MIT MEDIA LAB

CSAIL

Democratizing Data Science

Sophie Chou
@mpetitchou

William Li
@williampli

Ramesh Sridharan
@tweetsbyramesh

{soph,wpli,rameshvs}@mit.edu

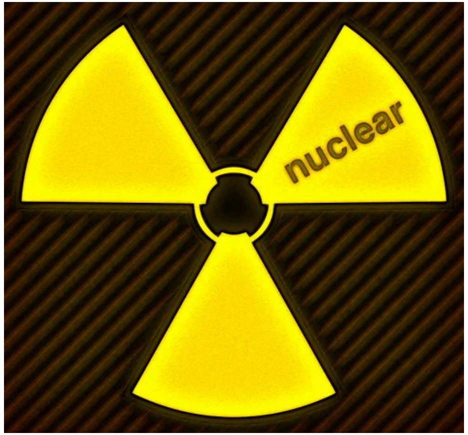
Some Links

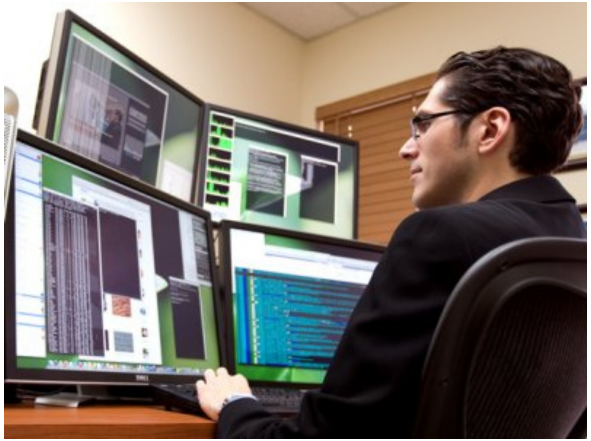
- Paper: **bit.ly/DDSpaper**

- Cathy O'Neil's Blog (@mathbabedotorg): **bit.ly/DDSblog**

- Twitter: **@mpetitchou, @williampli, @tweetsbyramesh**

[insert technology] for Social Good





Technology is a force multiplier,
for better or worse

What is Data Science?

- Our working definition: transforming data into insights/solutions/products
    1. collection & storage
    2. cleaning & structuring
    3. analyzing & finding patterns
    4. visualizing & communicating results

What is “Democratizing Data Science”?

The application of data science is undemocratic: problems that promote the common good receive insufficient attention.

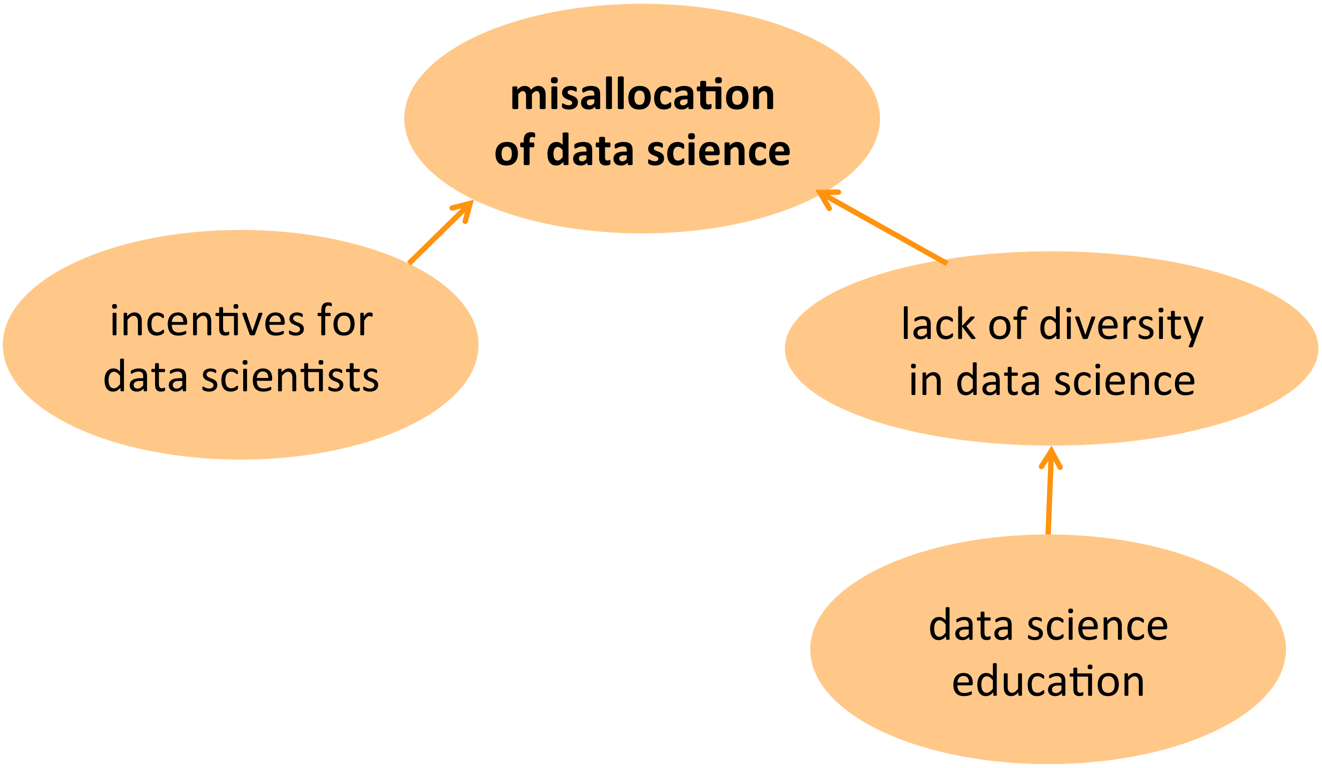
Ford/MacArthur Foundation, 2013

“Technology talent is a key need in government and civil society,

but the current state of the pipeline is inadequate to meet that need.”

Source: http://bit.ly/FordMacArthurReport

Why?



misallocation of data science

incentives for data scientists

lack of diversity in data science

data science education

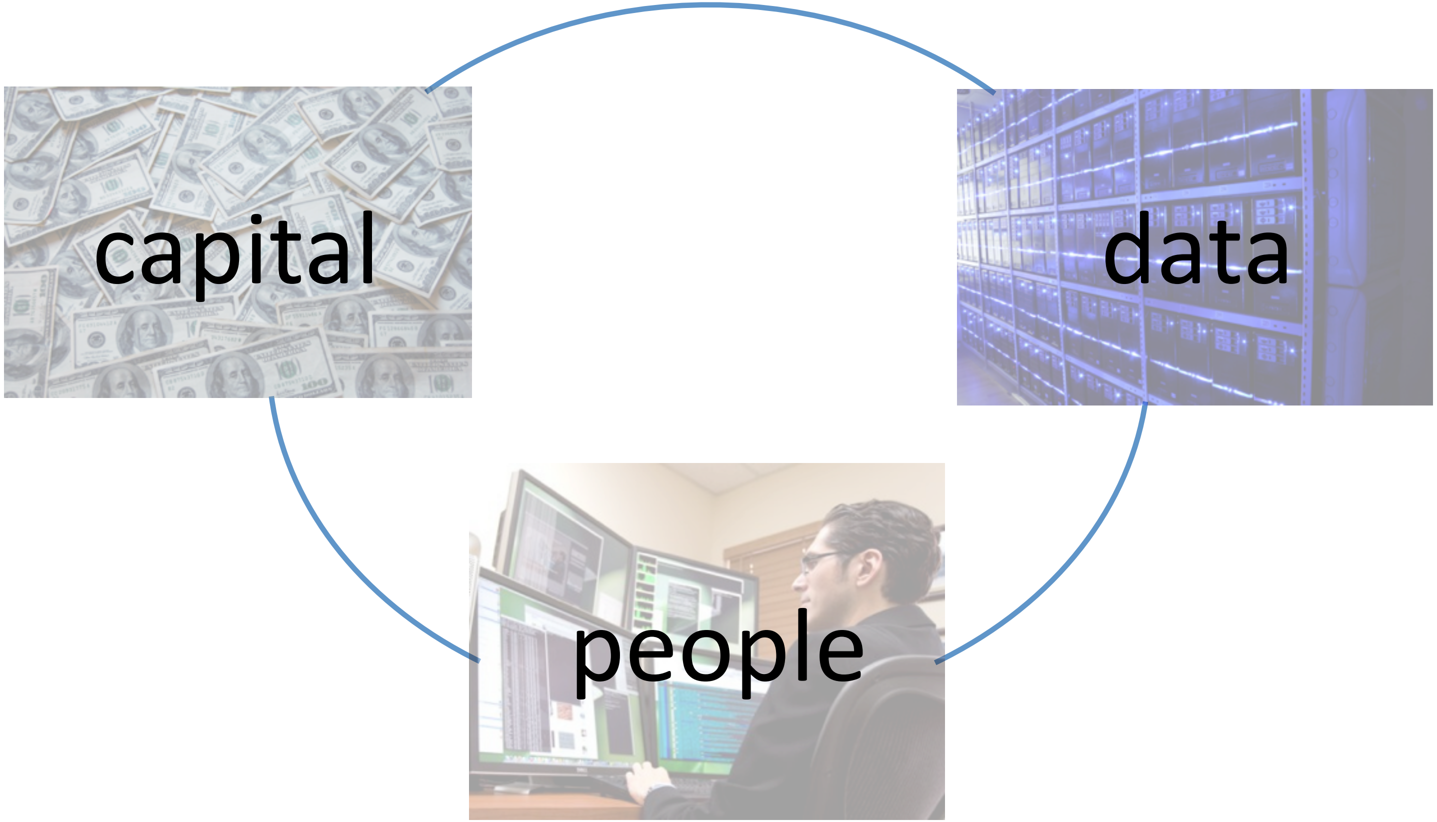
Outline

- Incentives for data scientists

- Democratizing data science education

- Potential solutions

Sources of Power in Data Science

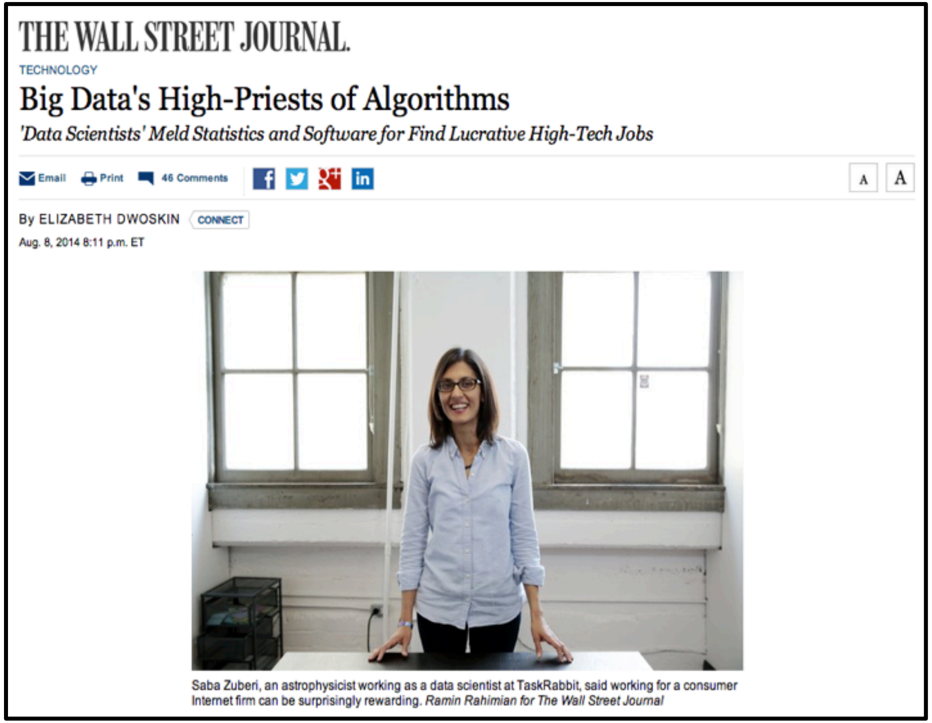


capital

data

people

Human expertise



THE WALL STREET JOURNAL.

TECHNOLOGY

Big Data's High-Priests of Algorithms

'Data Scientists' Meld Statistics and Software for Find Lucrative High-Tech Jobs

By ELIZABETH DWOSKIN

Aug. 8, 2014 8:11 p.m. ET

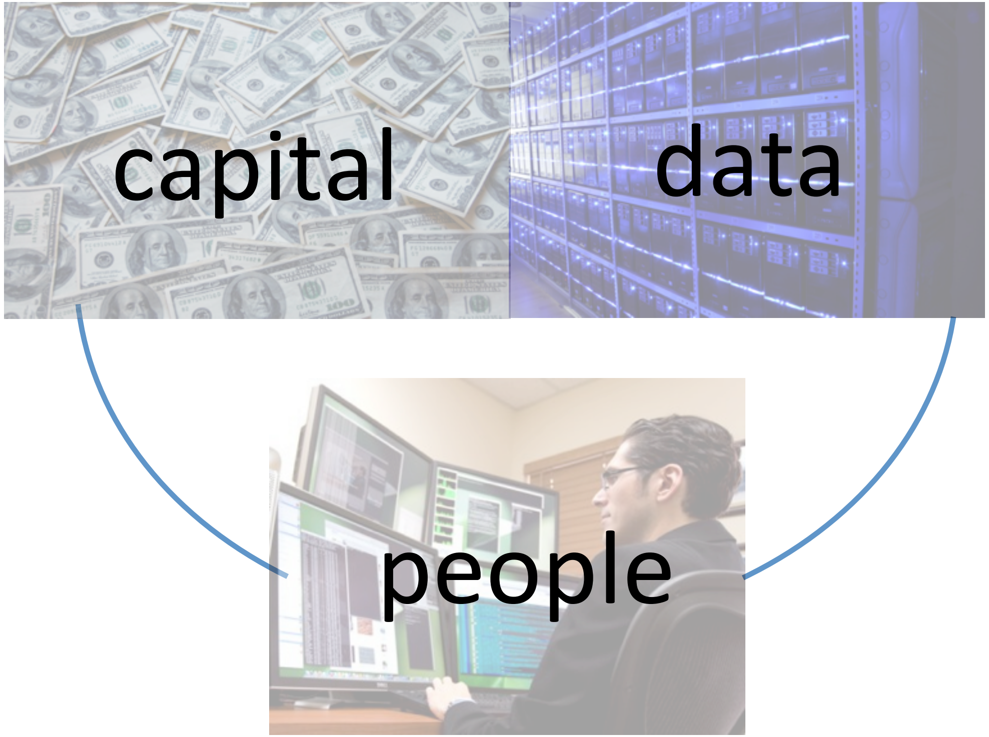
Saba Zuberi, an astrophysicist working as a data scientist at TaskRabbit, said working for a consumer Internet firm can be surprisingly rewarding. *Ramin Rahimian for The Wall Street Journal*

Source: http://bit.ly/WSJDataScience

"...[salaries] between $200,000 and $300,000 a year...100 recruiter emails a day"

"...working for a consumer Internet firm can be surprisingly rewarding."

Sources of Power in Data Science



capital

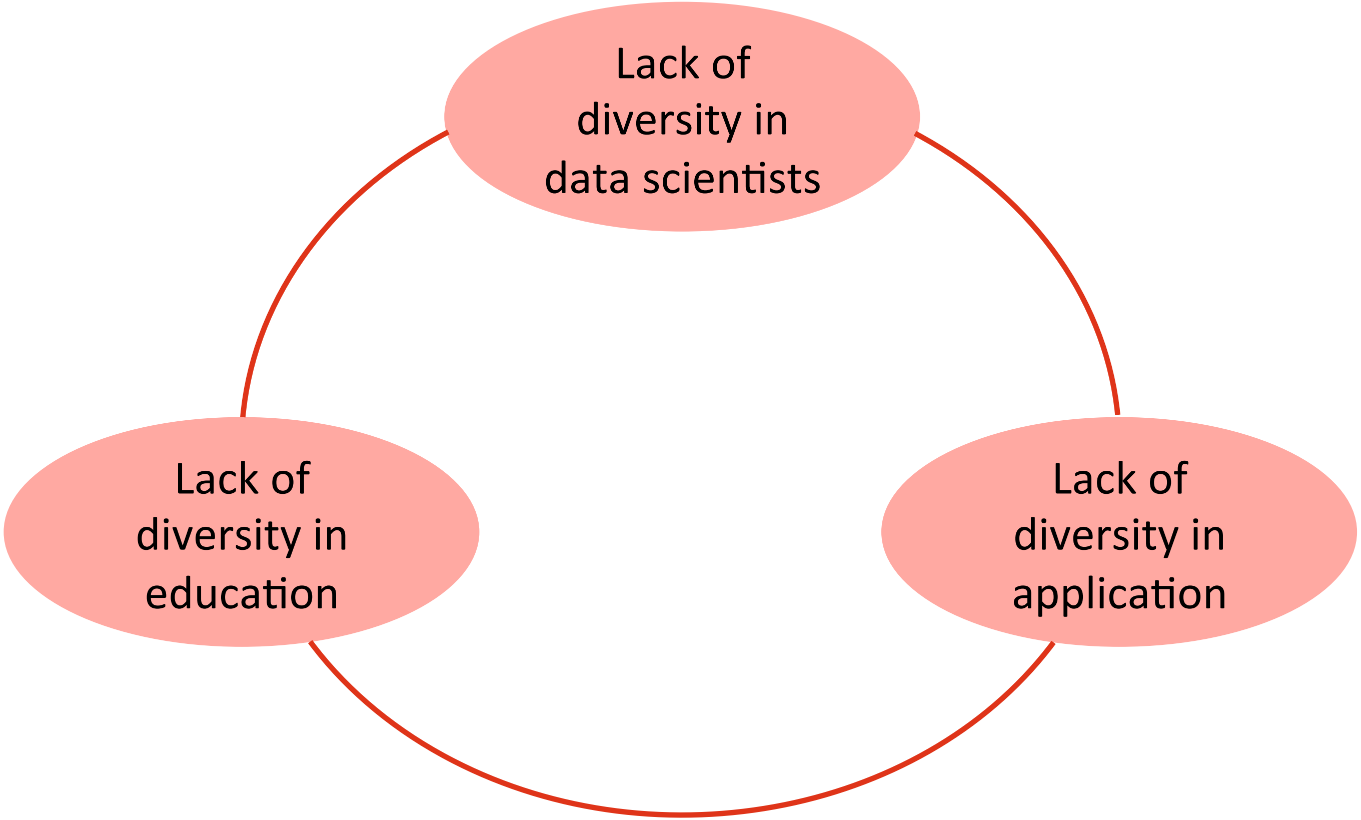
data

people

Outline

- Structural inequalities in data science
- Democratizing data science education
- Potential solutions

Negative feedback loop



Lack of diversity in data scientists

Lack of diversity in education

Lack of diversity in application

Why care?

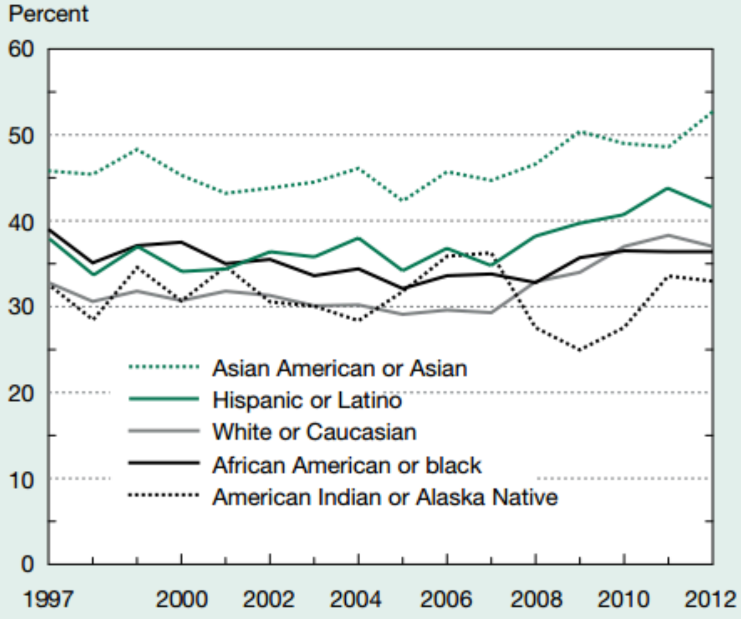
- Diversity is key to innovation (Forbes Insights, 2011)
- Lack of diversity perpetuates misallocation

Racial inequality in science and engineering

- Misrepresentation isn't going away

- Fewer minorities receive degrees

Freshmen intending S&E major, by race and ethnicity: 1997–2012



NOTE: In 2001, Native Hawaiian or Pacific Islander was added as a category under Asian American or Asian.

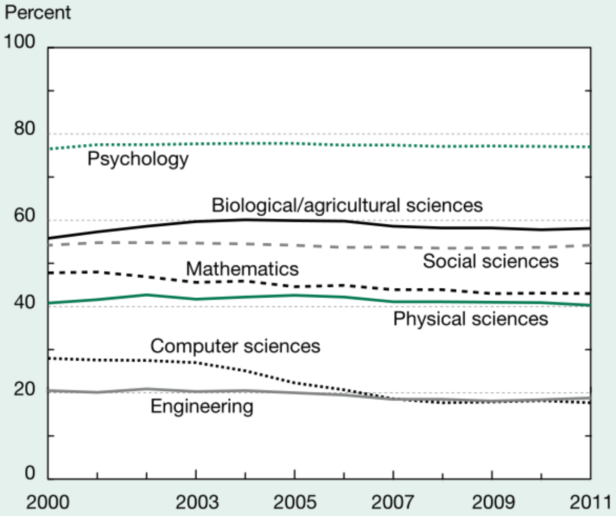
Source: NSF Science and Engineering Indicators 2014

Women in Computing

- 1985: 37%
  2000: 29%
  2014: 18%

- 2000 to 2014:
  - 5% decrease in math
  - 2% decrease in engineering

Women's share of S&E bachelor's degrees, by field: 2000–11

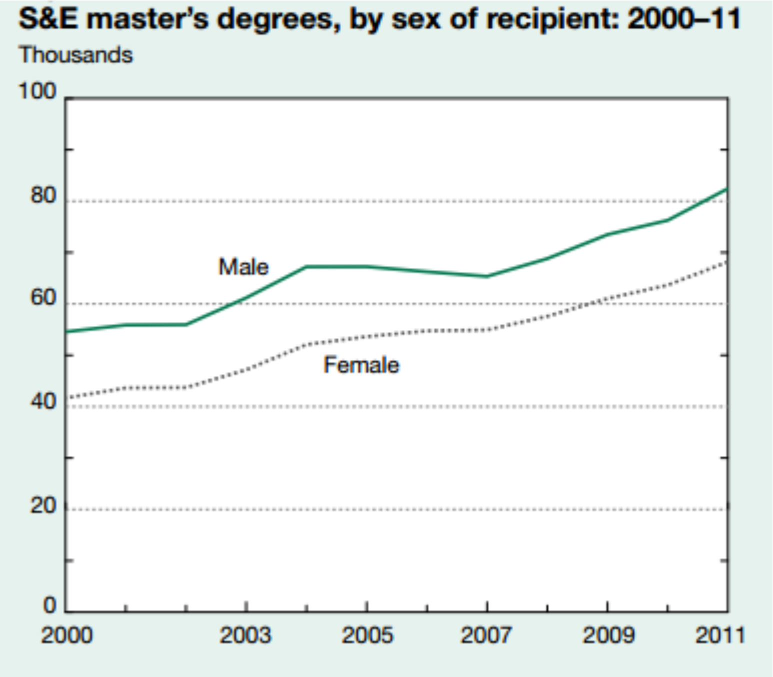


NOTE: Physical sciences include earth, atmospheric, and ocean sciences.

SOURCES: National Center for Education Statistics, Integrated Postsecondary Education Data System, Completions Survey; and National Science Foundation, National Center for Science and Engineering Statistics, WebCASPAR database, http://webcaspar.nsf.gov.

Source: NSF Science and Engineering Indicators 2014

Graduate degrees



- Women fare even worse
- 2x as many white males receiving degrees as *all minorities combined*
- Only 1 in 5 PhDs female

Source: NSF Science and Engineering Indicators 2014

Unlocking the Clubhouse: A Case Study

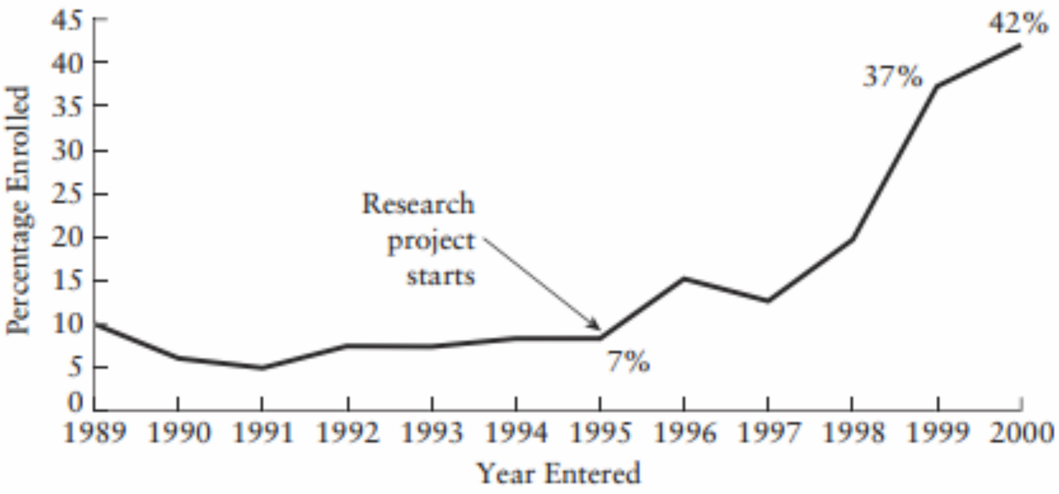


Figure 8.1
Enrollment trends for women entering the School of Computer Science

- 2014 incoming class: 40% women

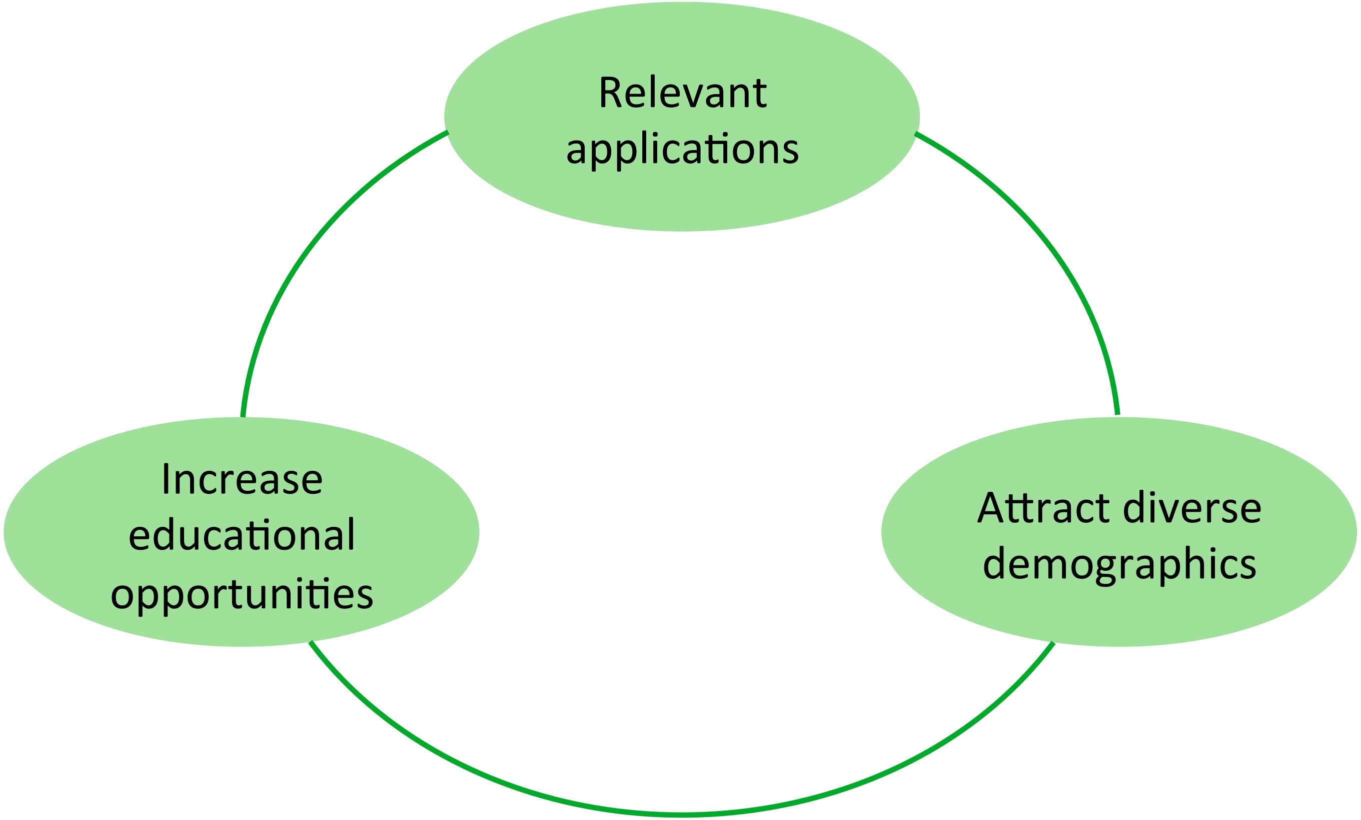
Source: Unlocking the Clubhouse, 2002

Triggering Positive Change

"insuring science and technology are considered in their social context may be the most important change that can be made in science teaching for all people, both male and female."

Source: Taggart & O'Gara, 2000

Positive feedback loop



Relevant applications

Attract diverse demographics

Increase educational opportunities

Outline

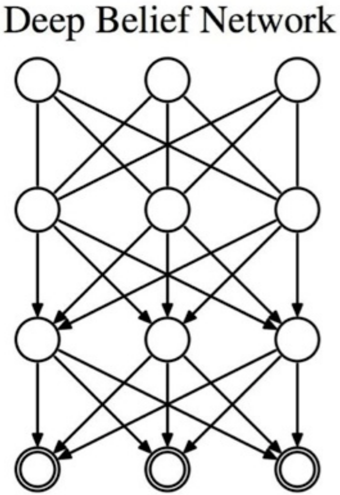
- Structural inequalities in data science

- Democratizing data science education

- Potential solutions

Technologies

- Recent technologies target broader audiences
- Often require significant technical literacy
- Can we broaden access further still?



amazon web services™

DataWrangler alpha

Crowd-based efforts

- Machine learning competitions
  - Can promote meaningful problems
  - Low barrier to entry
- Often run by for-profit entities
- Can we encourage more initiatives like KDD Cup 2014?

kaggle

2014 KDD Cup

Education: MOOCs

- Tremendous potential for reaching students
- Most often taken by professionals and people with graduate degrees

UDACITY

coursera

edX

Private Sector Opportunities

- Reaching out to underserved groups
  - Tap new markets

- Pro bono work could service groups and bring in new customers

- Meaningful “small data” to serve the long tail

Academic Research

- Research promoting social good is particularly accessible to academics
  - Social welfare problems often rely on public data
  - Academia is well-suited to interdisciplinary research
- Need for focus on meaningful problems

Your Solution Here!

- We believe the community has a responsibility to solve these problems
- Expertise in policy, business, statistics, healthcare, computer science will all be crucial

- Undemocratic inequalities persist in data science applications
- All of us can be part of the solution