Dimensions of Computing: A Data Science Perspective

Victoria Stodden
Associate Professor
School of Information Sciences
University of Illinois at Urbana-Champaign

Workshop on the Growth of CS Undergraduate Enrollments
Computer Science and Telecommunications Board
Division on Engineering and Physical Sciences
The National Academies of Sciences, Engineering, and Medicine

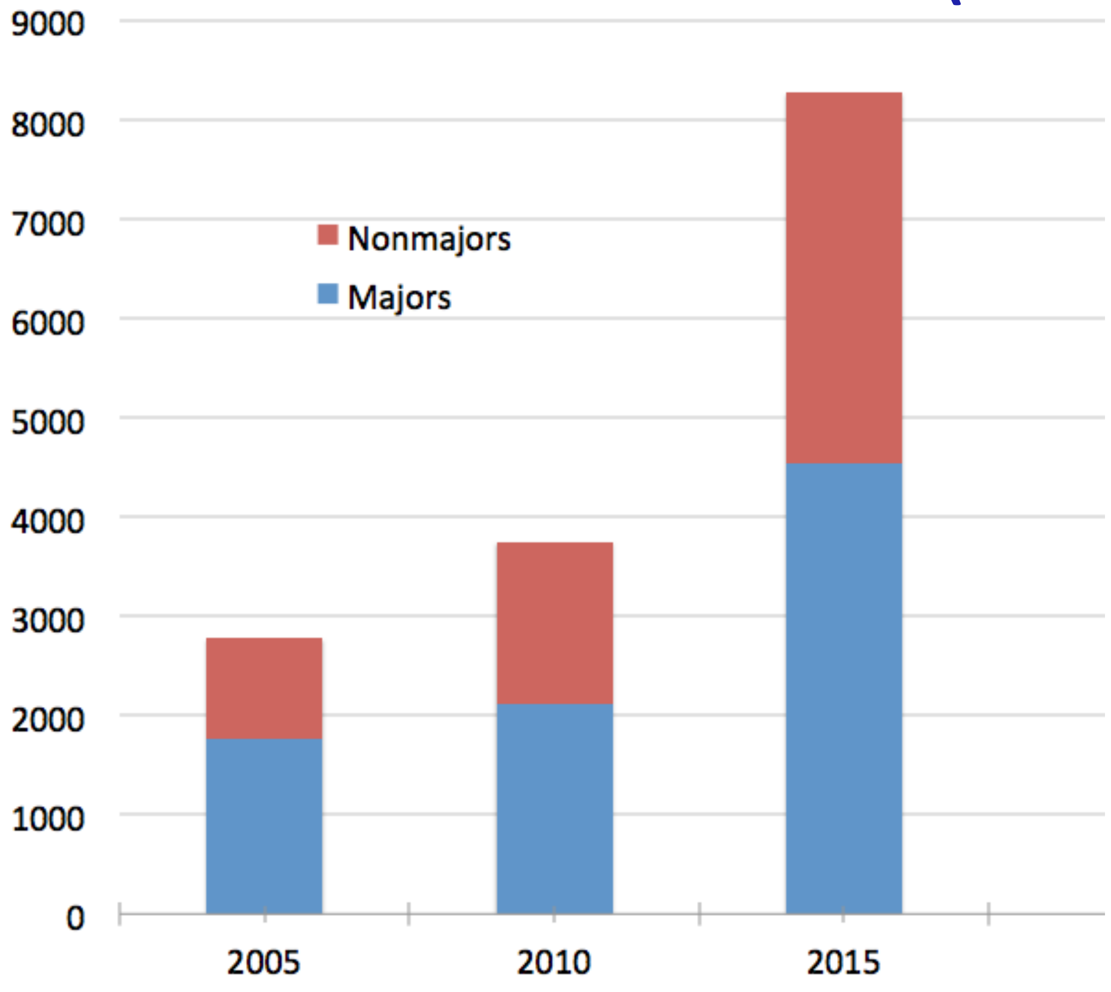
August 15, 2016

Non-majors Enrolling in CS Courses

Many reasons for the growth in non-major enrollment, e.g. programming and CS skills highly valued in the workplace.

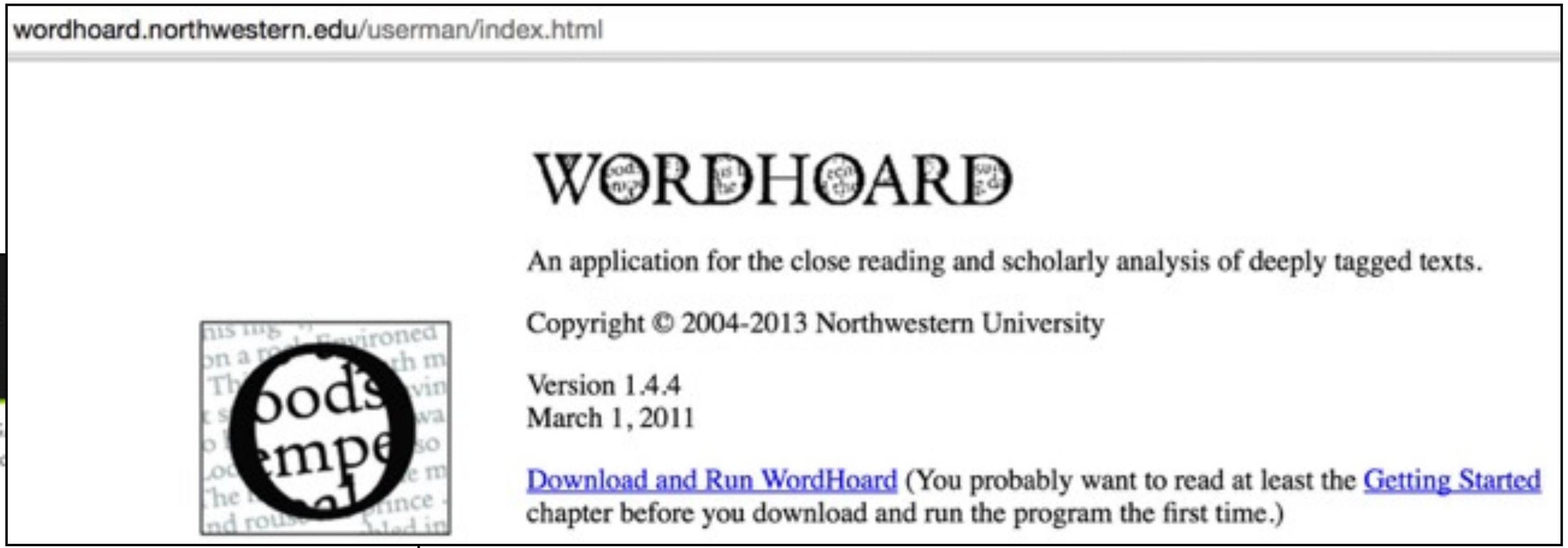
I focus on data science as a driver for non-major enrollment growth.

Students in 'Typical' Mid-Level Courses
(in 44 units)



Source: Tracy Camp, CRA Presentation 2016

Data- or Computationally-enabled Research is Pervasive



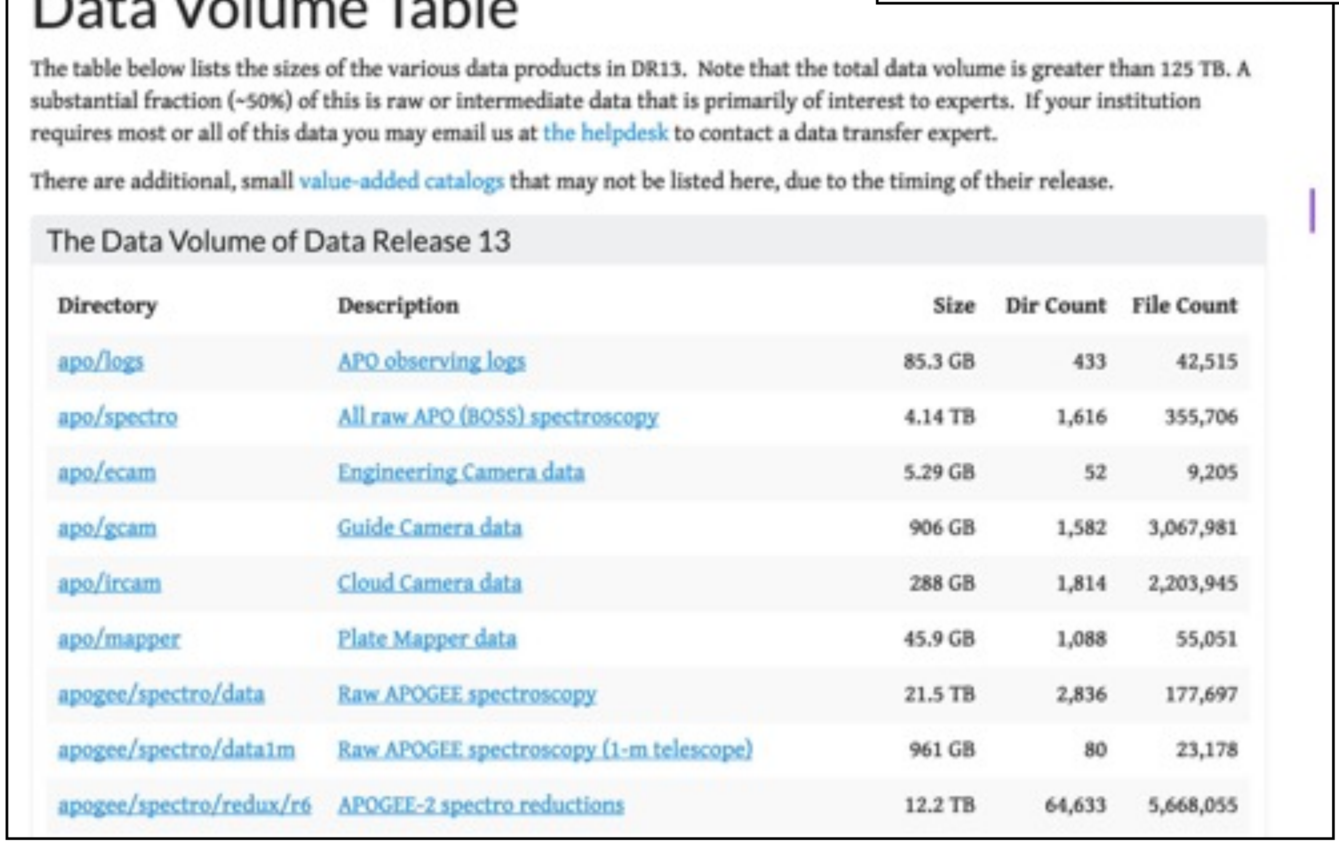
WORDHOARD

An application for the close reading and scholarly analysis of deeply tagged texts.

Copyright © 2004-2013 Northwestern University

Version 1.4.4
March 1, 2011

Download and Run WordHoard (You probably want to read at least the Getting Started chapter before you download and run the program the first time.)



Data Volume Table

The table below lists the sizes of the various data products in DR13. Note that the total data volume is greater than 125 TB. A substantial fraction (~50%) of this is raw or intermediate data that is primarily of interest to experts. If your institution requires most or all of this data you may email us at the helpdesk to contact a data transfer expert.

There are additional, small value-added catalogs that may not be listed here, due to the timing of their release.

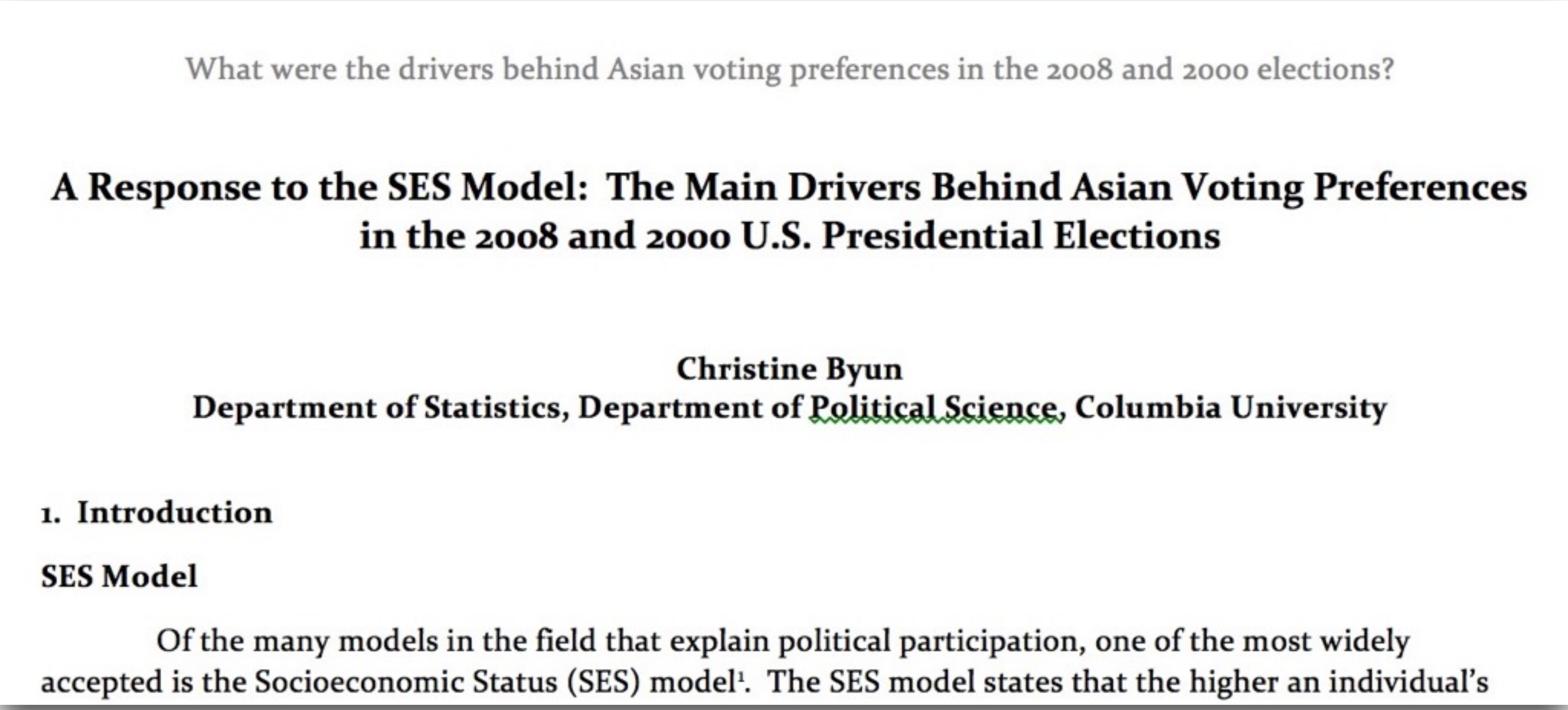
The Data Volume of Data Release 13

Directory	Description	Size	Dir Count	File Count
apo/logs	APO observing logs	85.3 GB	433	42,515
apo/spectro	All raw APO (BOSS) spectroscopy	4.14 TB	1,616	355,706
apo/ecam	Engineering Camera data	5.29 GB	52	9,205
apo/gcam	Guide Camera data	906 GB	1,582	3,067,981
apo/ircam	Cloud Camera data	288 GB	1,814	2,203,945
apo/mapper	Plate Mapper data	45.9 GB	1,088	55,051
apogee/spectro/data	Raw APOGEE spectroscopy	21.5 TB	2,836	177,697
apogee/spectro/data1m	Raw APOGEE spectroscopy (1-m telescope)	961 GB	80	23,178
apogee/spectro/redux/r6	APOGEE-2 spectro reductions	12.2 TB	64,633	5,668,055

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

The software contains "ideas that enable biology…" *Stories from the Supplement, 2013*

A Story of an Undergraduate



What were the drivers behind Asian voting preferences in the 2008 and 2000 elections?

A Response to the SES Model: The Main Drivers Behind Asian Voting Preferences in the 2008 and 2000 U.S. Presidential Elections

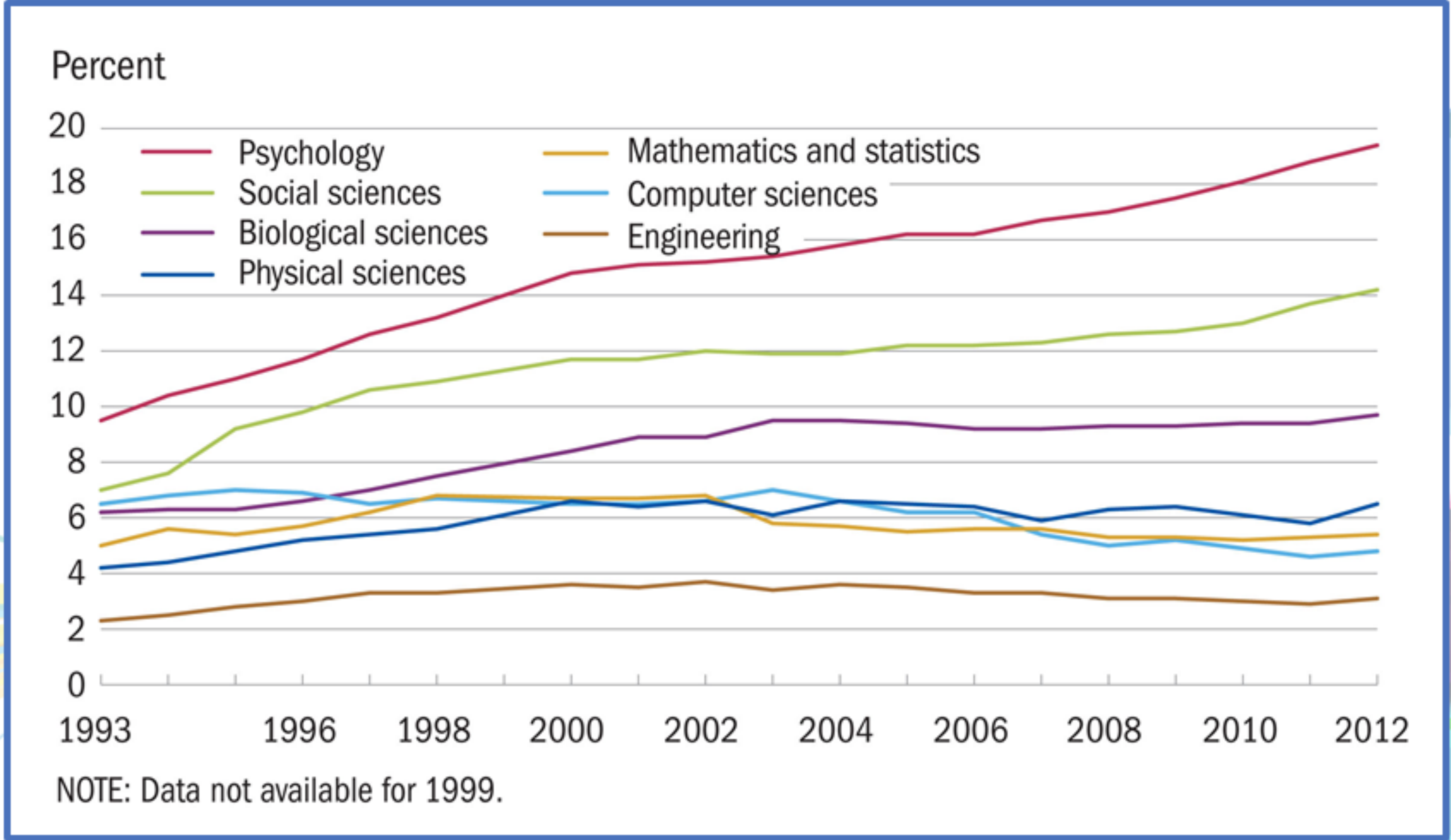
Christine Byun
Department of Statistics, Department of Political Science, Columbia University

1. Introduction

SES Model

Of the many models in the field that explain political participation, one of the most widely accepted is the Socioeconomic Status (SES) model[1]. The SES model states that the higher an individual's

Science and engineering bachelor's degrees earned by underrepresented minority women, by field: 1993–2012



NOTE: Data not available for 1999.

Women, Minorities, and Persons with Disabilities in Science and Engineering: 2015
www.nsf.gov/statistics/wmpd/

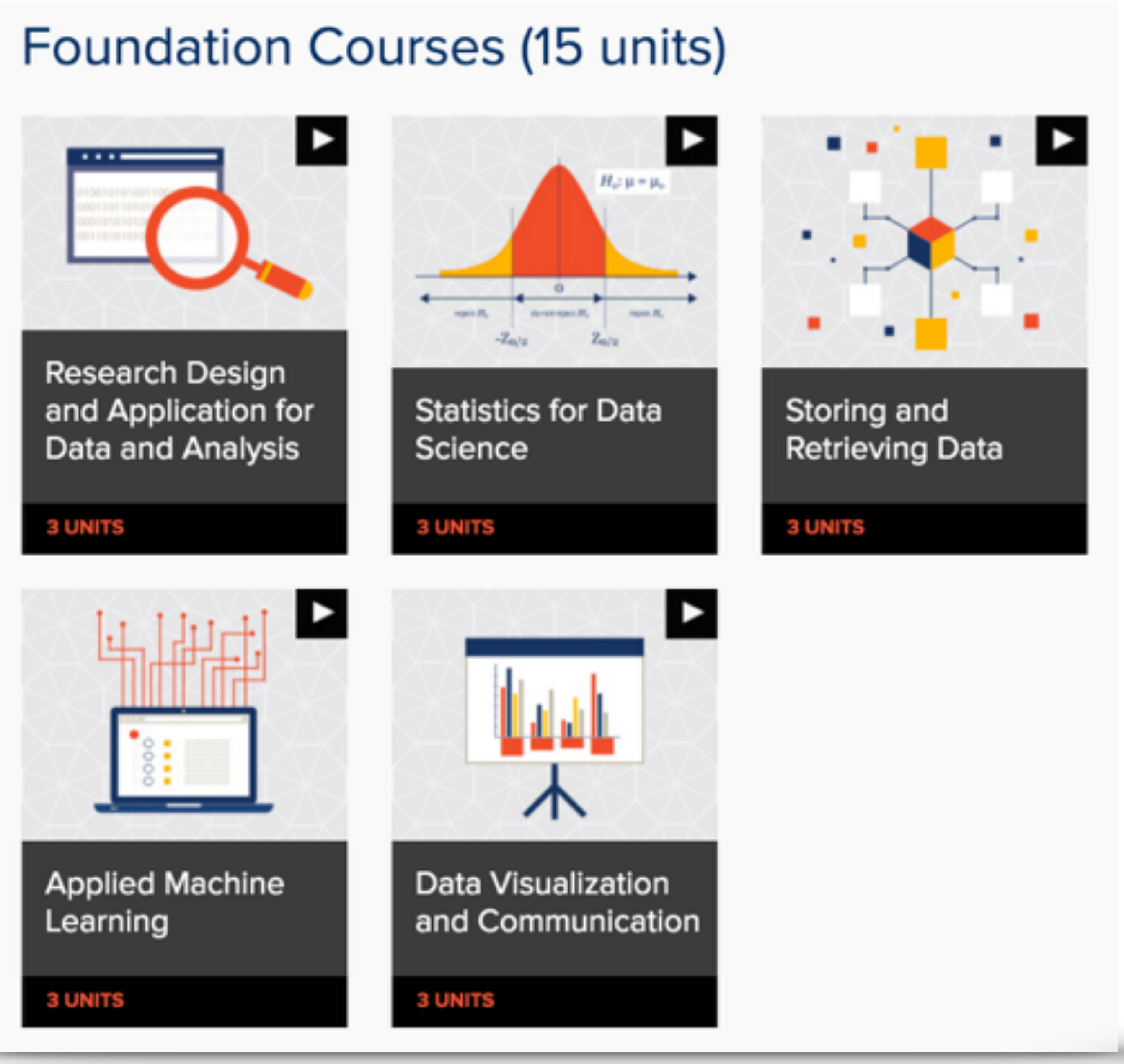
Why are Non-majors Enrolling?

Data Science a new and compelling interest for undergraduates, cuts across domain research areas (climate, energy use, water supply, voter patterns, etc),

Data Science uses foundational CS techniques, increasing demand for CS courses such as:

- software design, data structures, building packages and libraries, …
- interpreted languages: python, R, MATLAB, …
- algorithms, machine learning, scalability, …
- database management systems, …
- HPC and cloud computing, networks, …

Master of Information and Data Science, UC Berkeley



Foundation Courses (15 units)

Research Design and Application for Data and Analysis

3 UNITS

Statistics for Data Science

3 UNITS

Storing and Retrieving Data

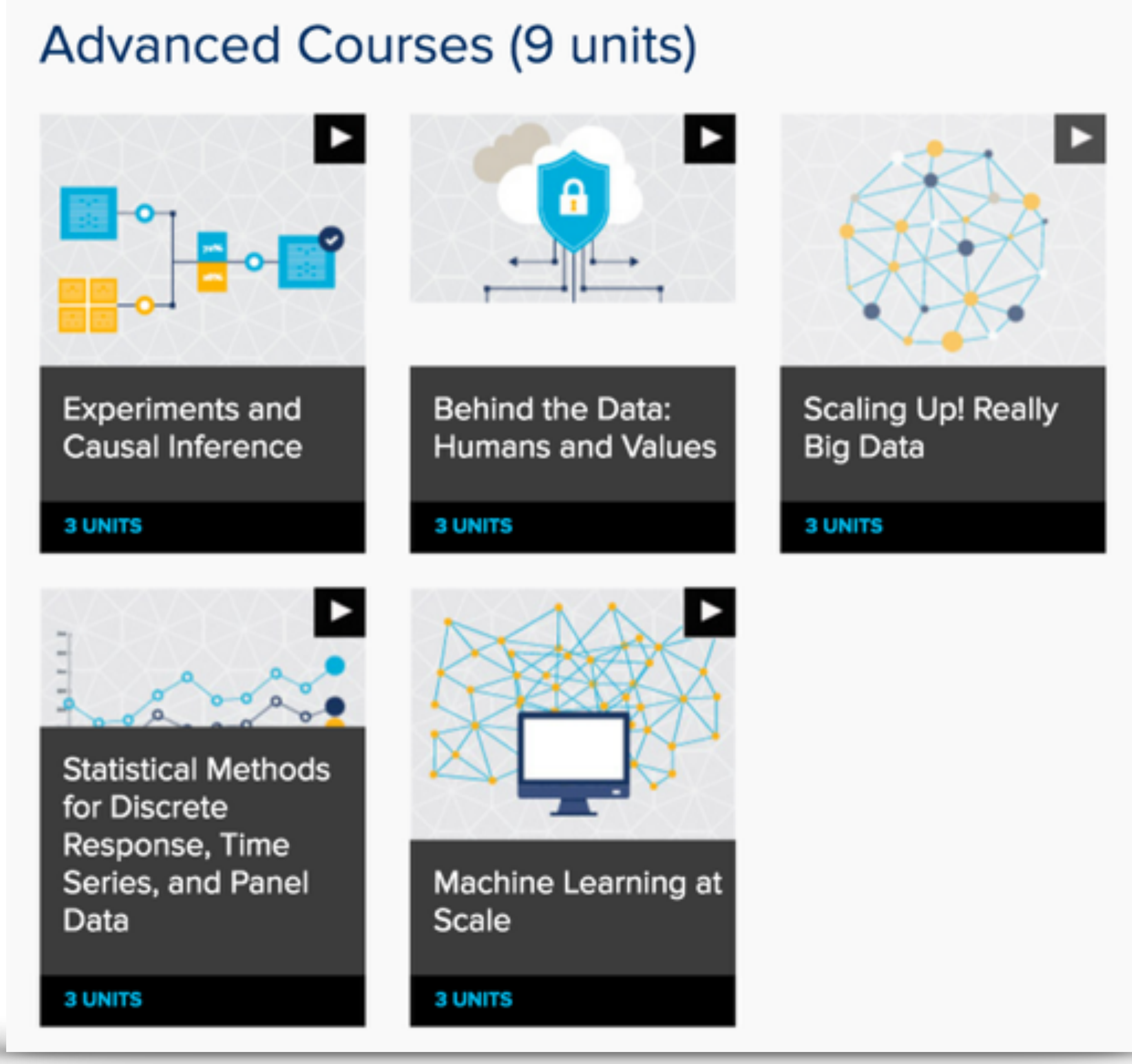
3 UNITS

Applied Machine Learning

3 UNITS

Data Visualization and Communication

3 UNITS



Advanced Courses (9 units)

Experiments and Causal Inference

3 UNITS

Behind the Data: Humans and Values

3 UNITS

Scaling Up! Really Big Data

3 UNITS

Statistical Methods for Discrete Response, Time Series, and Panel Data

3 UNITS

Machine Learning at Scale

3 UNITS

Emerging Computational Science Infrastructure

Dissemination Platforms

- ResearchCompendia.org
- MLOSS.org
- Open Science Framework
- IPOL
- thedatahub.org
- Madagascar
- nanoHUB.org
- RunMyCode.org

Workflow Tracking and Research Environments

- Vistrails
- Galaxy
- Pegasus
- Kepler
- GenePattern
- Kurator
- CDE
- Sumatra
- Jupyter
- Taverna

Embedded Publishing

- Verifiable Computational Research
- Collage Authoring Environment
- SOLE
- SHARE
- knitR
- Sweave
- clearScience
- Paper of the Future