



# The Data Science Track

Jeffrey Leek  
Johns Hopkins Bloomberg School of Public Health

# Why do data science?

"It is not the critic who counts: not the man who points out how the strong man stumbles or where the doer of deeds could have done better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood, who strives valiantly, who errs and comes up short again and again, because there is no effort without error or shortcoming, but who knows the great enthusiasms, the great devotions, who spends himself for a worthy cause; who, at the best, knows, in the end, the triumph of high achievement, and who, at the worst, if he fails, at least he fails while daring greatly, so that his place shall never be with those cold and timid souls who knew neither victory nor defeat."



*Theodore Roosevelt, 26th President of the United States*

[Statistics and the science game](#)

# The key challenge in data science

"Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew knew all of the given information in advance? Where you didn't have a surplus of information and have to filter it out, or you didn't have insufficient information and have to go find some?"



[Dan Myer, Mathematics Educator](#)

[The key word in data science is not data; it is science](#)

# About us

Data intensive statistics in biology and medicine

- Brian Caffo
  - Website <http://www.bcaffo.com/>
  - Twitter [@bcaffo](#)
  - Github <https://github.com/bcaffo>
- Jeff Leek
  - Website <http://biostat.jhsph.edu/~jleek/>, <http://simplystatistics.org/>
  - Twitter [@jtleek](#)
  - Github <https://github.com/jtleek>
- Roger Peng
  - Website <http://www.biostat.jhsph.edu/~rpeng/>, <http://simplystatistics.org/>
  - Twitter [@rdpeng](#)
  - Github <https://github.com/rdpeng>

# Why data science?



<http://www.economist.com/node/15579717>

# Why data science?

McKinsey Global Institute



June 2011

Big data: The next frontier  
for innovation, competition,  
and productivity

[http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

# Why statistical data science?

## The New York Times

### For Today's Graduate, Just One Word: Statistics


By **STEVE LOHR**

Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

 TWITTER

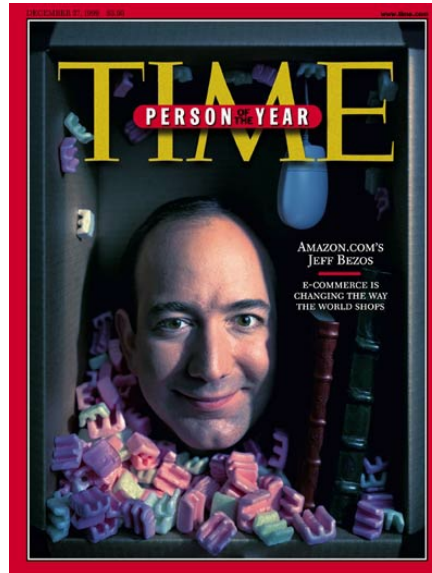
 LINKEDIN

 COMMENTS  
(58)

 SIGN IN TO E-MAIL

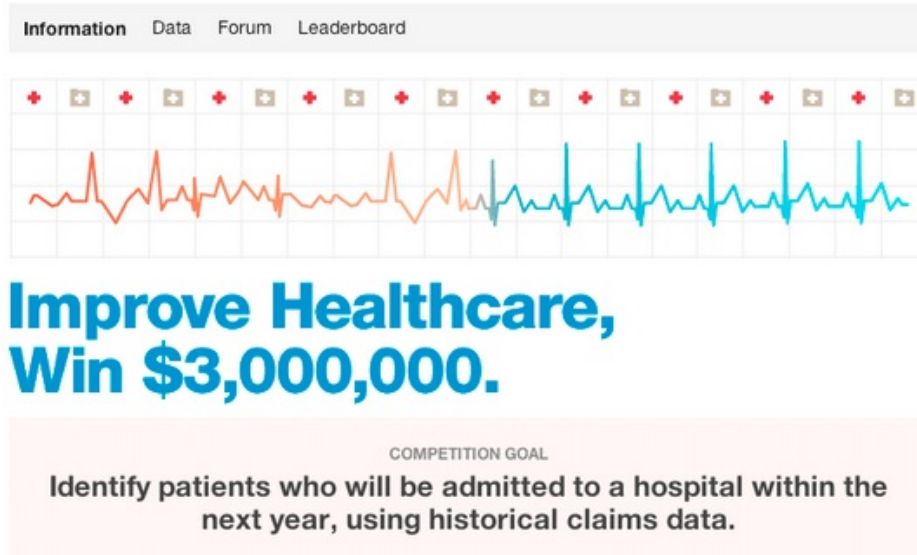
[http://www.nytimes.com/2009/08/06/technology/06stats.html?\\_r=0](http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=0)

# Why are you lucky?





# Why are you lucky?



[Heritage Health Prize](#)

# Why R?

The New York Times

## Business Computing

Search All NYTimes.com

Go



WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUTOS



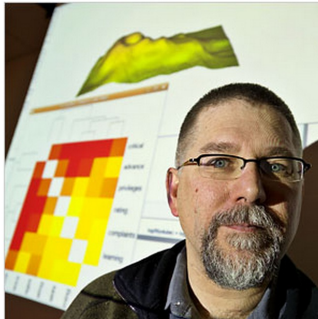
PROGRESS IS EVERYONE'S BUSINESS

See how Goldman Sachs has helped Hologic enable better outcomes for patients.

WATCH THE VIDEO

Goldman Sachs

### Data Analysts Captivated by R's Power



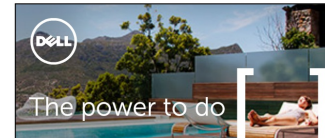
Log in to see what your friends are sharing on nytimes.com. [Privacy Policy](#) | [What's This?](#) Log In With Facebook

#### What's Popular Now

Amiri Baraka, Polarizing Poet and Playwright, Dies at 79



'Very Sad' Chris Christie Extends Apology in Bridge Scandal



<http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all>

# Why R?

- It is free
- It has a comprehensive set of packages
  - Data access
  - Data cleaning
  - Analysis
  - Data reporting
- It has one of the best development environments - Rstudio <http://www.rstudio.com/>
- It has an amazing ecosystem of developers
- Packages are easy to install and "play nicely together"

# Who is a data scientist?



[Daryl Morey](#)

# Who is a data scientist?



[Hilary Mason](#)

# Who is a data scientist?



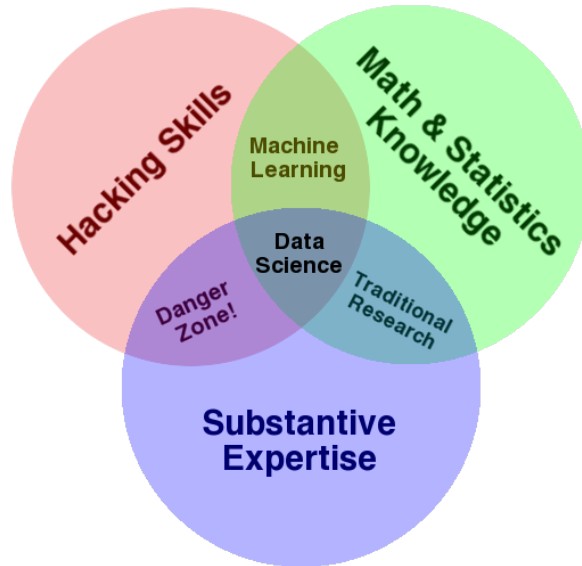
[Daphne Koller](#)

# Who is a data scientist?



[Nate Silver](#)

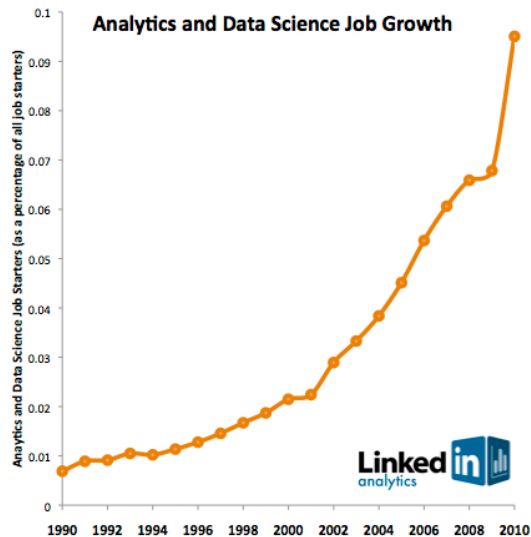
# Our goal



[Drew Conway](#)



# Plus jobs



<http://radar.oreilly.com/2011/09/building-data-science-teams.html>

# This course

- Introducing you to the track
- Getting tools set up
- Giving you basic background