ReSnag

Scraping NSF Awards to Create Database of Active STEM Researchers

Nathan Stevens

Agenda

- Introduction
- Challenges
- Approach
- Use Cases
- Next Steps

Introduction

There a numerous use cases for having a searchable database of active STEM faculty/researchers. For example:

- Targeted Marketing
- Selecting graduate program and mentor
- Getting overview of active research areas

Challenge

Aside from manually populating such a database by visiting faculty profiles, a more efficient approach is to use web scraping. However, web scraping profiles presented multiple challenges.

1. No central list of faculty with links to their profiles
2. No standardized format for profile pages
3. No easy way to get activity level

Approach

Follow The Money!

BeautifulSoup

Keywords Page

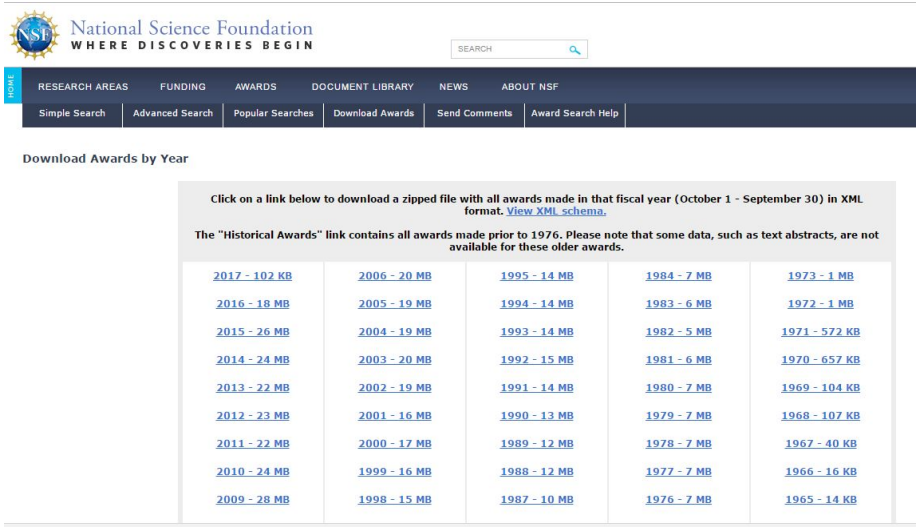
NSF Awards Page → Spider → Pipelines

- Download Zip File
- Process XML
- Generate Item Records

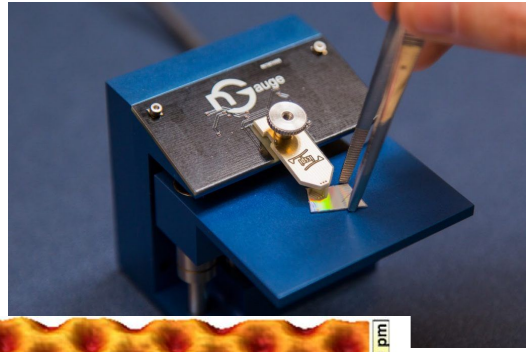
Save To MongoDB
Years: 2000 - 2017
~200,000 Documents

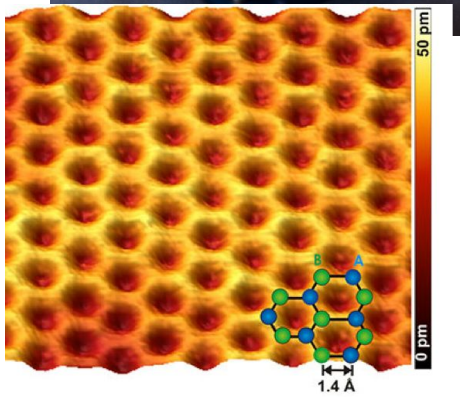
R Script

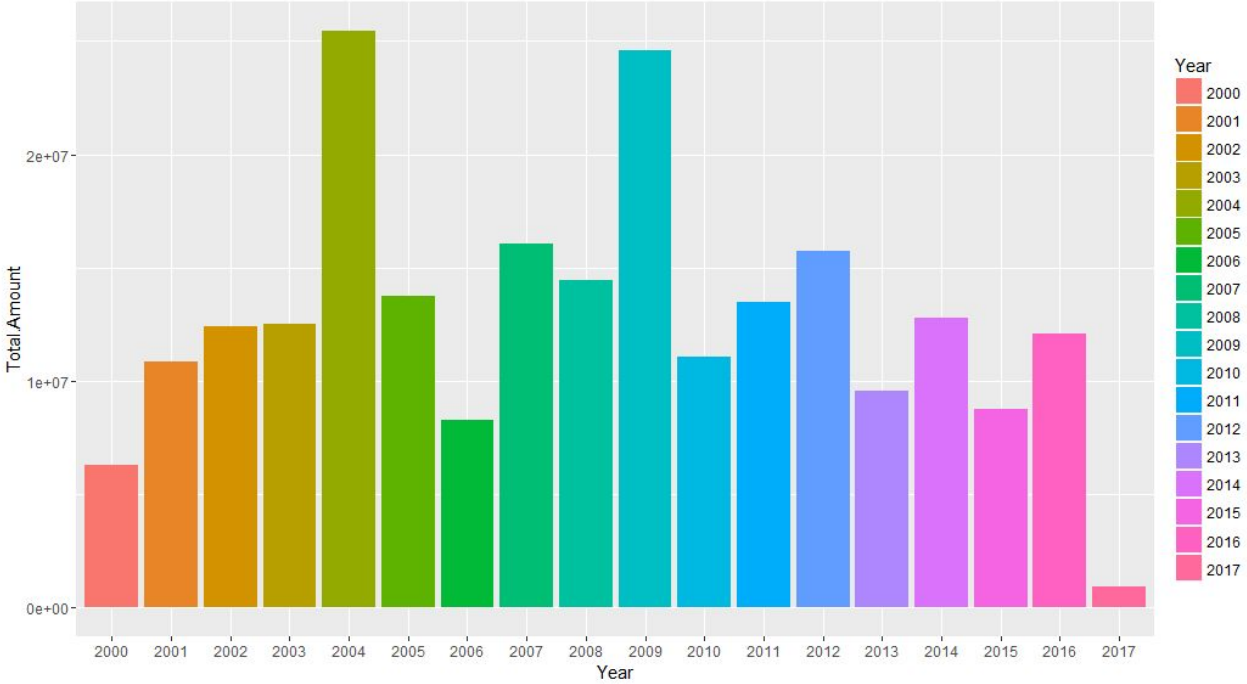
Python



Use Case -- Acme AFMs





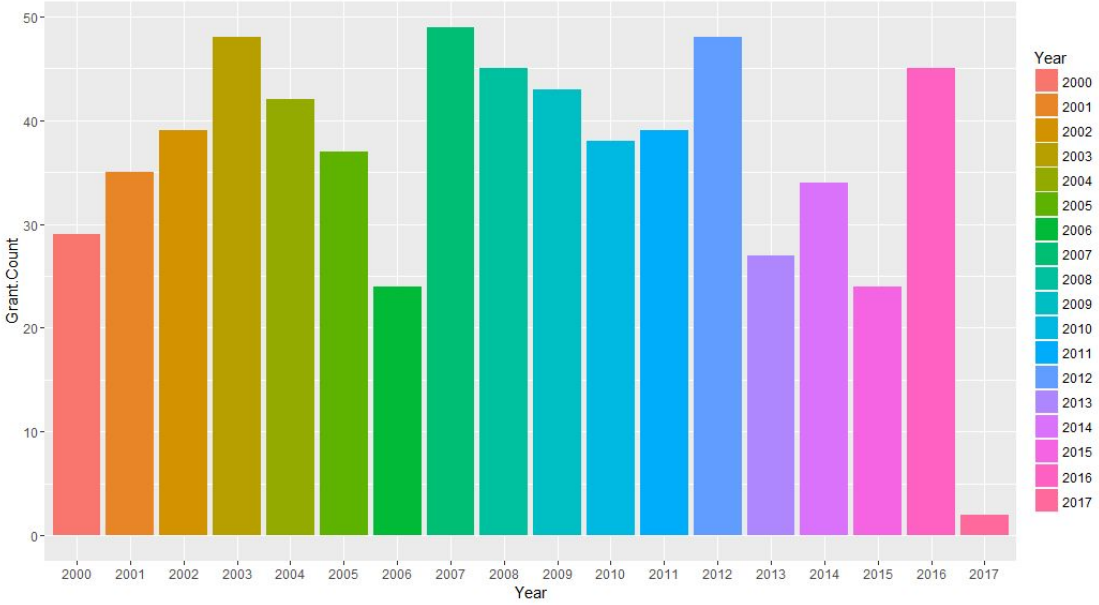


Credit: zmescience.com

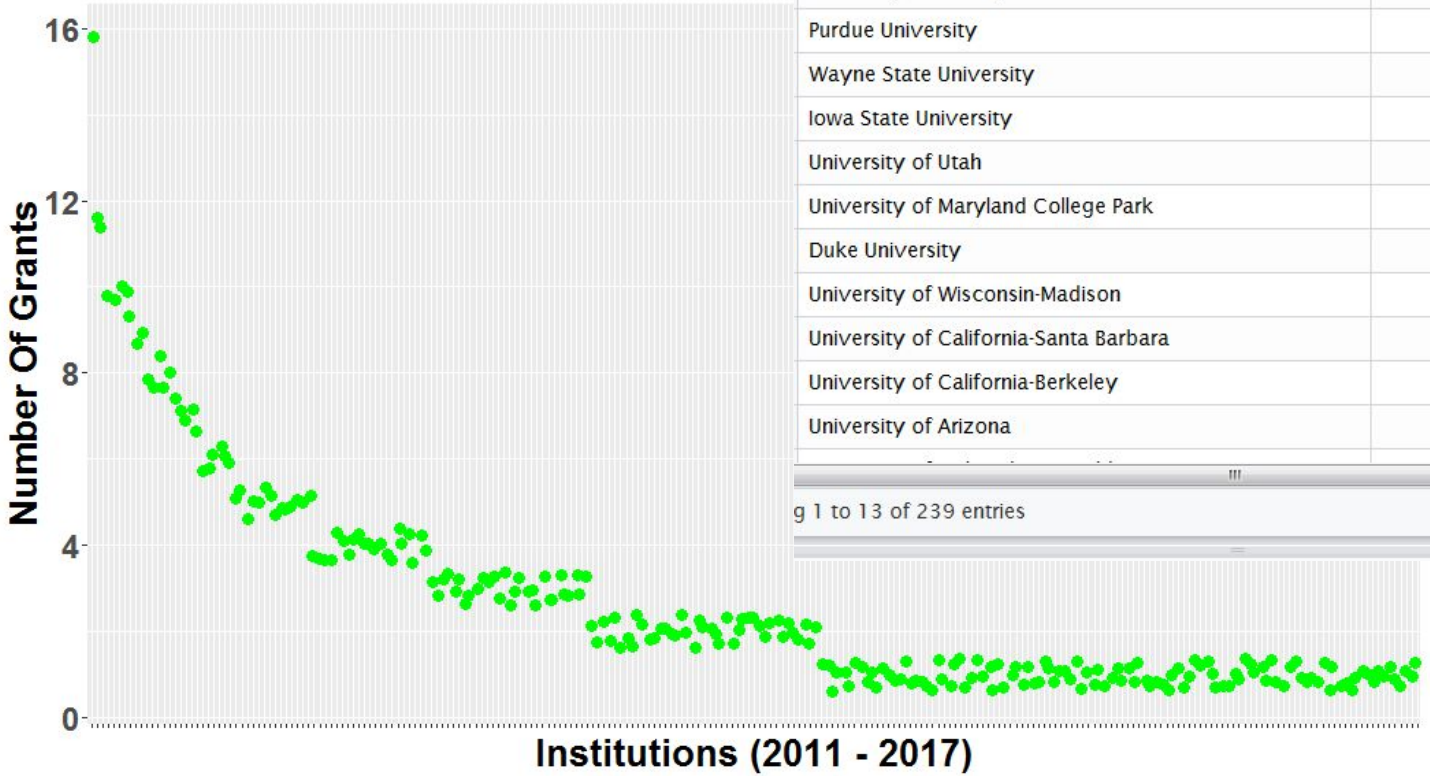
Is There Demand?

Filter

	Year	Total.Amount	Grant.Count
6	2005	13754452	37
7	2006	8273990	24
8	2007	16066467	49
9	2008	14422190	45
10	2009	24567272	43
11	2010	11079239	38
12	2011	13478522	39
13	2012	15752549	48
14	2013	9569074	27
15	2014	12793917	34
16	2015	8741193	24
17	2016	12065259	45
18	2017	945636	2



Who To Speak To?



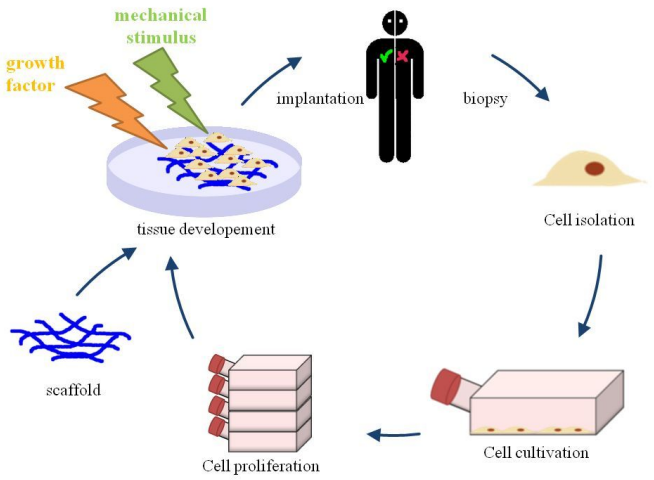
Filter

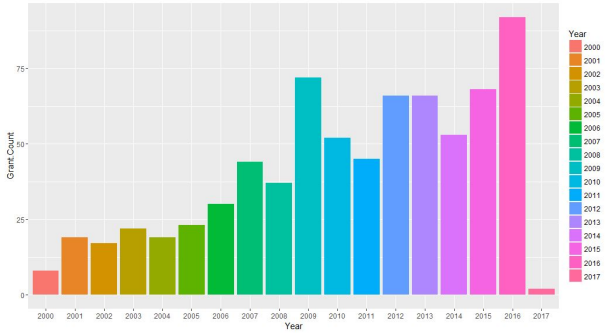
Institution	Total.Amount	Grant.Count
Georgia Tech Research Corporation	6381313	16
University of Pennsylvania	4266303	12
Purdue University	3647076	11
Wayne State University	3454985	10
Iowa State University	2590193	10
University of Utah	2512572	10
University of Maryland College Park	2487011	10
Duke University	4307761	9
University of Wisconsin-Madison	2648613	9
University of California-Santa Barbara	2638962	9
University of California-Berkeley	16042686	8
University of Arizona	3098679	8

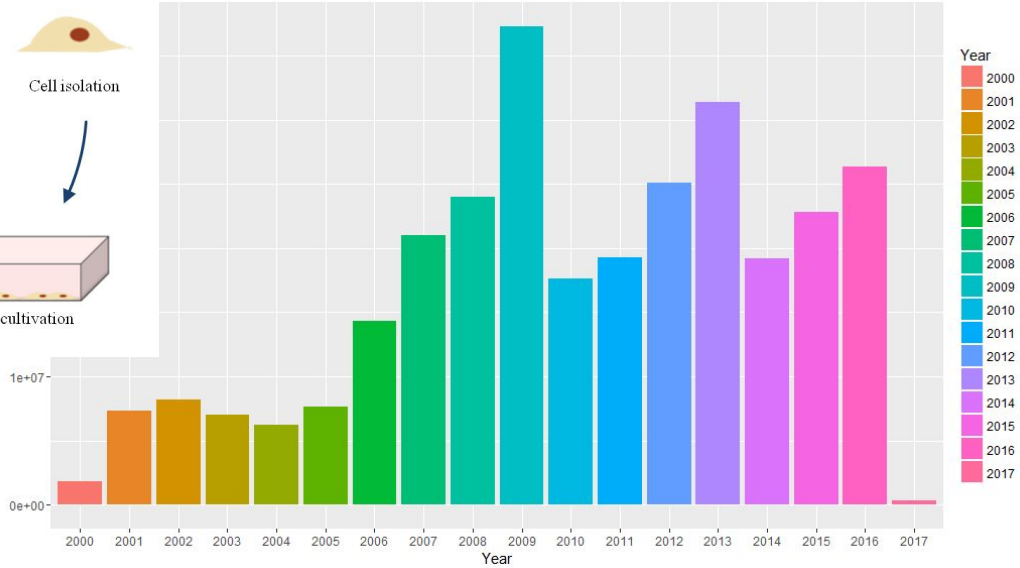
g 1 to 13 of 239 entries

Use Case 2 -- Which Grad School?

Tissue Engineering







Who Are The Top Schools?

Institution	Total.Amount	Grant.Count
William Marsh Rice University	1670000	3
Trustees of Boston University	1449731	2
Massachusetts Institute of Technology	1445000	3
University of Pennsylvania	1337345	2
California Institute of Technology	1132649	2
Southern Methodist University	650000	1
Virginia Polytechnic Institute and State University	649999	2
Colorado State University	638997	1
University of Colorado at Boulder	609438	2
Arizona State University	594884	2
Princeton University	586140	1
Lehigh University	510240	1
Tuskegee University	507374	1

1 to 13 of 52 entries

Year 2015

Year 2016

Filter

Institution	Total.Amount	Grant.Count
University of New Mexico	3999914	2
Rutgers University New Brunswick	1619473	5
University of Akron	1356820	6
University of Alabama at Birmingham	1089790	3
University of Notre Dame	1000000	2
Washington State University	950002	2
University of California-San Diego	923899	3
Rowan University	903000	2
University of Washington	873086	2
University of Delaware	850000	2
University of Pittsburgh	808374	2
William Marsh Rice University	804880	2
North Carolina State University	799876	2

1 to 13 of 43 entries

Next Steps

- Develop Interactive web application
- Use machine learning for keyword tagging
- Add additional data: more years, other agencies, publication
- Explore predictive modeling