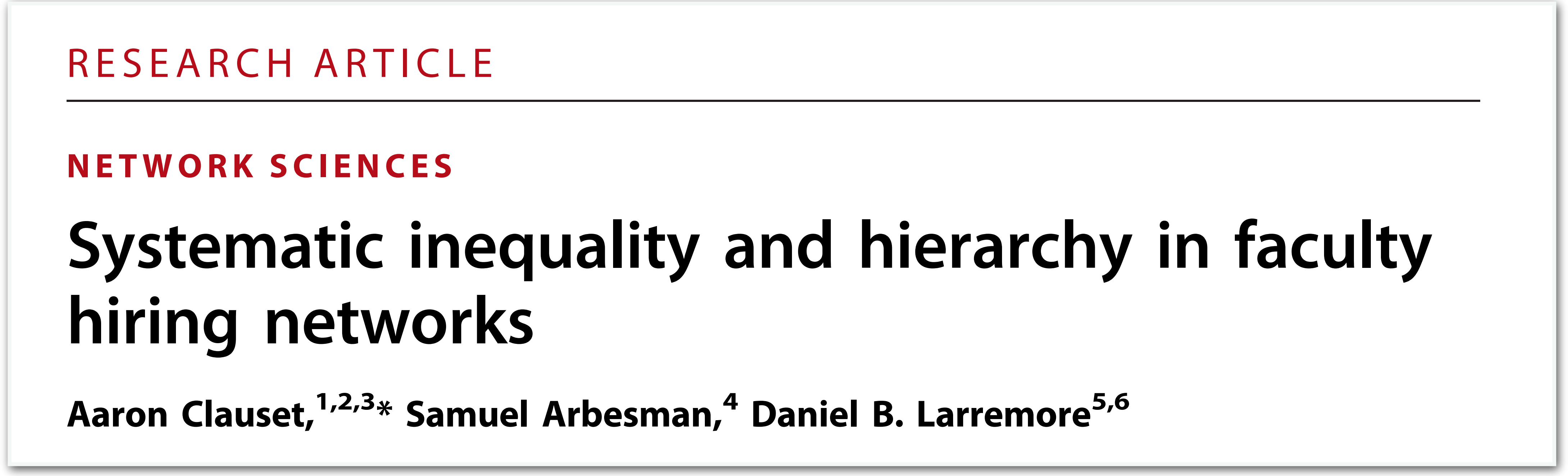
Allison Morgan, Samuel Way, Aaron Clauset
University of Colorado Boulder

About Me

Third year PhD Student in CS at CU Boulder

Collaborators and I study the "sociology of science"

Interested in computational methods to study under-representation in academia



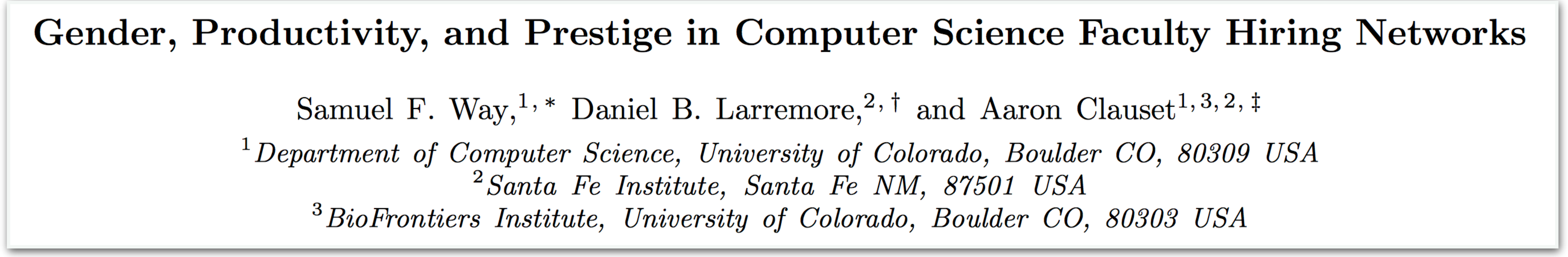
RESEARCH ARTICLE

NETWORK SCIENCES

Systematic inequality and hierarchy in faculty hiring networks

Aaron Clauset,[1,2,3]* Samuel Arbesman,[4] Daniel B. Larremore[5,6]

Science Advances 1(1), e1400005, (2015).

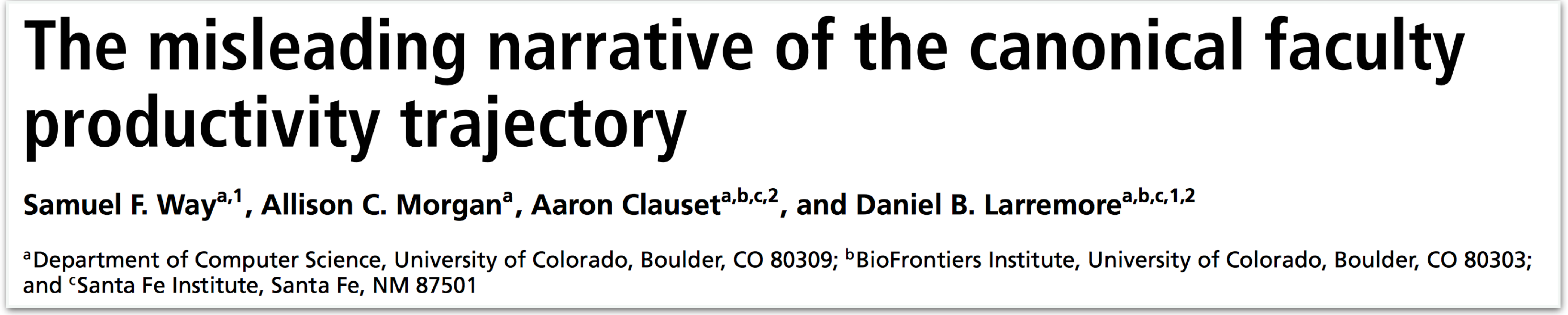


Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks

Samuel F. Way,[1,*] Daniel B. Larremore,[2,†] and Aaron Clauset[1,3,2,‡]

[1] *Department of Computer Science, University of Colorado, Boulder CO, 80309 USA*
[2] *Santa Fe Institute, Santa Fe NM, 87501 USA*
[3] *BioFrontiers Institute, University of Colorado, Boulder CO, 80303 USA*

Proc. 25th Int'l World Wide Web Conf. (WWW), (2016)



The misleading narrative of the canonical faculty productivity trajectory

Samuel F. Way[a,1], Allison C. Morgan[a], Aaron Clauset[a,b,c,2], and Daniel B. Larremore[a,b,c,1,2]

[a]Department of Computer Science, University of Colorado, Boulder, CO 80309; [b]BioFrontiers Institute, University of Colorado, Boulder, CO 80303; and [c]Santa Fe Institute, Santa Fe, NM 87501

Proceedings of the National Academy of Sciences Oct 2017, 201702121

Motivation

Nobel Prize winners

Chemists

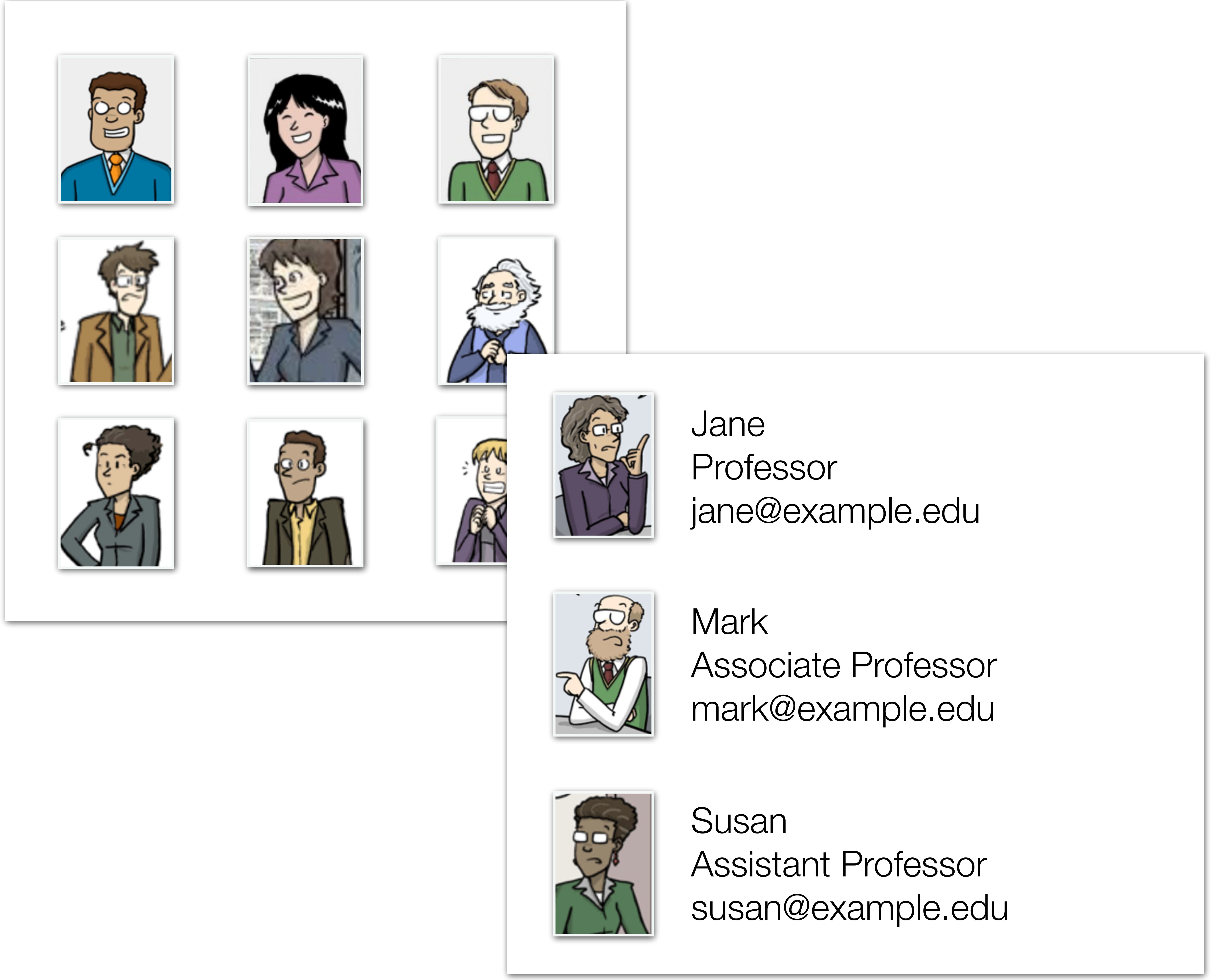
and those who leave academia

Cartoons by Jorge Chan; phdcomics.com

Much of the sociology of science studies small samples of the academic workforce at a single point in time.

Can we build a tool to efficiently collect the employment information of **all faculty** across institutions, **across time**?

Challenge



Jane
Professor
jane@example.edu

Mark
Associate Professor
mark@example.edu

Susan
Assistant Professor
susan@example.edu

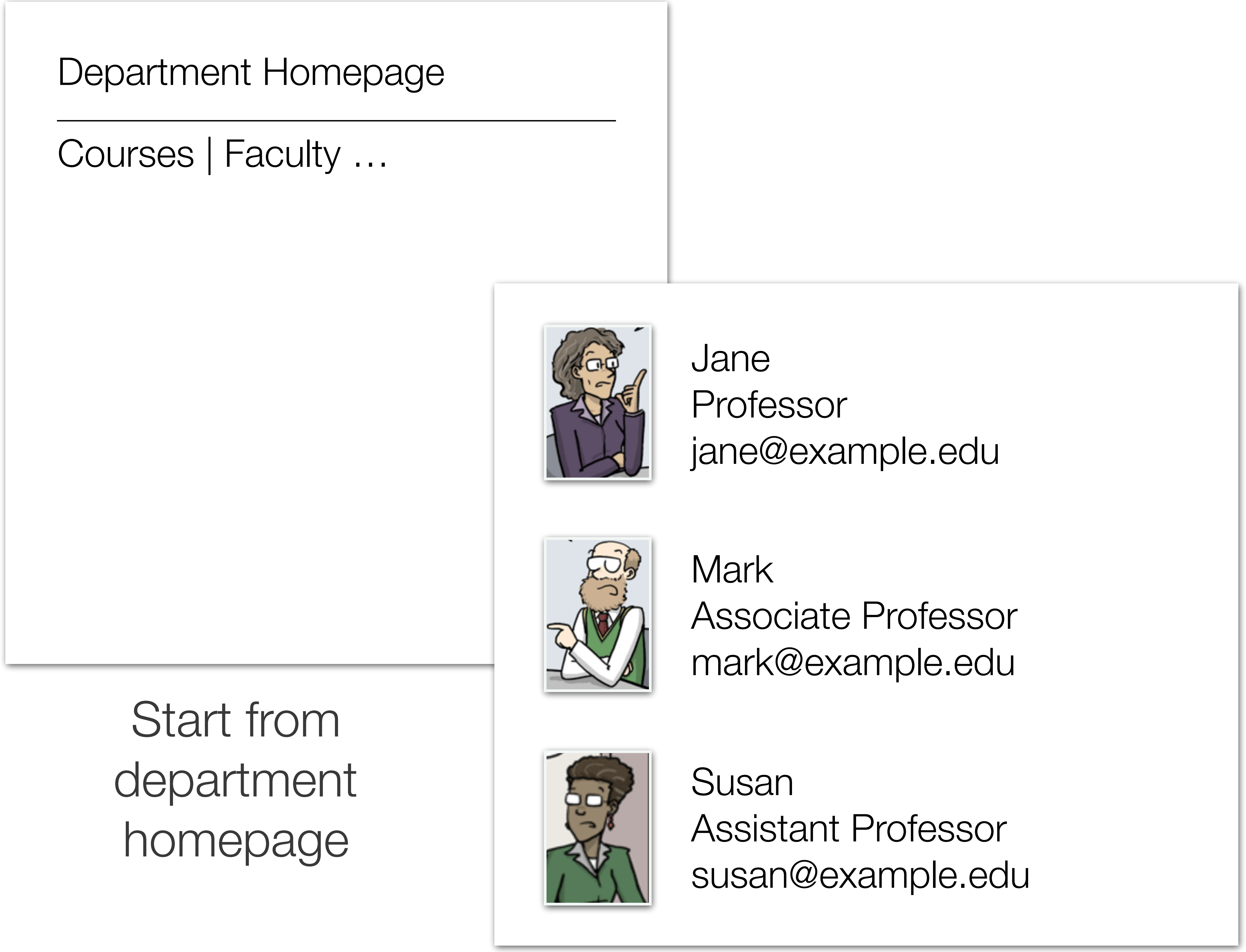
Every department contains a public directory of its faculty

With the same information: names, titles, email addresses, and webpages

But, information is distributed and not well structured

Cartoons by Jorge Chan; phdcomics.com

Our Approach



Department Homepage

Courses | Faculty …

Jane
Professor
jane@example.edu

Mark
Associate Professor
mark@example.edu

Susan
Assistant Professor
susan@example.edu

Start from
department
homepage

Navigate to its
faculty directory

Identify the directory's
HTML structure & extract
faculty information

```
faculty_name: Jane
title: Professor
website: ...
email: ...
```

Filter non-tenure-track
faculty for further analyses

title: Assistant Professor
~~title: Research Professor~~
title: Full Professor
~~title: Instructor~~

Cartoons by Jorge Chan; phdcomics.com

Our Approach

Start at department homepage

Collect all outward links

Sort links

Pick a link: is this a directory?

If not, try the next likely link

Given the directory URL

(i) Navigate to the directory

Parse tables

Parse lists

Parse divs

Parse articles

HTML structure has been identified

(ii) Identify the HTML structure of the directory

Identify names

Identify titles

Identify webpages

Identify emails

Directory with every person on the page

(iii) Identify faculty members

For each person

Is their title tenure-track?

If not, remove entry from directory

(iv) Sample the relevant faculty members

Filtered directory

Department Homepage
Courses | Prospective Students | Faculty
Start from department homepage
Jane
Professor
jane@example.edu
Mark
Associate Professor
mark@example.edu
Susan
Assistant Professor
susan@example.edu
Navigate to its faculty directory
Identify the directory's HTML structure & extract faculty information
faculty name: Jane
title: Professor
website: ...
email: ...
Filter non-tenure-track faculty for further analyses
title: Assistant Professor
~~title: Research Professor~~
title: Full Professor
~~title: Instructor~~

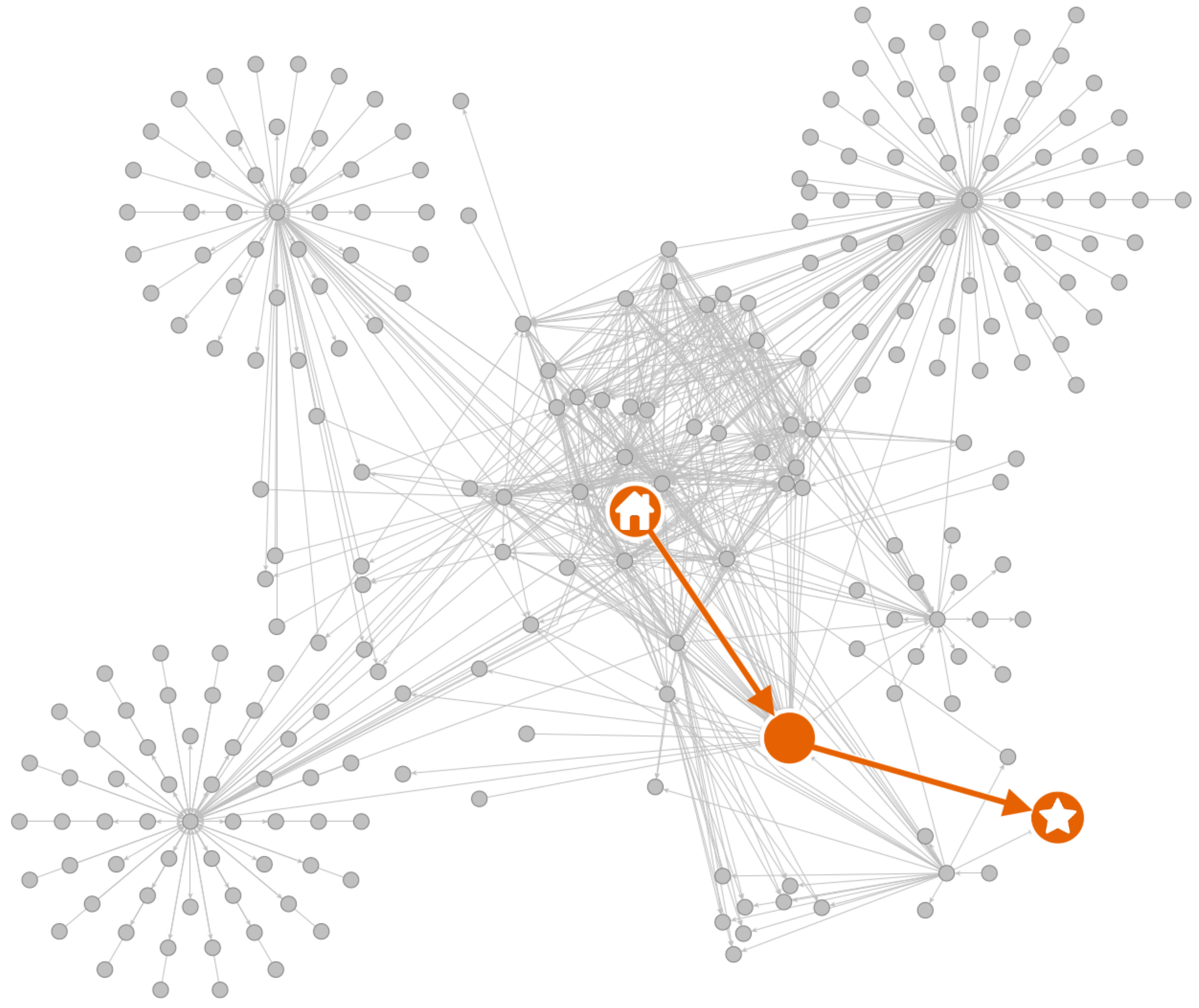
Cartoons by Jorge Chan; phdcomics.com

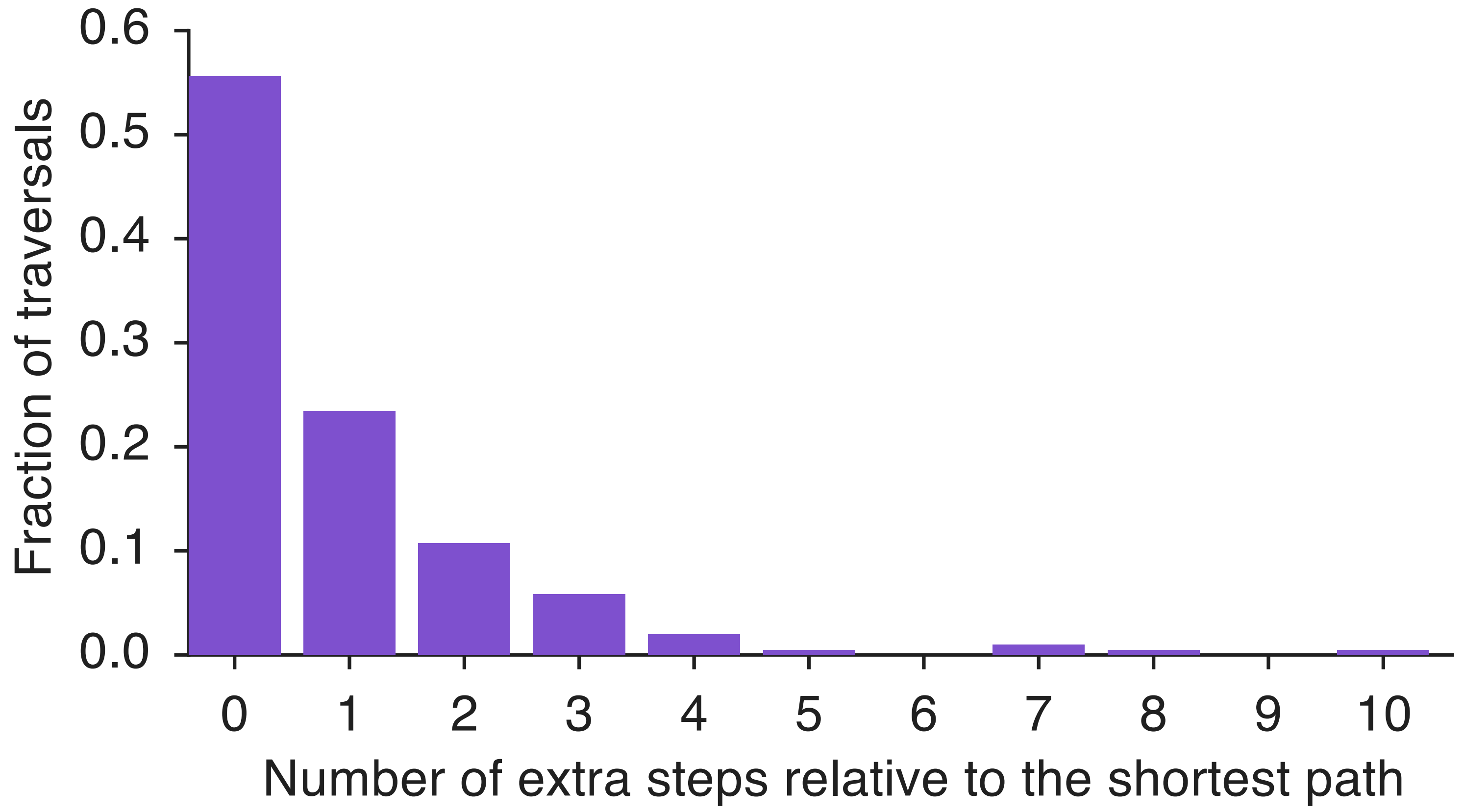
Navigation

From a department homepage, sort all outgoing links by keywords:

[“professor”, "faculty", “people", "directory", “personnel", “staff” …]

For more than half of departments, this heuristic results in the shortest path.





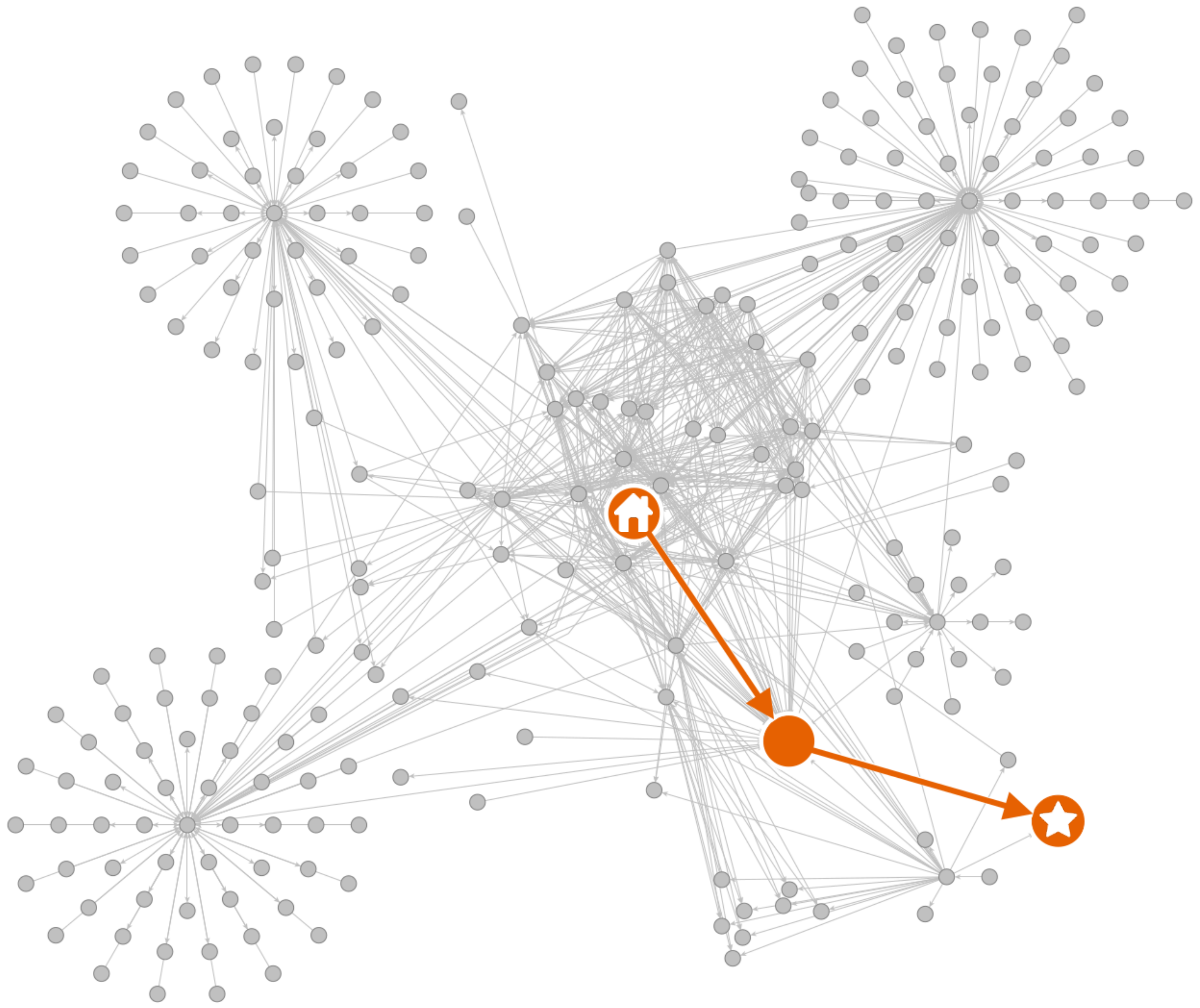
Showing http://www.cs.ucdavis.edu to http://www.cs.ucdavis.edu/people/faculty/

Navigation

To stop at directories, we use a random forest classifier trained on all directory pages, and a sample of non-directory pages.

Important features: [“NAME”, “TITLE”, “EMAIL”, “PHONE”, “website”, “profile”, “office”, “interest”]

Average accuracy is 82%*



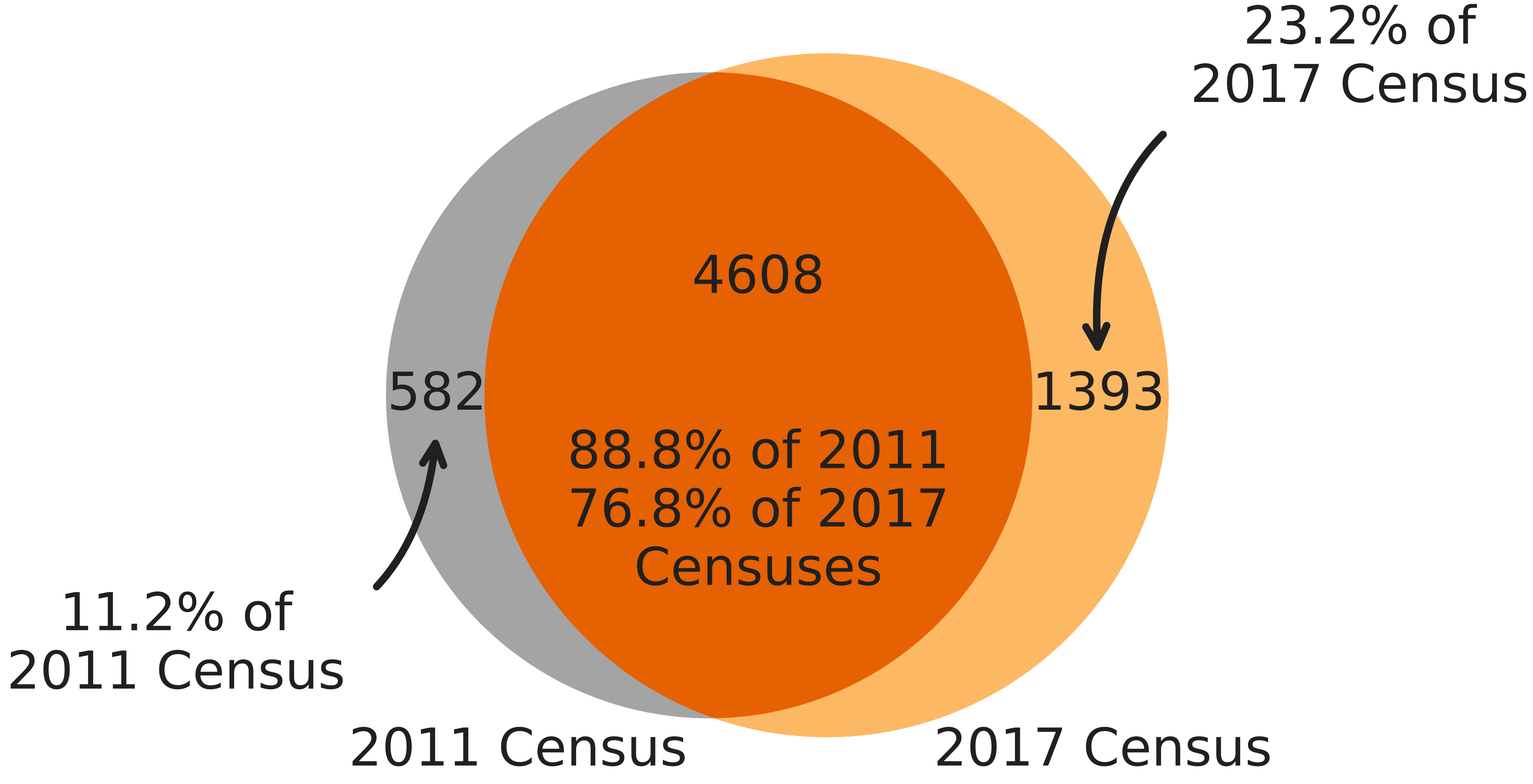
* To avoid skipping directory pages, we parse any page which has a likelihood of being a directory > 0 . Results in perfect recall, at the expense of precision.

Summary of Engineering Results

Fast: average < 1 minute vs ~8 hours to produce a single department's faculty directory

Accurate: 99% recall (nearly all tenure-track faculty are retrieved) and precision (few non-tenure-track faculty are retrieved)*

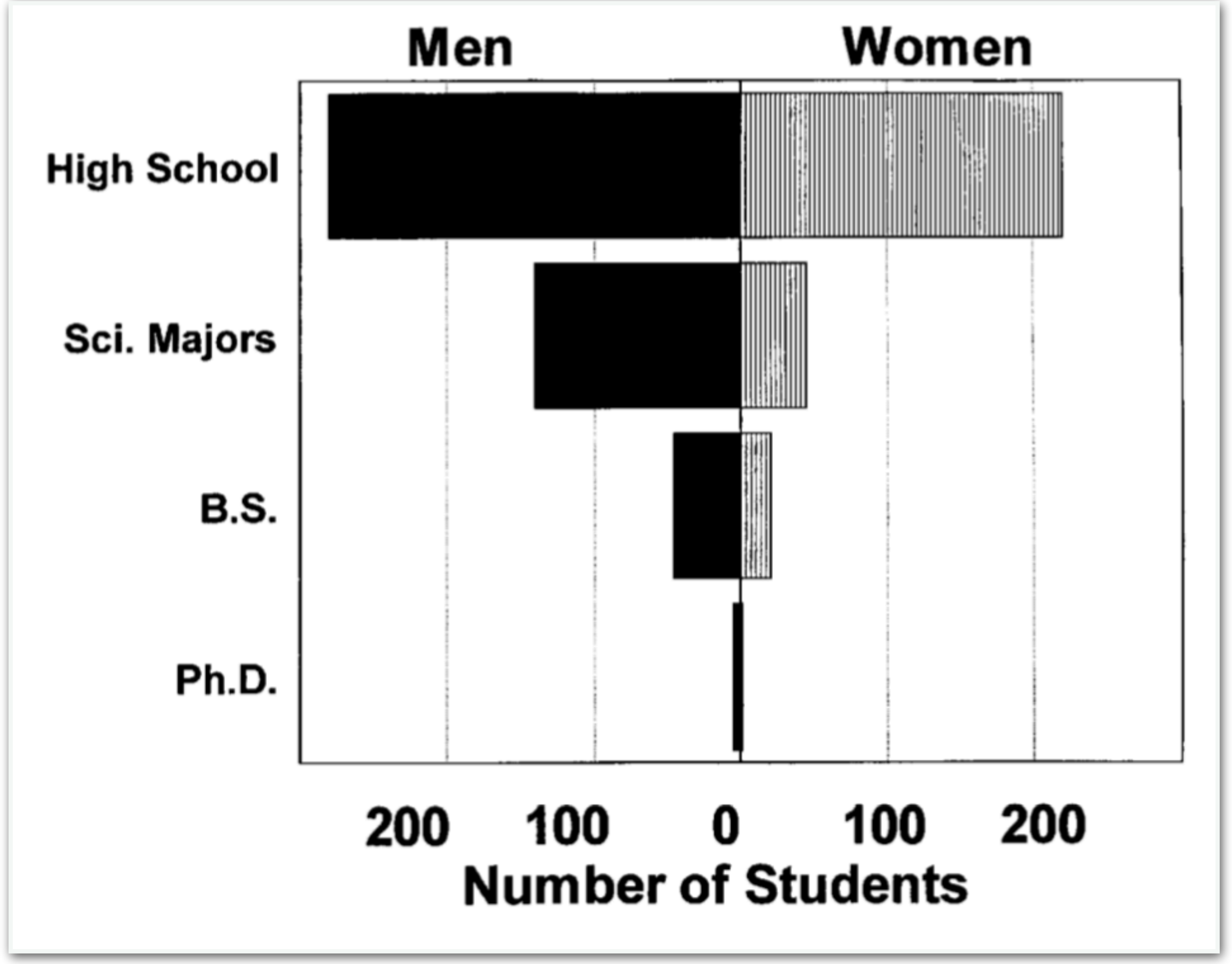
Comparable to findings of major survey organization: 16% vs 11% net growth in the number of faculty from the CRA



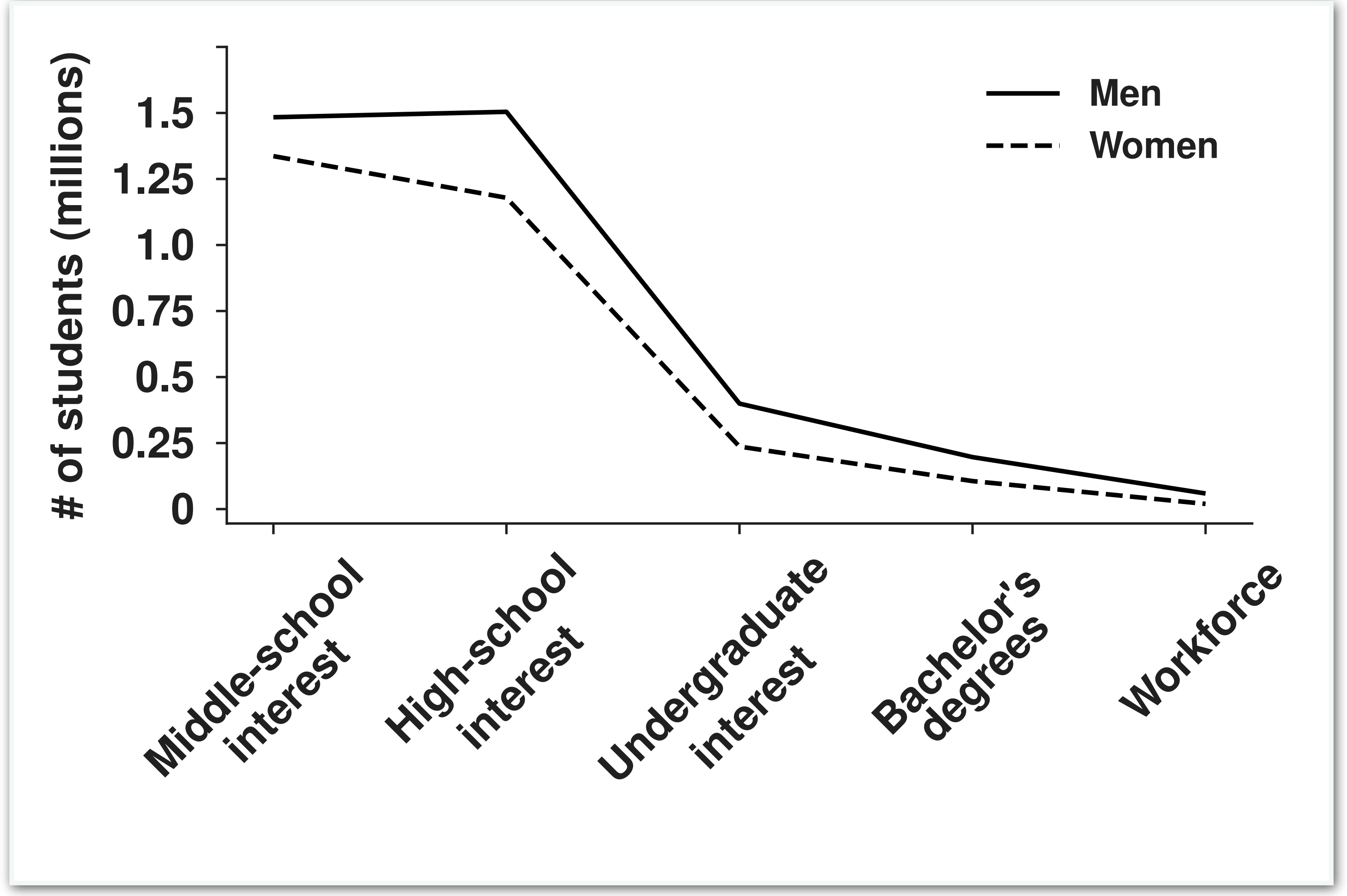
* *Manually checked against a third of departments; Computing Research Association: https://cra.org*

So what can we do with this tool?

We investigate the **“leaky pipeline”**: women leave STEM at various career stages, resulting in their under-representation at the faculty level

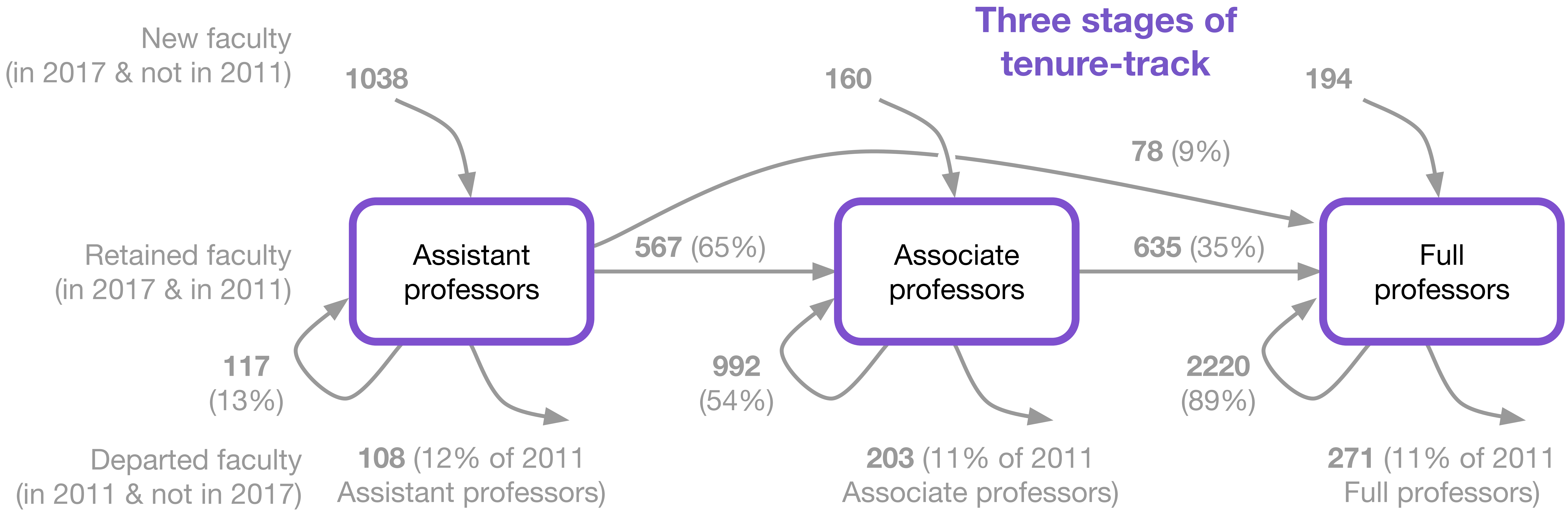


Journal of Animal Science, 74(11), 2843-2848, 1996



PloS ONE, *11*(7), e0157447, 2016

Leaky Pipeline



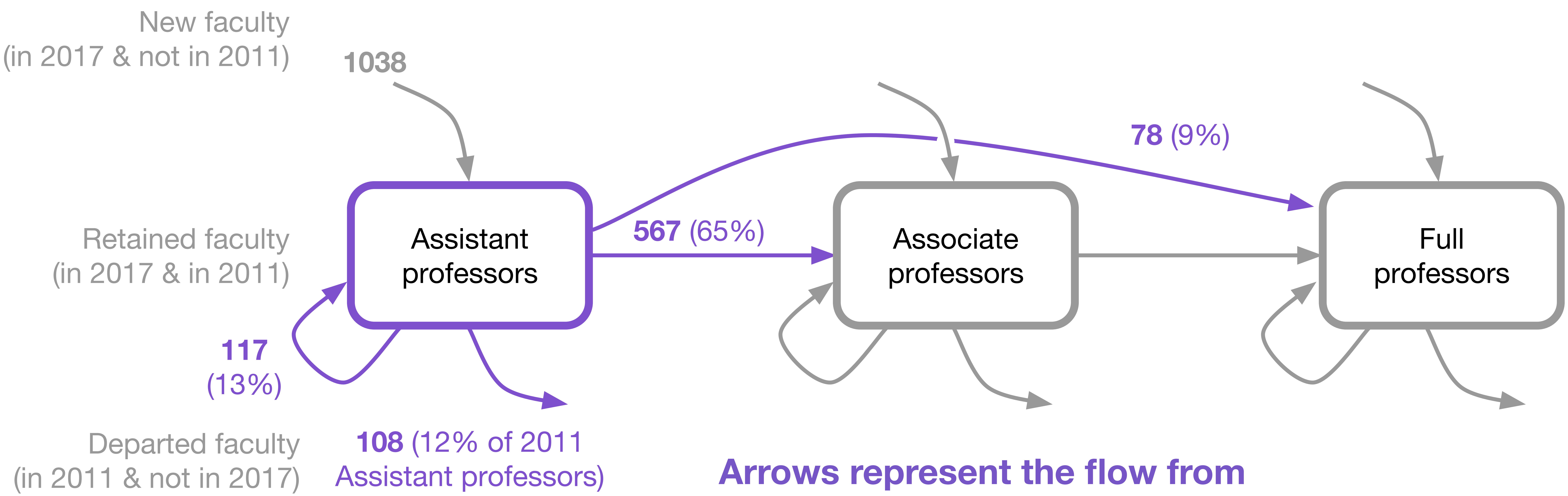
Three stages of tenure-track

New faculty (in 2017 & not in 2011)

Retained faculty (in 2017 & in 2011)

Departed faculty (in 2011 & not in 2017)

Leaky Pipeline



New faculty
(in 2017 & not in 2011)

1038

Retained faculty
(in 2017 & in 2011)

117
(13%)

Departed faculty
(in 2011 & not in 2017)

108 (12% of 2011
Assistant professors)

Assistant
professors

567 (65%)

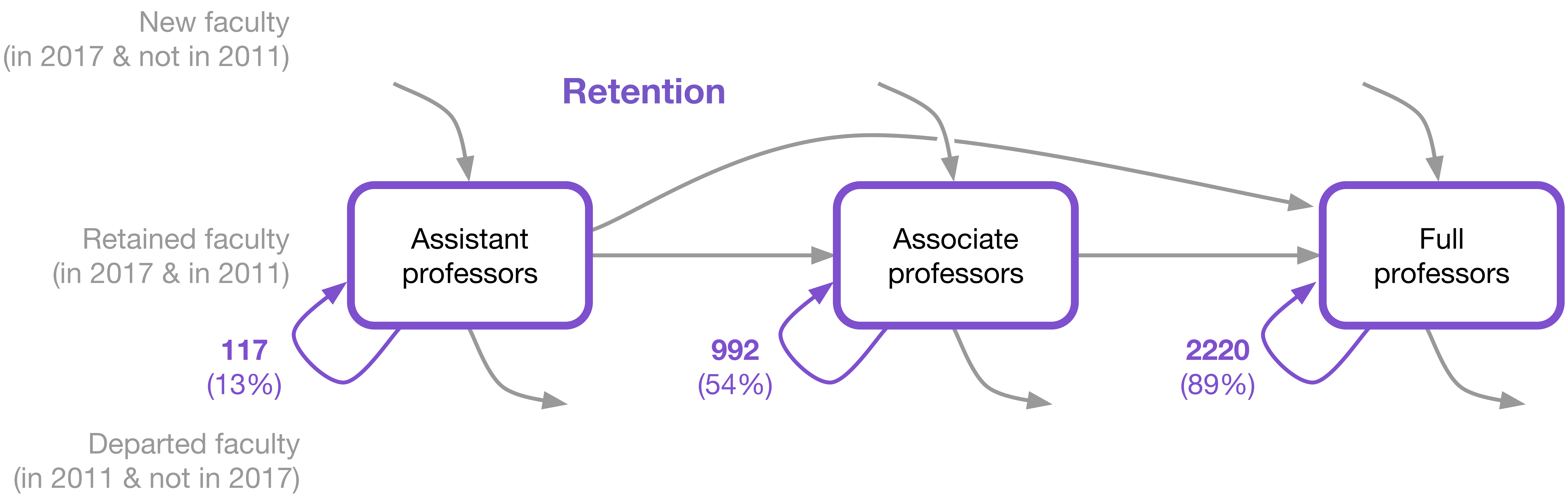
Associate
professors

78 (9%)

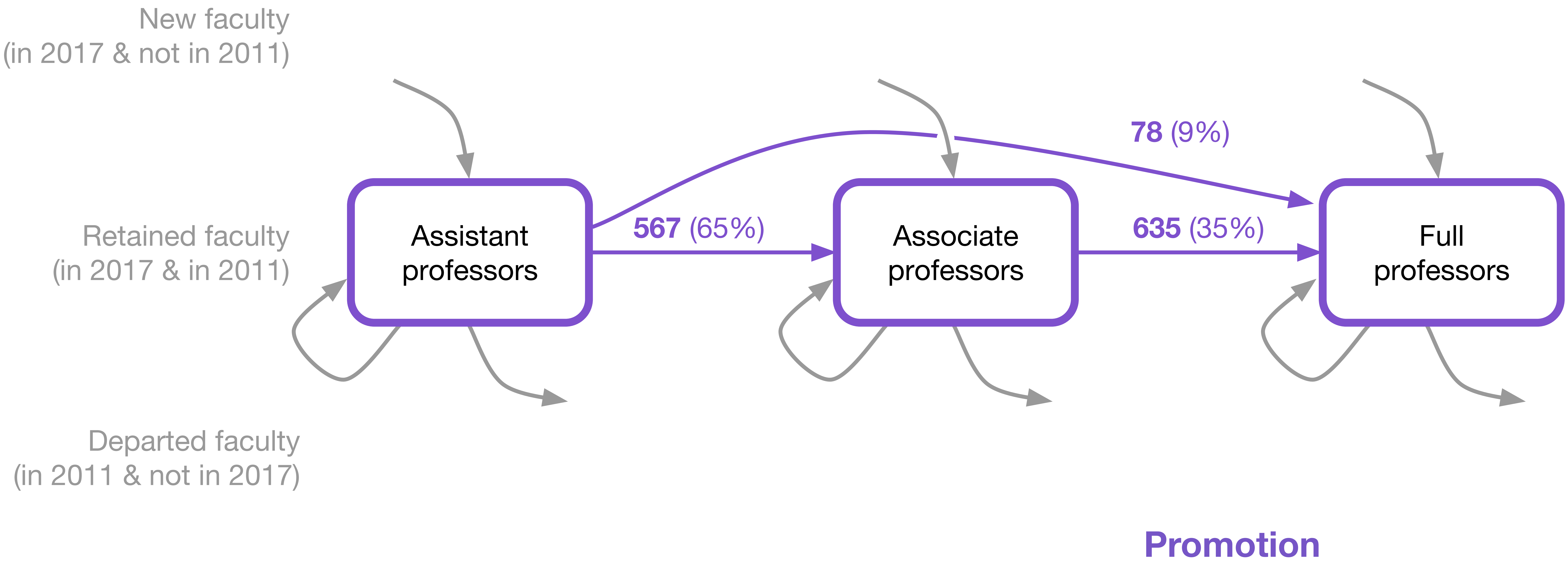
Full
professors

Arrows represent the flow from tenure-track stage in 2011 to 2017

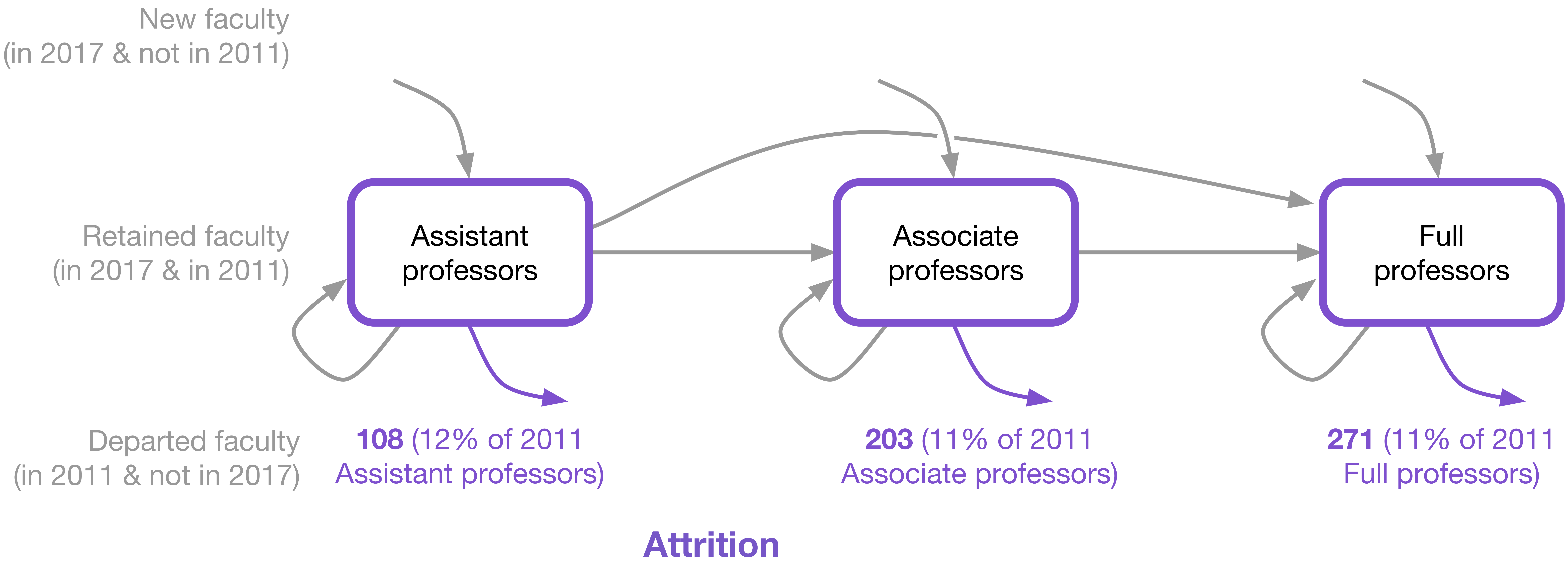
Leaky Pipeline



Leaky Pipeline



Leaky Pipeline



New faculty
(in 2017 & not in 2011)

Retained faculty
(in 2017 & in 2011)

Departed faculty
(in 2011 & not in 2017)

Assistant professors

Associate professors

Full professors

108 (12% of 2011 Assistant professors)

203 (11% of 2011 Associate professors)

271 (11% of 2011 Full professors)

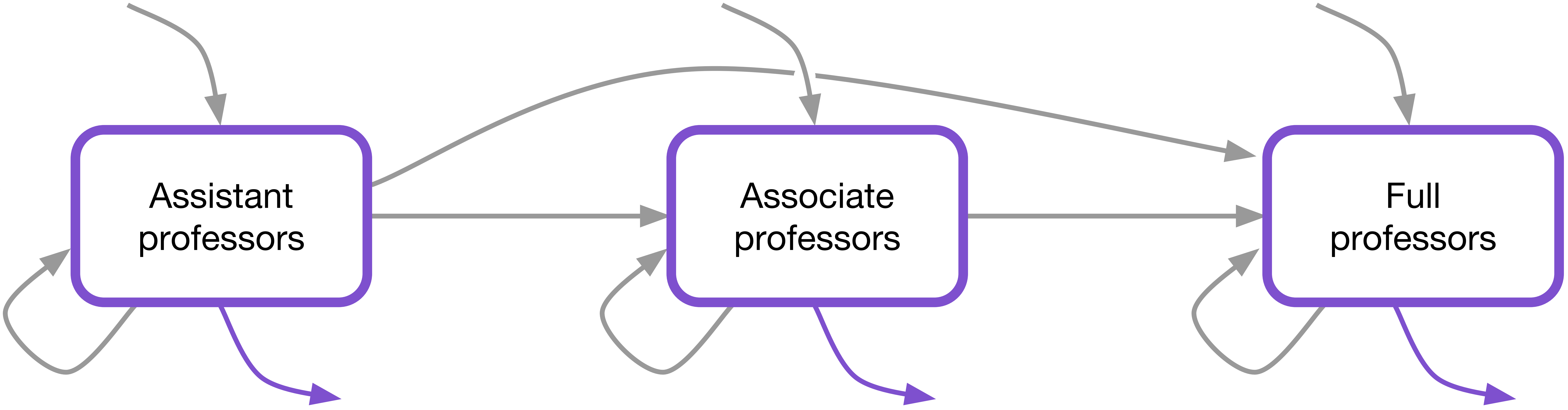
Attrition

Leaky Pipeline

New faculty
(in 2017 & not in 2011)

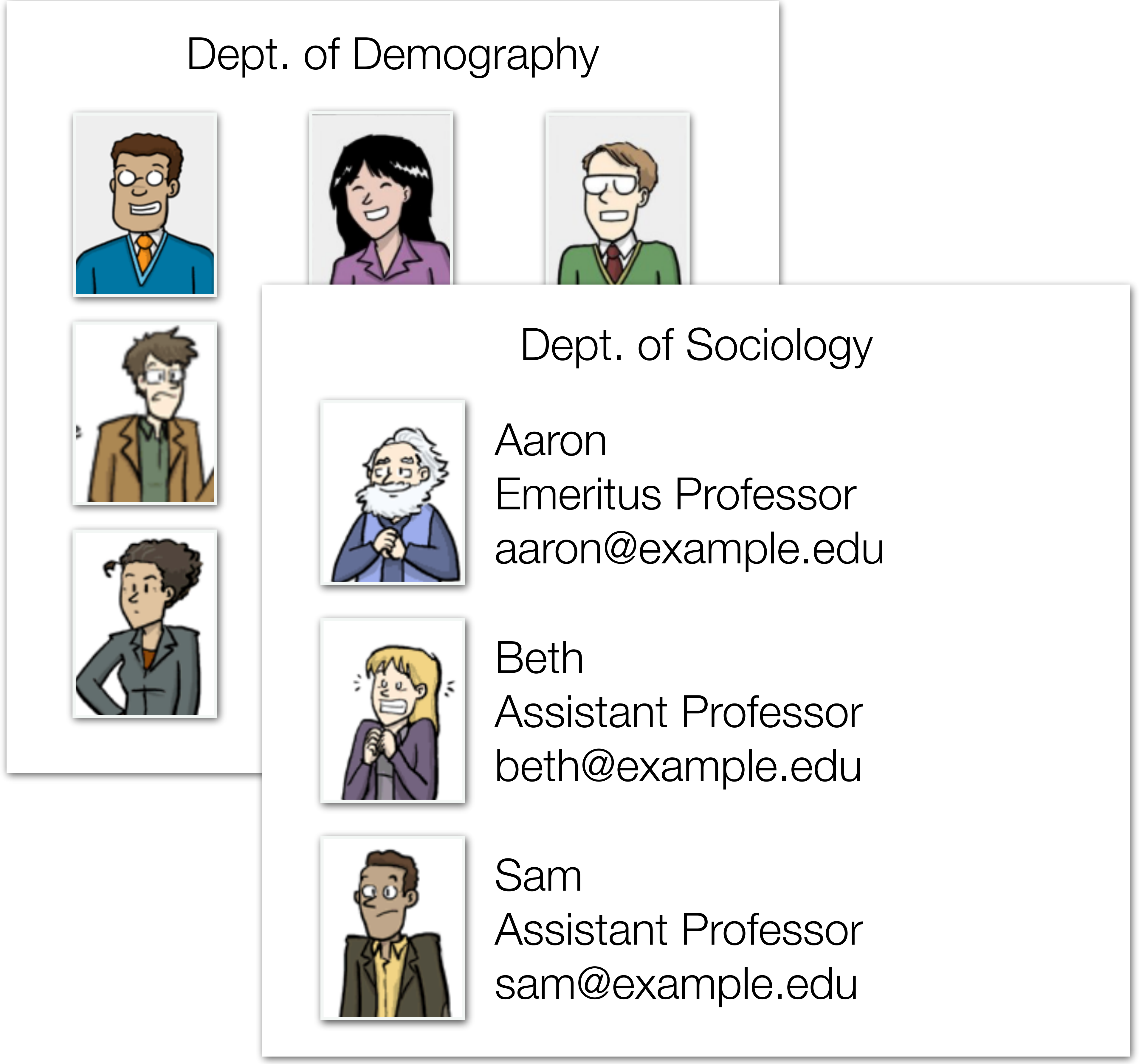
Retained faculty
(in 2017 & in 2011)

Departed faculty
(in 2011 & not in 2017)



Overall attrition for women is slightly higher than men (15.5% vs 14.3%)

Future Work



Dept. of Demography

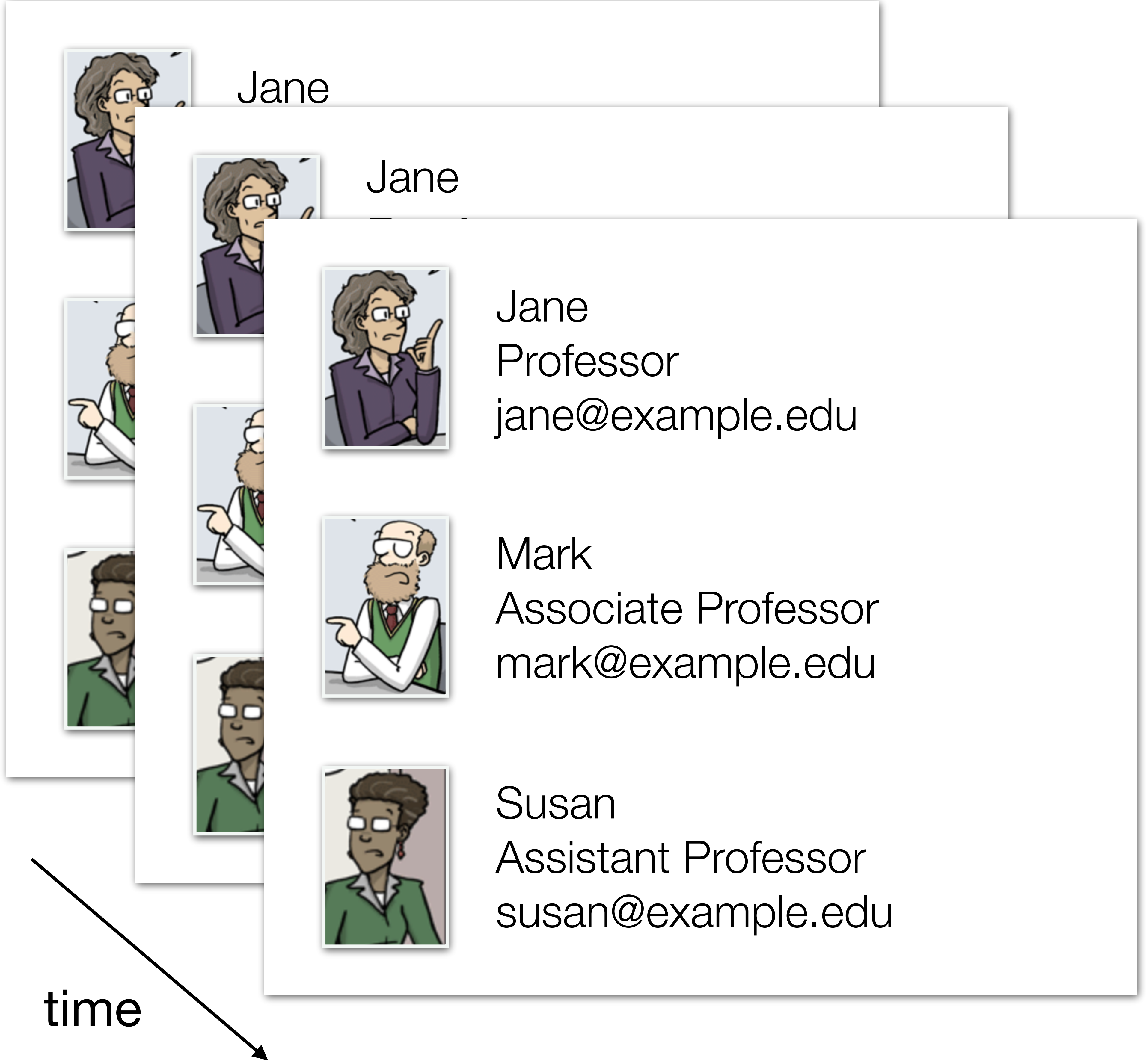
Dept. of Sociology

Aaron
Emeritus Professor
aaron@example.edu

Beth
Assistant Professor
beth@example.edu

Sam
Assistant Professor
sam@example.edu

Expand support to other academic fields



Jane

Jane

Jane
Professor
jane@example.edu

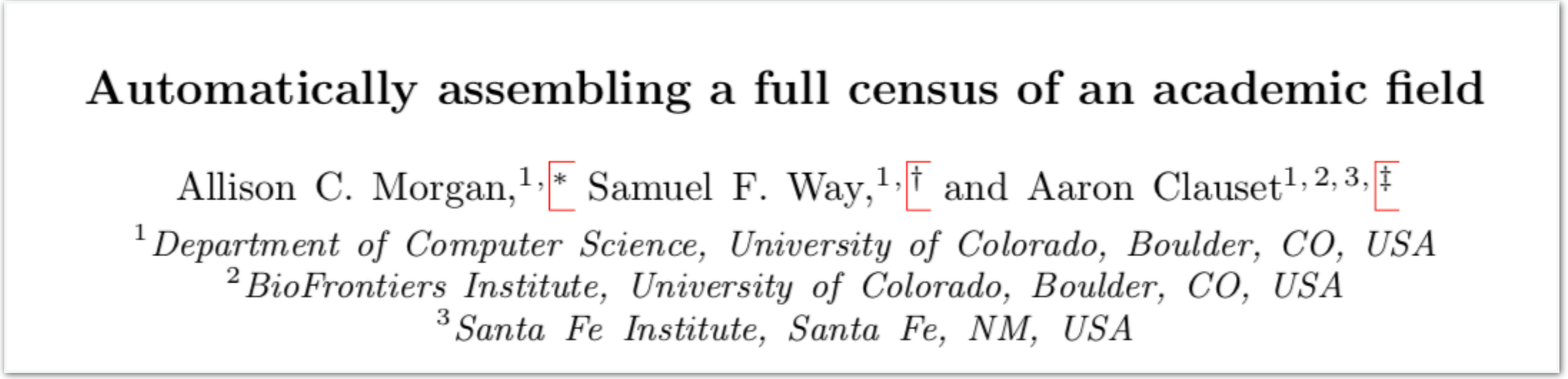
Mark
Associate Professor
mark@example.edu

Susan
Assistant Professor
susan@example.edu

Use the InternetArchive to collect the historical data

Cartoons by Jorge Chan; phdcomics.com

Thanks!



Automatically assembling a full census of an academic field

Allison C. Morgan,[1,*] Samuel F. Way,[1,†] and Aaron Clauset[1,2,3,‡]

[1] *Department of Computer Science, University of Colorado, Boulder, CO, USA*
[2] *BioFrontiers Institute, University of Colorado, Boulder, CO, USA*
[3] *Santa Fe Institute, Santa Fe, NM, USA*

https://arxiv.org/abs/1804.02760

Prof. Aaron Clauset
PhD Computer Science
aaron.clauset@colorado.edu

Dr. Sam Way
PhD Computer Science
samuel.way@colorado.edu

University of Colorado **Boulder**