IS 6489: Statistics and Predictive Analytics

Class 1

Jeff Webb

Tonight's agenda

- What is this course about?

Tonight's agenda

- What is this course about?
- Course elements.

Tonight's agenda

- What is this course about?
- Course elements.
- Why study data science?

Tonight's agenda

- What is this course about?
- Course elements.
- Why study data science?
- Class 1 script: introduction to R and RStudio

What is this course about?

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis
- Statistical inference

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis
- Statistical inference
- Linear regression (including model assumptions and diagnostics)

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis
- Statistical inference
- Linear regression (including model assumptions and diagnostics)
- Model selection and regularization

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis
- Statistical inference
- Linear regression (including model assumptions and diagnostics)
- Model selection and regularization
- Using models for prediction (and related issues such as overfitting and cross-validation)

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis
- Statistical inference
- Linear regression (including model assumptions and diagnostics)
- Model selection and regularization
- Using models for prediction (and related issues such as overfitting and cross-validation)
- Logistic regression

Course topics

This is a graduate level course in applied statistics using R, with an emphasis on linear and logistic regression models. For comparison we will also briefly discuss some machine learning approaches to regression and classification tasks. The engaged student should expect to develop foundational skills for data analysis. Core statistical topics covered will include:

- Exploratory data analysis
- Statistical inference
- Linear regression (including model assumptions and diagnostics)
- Model selection and regularization
- Using models for prediction (and related issues such as overfitting and cross-validation)
- Logistic regression
- Statistical communication

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;
- Pick statistical methods that are appropriate for the data and the research question;

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;
- Pick statistical methods that are appropriate for the data and the research question;
- Develop and compare multiple models, checking for violations of model assumptions and assessing model fit;

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;
- Pick statistical methods that are appropriate for the data and the research question;
- Develop and compare multiple models, checking for violations of model assumptions and assessing model fit;
- Choose the best model for the analytic context;

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;
- Pick statistical methods that are appropriate for the data and the research question;
- Develop and compare multiple models, checking for violations of model assumptions and assessing model fit;
- Choose the best model for the analytic context;
- Interpret and translate results for non-expert audiences;

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;
- Pick statistical methods that are appropriate for the data and the research question;
- Develop and compare multiple models, checking for violations of model assumptions and assessing model fit;
- Choose the best model for the analytic context;
- Interpret and translate results for non-expert audiences;
- Make your research reproducible.

Course learning objectives

This course will help you develop the skills necessary to be a working data analyst or data scientist. You will learn how to:

- Explore, summarize, and visualize data using appropriate descriptive techniques;
- Pick statistical methods that are appropriate for the data and the research question;
- Develop and compare multiple models, checking for violations of model assumptions and assessing model fit;
- Choose the best model for the analytic context;
- Interpret and translate results for non-expert audiences;
- Make your research reproducible.
- Above all, I hope you will learn how to **think with data** by asking questions to guide your analysis and then, having completed that analysis, being able to understand and communicate the business value of your results.

Regression example

- This course will be focused on regression as a staple technique for data analysis.

Regression example

- This course will be focused on regression as a staple technique for data analysis.
- With regression we can (among other things):

Regression example

- This course will be focused on regression as a staple technique for data analysis.
- With regression we can (among other things):
  - **describe** the relationships between variables in a data sample (and assess whether those relationships are artifacts of the sample).

Regression example

- This course will be focused on regression as a staple technique for data analysis.
- With regression we can (among other things):
  - **describe** the relationships between variables in a data sample (and assess whether those relationships are artifacts of the sample).
  - create a model to **predict** unknown values of the outcome variable given known inputs.

Regression example

- This course will be focused on regression as a staple technique for data analysis.
- With regression we can (among other things):
  - **describe** the relationships between variables in a data sample (and assess whether those relationships are artifacts of the sample).
  - create a model to **predict** unknown values of the outcome variable given known inputs.
- Regression models are easy to fit and extremely powerful, *yet they are also easy to misuse and misinterpret*

mtcars dataset (first six rows)

	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6	160	110	3.90	2.620
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875
Datsun 710	22.8	4	108	93	3.85	2.320
Hornet 4 Drive	21.4	6	258	110	3.08	3.215
Hornet Sportabout	18.7	8	360	175	3.15	3.440
Valiant	18.1	6	225	105	2.76	3.460

Let's create a simple model of mpg

```
## lm(formula = mpg ~ cyl + hp + wt, data = mtcars)
##             coef.est coef.se
## (Intercept) 38.75     1.79
## cyl         -0.94     0.55
## hp          -0.02     0.01
## wt          -3.17     0.74
## ---
## n = 32, k = 4
## residual sd = 2.51, R-Squared = 0.84
```

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?
- Is this a good model? If we wanted to make it better, which variables should be added or removed?

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?
- Is this a good model? If we wanted to make it better, which variables should be added or removed?
- How would we know if adding or removing variables improved the model?

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?
- Is this a good model? If we wanted to make it better, which variables should be added or removed?
- How would we know if adding or removing variables improved the model?
- Should any of these variables be transformed or should any outlying observations be removed from the dataset?

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?
- Is this a good model? If we wanted to make it better, which variables should be added or removed?
- How would we know if adding or removing variables improved the model?
- Should any of these variables be transformed or should any outlying observations be removed from the dataset?
- Does this model violate any of the mathematical assumptions of linear regression?

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?
- Is this a good model? If we wanted to make it better, which variables should be added or removed?
- How would we know if adding or removing variables improved the model?
- Should any of these variables be transformed or should any outlying observations be removed from the dataset?
- Does this model violate any of the mathematical assumptions of linear regression?
- **Using modern statistical software to fit models is easy, but understanding, validating, improving and communicating your results can be a challenge.**

Questions

- What do these coefficients mean exactly? How would we translate them into meaningful quantities for a client with no background in statistics?
- Is this a good model? If we wanted to make it better, which variables should be added or removed?
- How would we know if adding or removing variables improved the model?
- Should any of these variables be transformed or should any outlying observations be removed from the dataset?
- Does this model violate any of the mathematical assumptions of linear regression?
- **Using modern statistical software to fit models is easy, but understanding, validating, improving and communicating your results can be a challenge.**
- This course will equip you for that challenge.

Course elements

Main course texts

- **Datacamp.** Students have free access to all of the content at Datacamp through the end of the semester. (Email me if you have not received an invitation to the IS 6489 group at Datacamp or experience problems with your account.)

Main course texts

- **Datacamp.** Students have free access to all of the content at Datacamp through the end of the semester. (Email me if you have not received an invitation to the IS 6489 group at Datacamp or experience problems with your account.)
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). **An introduction to statistical learning**. Springer. This is the main textbook for the course. It is available to download for free at the above link (look in the upper right corner of the page: "Download the book PDF"). The print book is available from Amazon.

Supplementary course texts

- Gelman, A., and Hill, J. (2007). **Data analysis using regression/hierarchical models**. Cambridge: Cambridge UP. Several chapters from this book will be posted on Canvas as a supplementary resource.

Supplementary course texts

- Gelman, A., and Hill, J. (2007). **Data analysis using regression/hierarchical models**. Cambridge: Cambridge UP. Several chapters from this book will be posted on Canvas as a supplementary resource.
- Webb, J. (2017). **Course Notes for IS-6489, Statistics and Predictive**. The notes cover the course material in a lot of detail, with many specific code examples.

Course schedule

- Thursdays, 6 - 10 PM, during semester terms II and III.

Course schedule

- Thursdays, 6 - 10 PM, during semester terms II and III.
- However, this will be a hybrid course, with some lecture material available online, to be watched before class. Our nightly schedule will usually go from 6 - 8:30 PM.

Homework

- **Video lectures.** You should plan to watch the weekly videos before class. The lectures include embedded comprehension quizzes.

Homework

- **Video lectures.** You should plan to watch the weekly videos before class. The lectures include embedded comprehension quizzes.
- **Readings.** Weekly readings from *An introduction to statistical learning* should be completed before class.

Homework

- **Video lectures.** You should plan to watch the weekly videos before class. The lectures include embedded comprehension quizzes.
- **Readings.** Weekly readings from *An introduction to statistical learning* should be completed before class.
- **Labs.** There will be weekly labs consisting in questions embedded in interactive R notebooks.

Homework

- **Video lectures.** You should plan to watch the weekly videos before class. The lectures include embedded comprehension quizzes.
- **Readings.** Weekly readings from *An introduction to statistical learning* should be completed before class.
- **Labs.** There will be weekly labs consisting in questions embedded in interactive R notebooks.
- **Weekly quizzes.** To ensure that you have understood the material in the labs, there will be short weekly quizzes covering the same material.

Project

The final project will consist in a prediction competition that will require you to practice the skills you learned in the class.

- You can choose to work in a group no larger than three or, if you prefer, by yourself.

Project

The final project will consist in a prediction competition that will require you to practice the skills you learned in the class.

- You can choose to work in a group no larger than three or, if you prefer, by yourself.
- There will be an interim report due midway through the semester to ensure that you're making progress on the project, and a final report due a week after the last class.

Methods of instruction

- Class sessions will be a mix of review lecture and practice. I will rehearse concepts and tools from the lecture videos and then, depending on the topic, we will practice live coding or work through and discuss data analysis problems and exercises.

Methods of instruction

- Class sessions will be a mix of review lecture and practice. I will rehearse concepts and tools from the lecture videos and then, depending on the topic, we will practice live coding or work through and discuss data analysis problems and exercises.
- I like to teach interactively, so please do not hesitate to ask questions during the class. Chances are, if you have a question about the material, someone else does too: you will be doing everyone a favor by asking your question.

Methods of instruction

- Class sessions will be a mix of review lecture and practice. I will rehearse concepts and tools from the lecture videos and then, depending on the topic, we will practice live coding or work through and discuss data analysis problems and exercises.
- I like to teach interactively, so please do not hesitate to ask questions during the class. Chances are, if you have a question about the material, someone else does too: you will be doing everyone a favor by asking your question.
- You should expect to work in small groups occasionally and to present your findings to the class. I will also occasionally cold call on students for answers to problems.

Methods of instruction

- Class sessions will be a mix of review lecture and practice. I will rehearse concepts and tools from the lecture videos and then, depending on the topic, we will practice live coding or work through and discuss data analysis problems and exercises.
- I like to teach interactively, so please do not hesitate to ask questions during the class. Chances are, if you have a question about the material, someone else does too: you will be doing everyone a favor by asking your question.
- You should expect to work in small groups occasionally and to present your findings to the class. I will also occasionally cold call on students for answers to problems.
- Slides will be available prior to class for download, if you want to follow along on your own computer.

Methods of instruction

- Class sessions will be a mix of review lecture and practice. I will rehearse concepts and tools from the lecture videos and then, depending on the topic, we will practice live coding or work through and discuss data analysis problems and exercises.
- I like to teach interactively, so please do not hesitate to ask questions during the class. Chances are, if you have a question about the material, someone else does too: you will be doing everyone a favor by asking your question.
- You should expect to work in small groups occasionally and to present your findings to the class. I will also occasionally cold call on students for answers to problems.
- Slides will be available prior to class for download, if you want to follow along on your own computer.
- The script created during live coding will be posted to Canvas afterwards for your reference.

Learning R

- If you have never done any programming, learning R can be a challenge.

Learning R

- If you have never done any programming, learning R can be a challenge.
- Be tenacious in puzzling through the examples in class and use the tutorials at Datacamp. You will pick it up quickly.

Learning R

- If you have never done any programming, learning R can be a challenge.
- Be tenacious in puzzling through the examples in class and use the tutorials at Datacamp. You will pick it up quickly.
- Make use of office hours and the TA.

Data Science

Why data science? Free and ubiquitous data.

"The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate— will be a hugely important skill in the next decades, because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it. **I keep saying the sexy job in the next ten years will be statisticians**."

Hal Varian, Google Chief Economist and UC Berkeley Professor, *The McKinsey Quarterly*, January 2009

Why data science? A trend of more and more data.

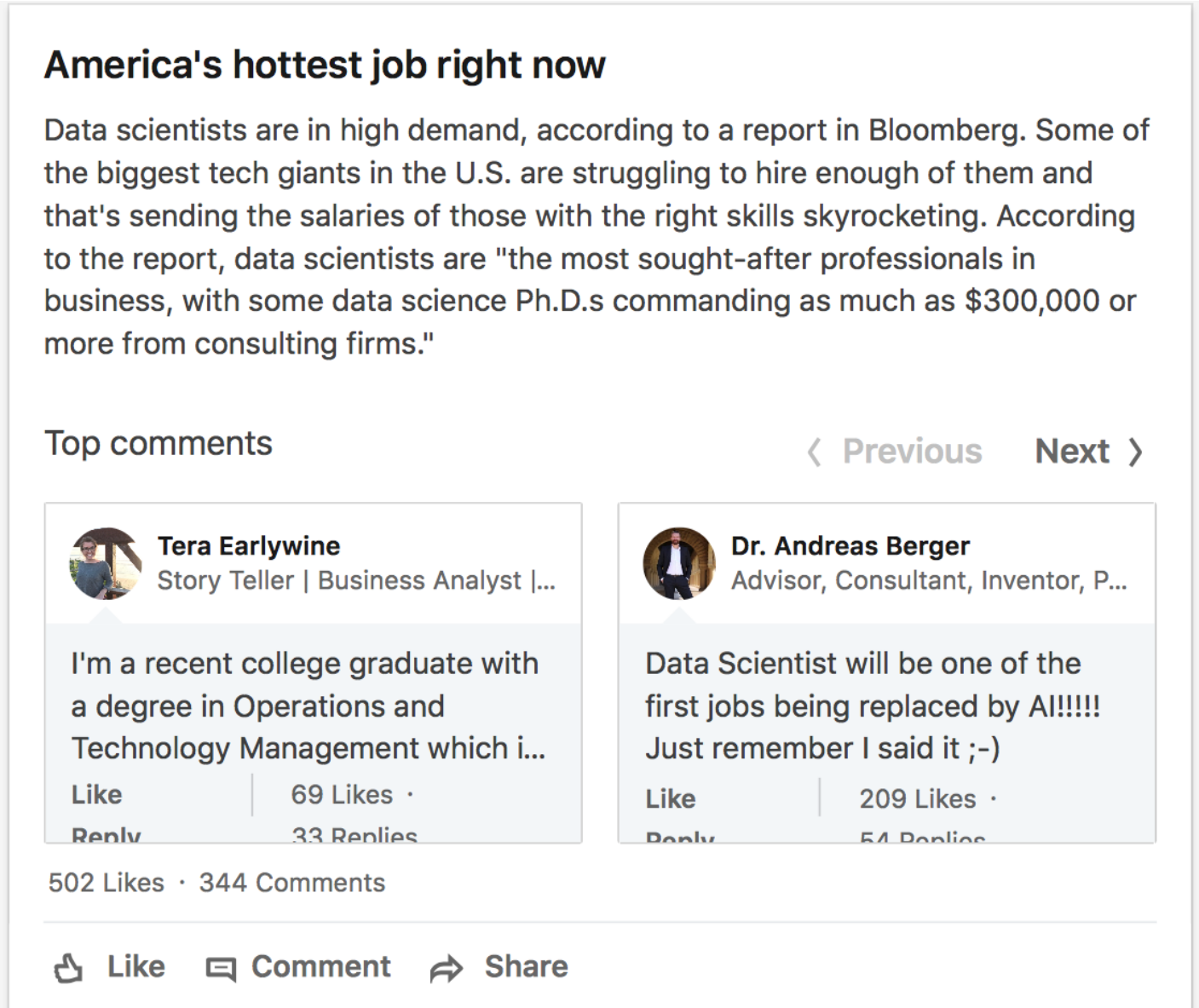
From 2011:

"By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

Mckinsey & Company, *Big data: The next frontier for innovation, competition, and productivity* (2011).

Why data science? The trend continues.

From LinkedIn, May 2018:



America's hottest job right now

Data scientists are in high demand, according to a report in Bloomberg. Some of the biggest tech giants in the U.S. are struggling to hire enough of them and that's sending the salaries of those with the right skills skyrocketing. According to the report, data scientists are "the most sought-after professionals in business, with some data science Ph.D.s commanding as much as $300,000 or more from consulting firms."

Top comments ‹ Previous Next ›

Tera Earlywine
Story Teller | Business Analyst |...

I'm a recent college graduate with a degree in Operations and Technology Management which i...

Like 69 Likes ·
Reply 33 Replies

Dr. Andreas Berger
Advisor, Consultant, Inventor, P...

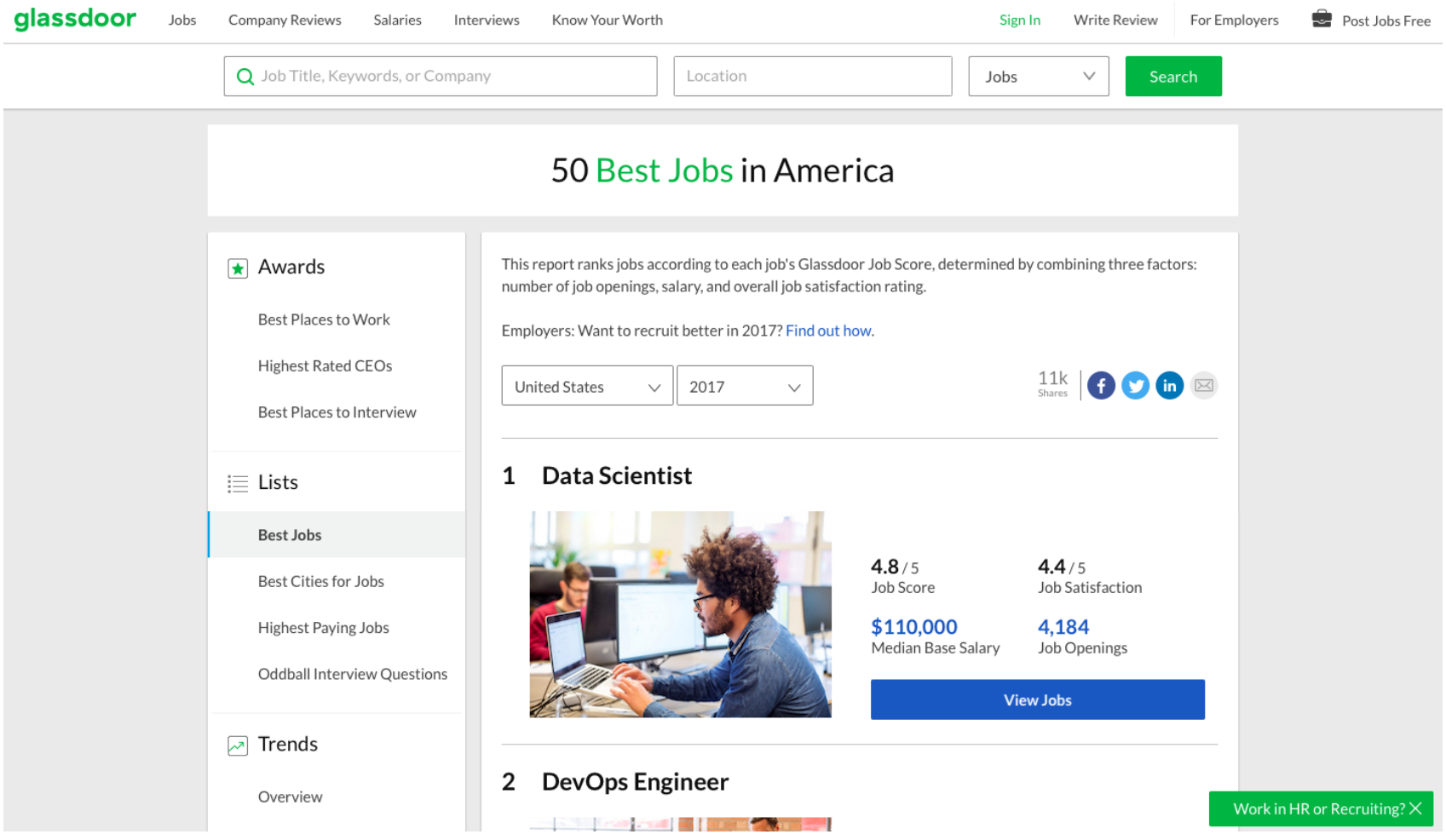
Data Scientist will be one of the first jobs being replaced by AI!!!!! Just remember I said it ;-)

Like 209 Likes ·
Reply 54 Replies

502 Likes · 344 Comments

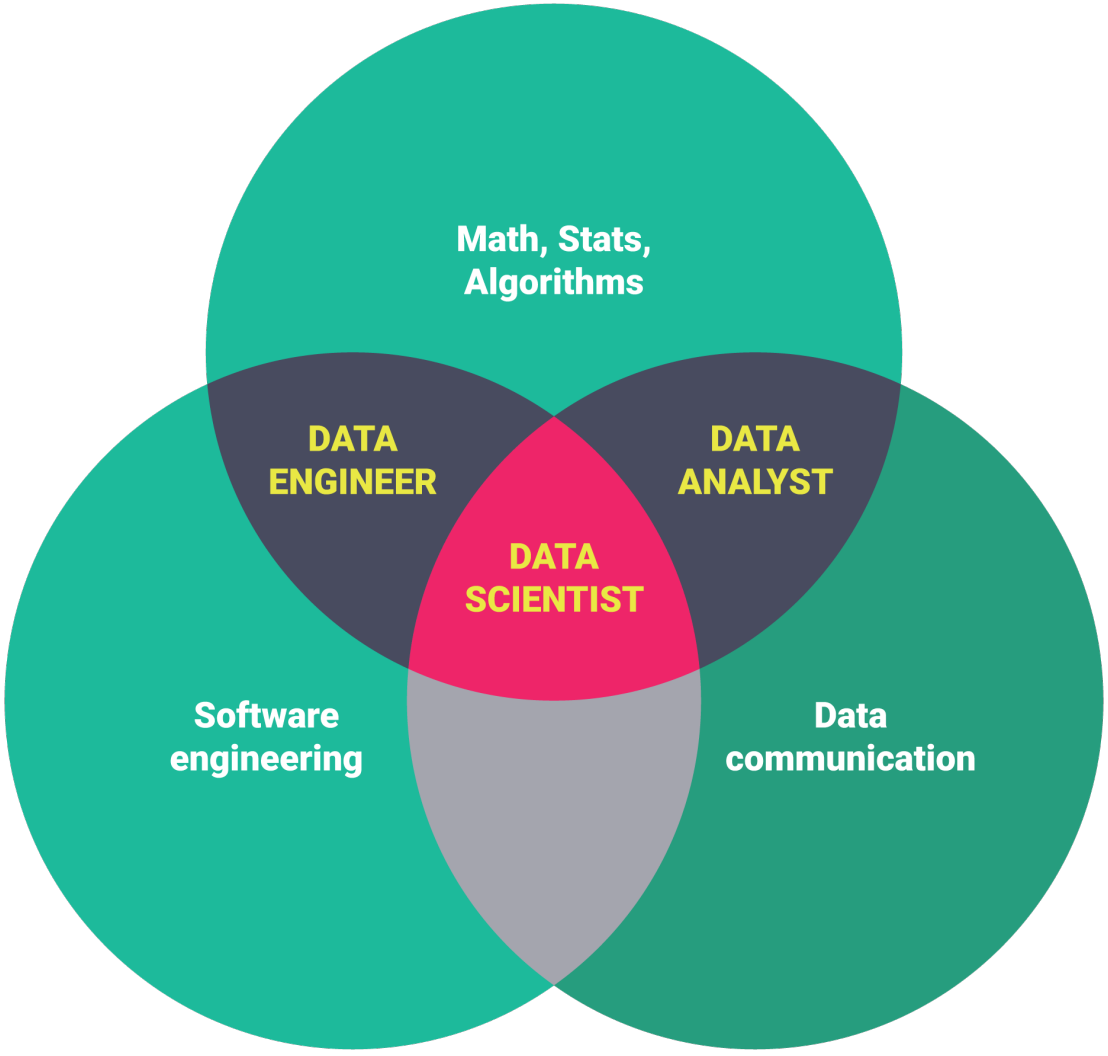
Like Comment Share

Why data science? Interesting work.



What is a data scientist?

Data scientists extract, visualize and communicate insights from data. They are skilled at statistics, programming and telling stories.



Why R?

- R is an open-source, object-oriented programming language that was invented to do statistics and is widely used.

Why R?

- R is an open-source, object-oriented programming language that was invented to do statistics and is widely used.
- Solutions to coding problems abound on the web.

Why R?

- R is an open-source, object-oriented programming language that was invented to do statistics and is widely used.
- Solutions to coding problems abound on the web.
- Cutting edge techniques are immediately available as packages (long before they are incorporated into commercial software).

Why R?

- R is an open-source, object-oriented programming language that was invented to do statistics and is widely used.
- Solutions to coding problems abound on the web.
- Cutting edge techniques are immediately available as packages (long before they are incorporated into commercial software).
- Makes collaboration and peer review easy (if your colleagues are using R).

Why R?

- R is an open-source, object-oriented programming language that was invented to do statistics and is widely used.
- Solutions to coding problems abound on the web.
- Cutting edge techniques are immediately available as packages (long before they are incorporated into commercial software).
- Makes collaboration and peer review easy (if your colleagues are using R).
- RStudio!

Why R?

- R is an open-source, object-oriented programming language that was invented to do statistics and is widely used.
- Solutions to coding problems abound on the web.
- Cutting edge techniques are immediately available as packages (long before they are incorporated into commercial software).
- Makes collaboration and peer review easy (if your colleagues are using R).
- RStudio!
- The tidyverse collection of packages: dplyr, ggplot2, tidyr ....

Homework

- Datacamp, *Introduction to R* and *Intermediate R* (through chapter 3). www.datacamp.com.

Homework

- Datacamp, *Introduction to R* and *Intermediate R* (through chapter 3). www.datacamp.com.
- Supplementary reading: *Course Notes*, chapters 1-3. https://bookdown.org/jefftemplewebb/IS-6489/

Homework

- Datacamp, *Introduction to R* and *Intermediate R* (through chapter 3). www.datacamp.com.
- Supplementary reading: *Course Notes*, chapters 1-3. https://bookdown.org/jefftemplewebb/IS-6489/
- Week 2 lecture videos.

Class 1 tutorial script: Introduction to RStudio and R

Find the script on Canvas: Files => Class 1