

Beyond the Buzzword

A look at data science in practice and
how you can be a part of it

Diversity in Data Science and Machine Learning

October 18, 2019



bit.ly/beyond-the-buzzword

Maria Tackett
Duke University



B.S. in Mathematics

✗ Minor in Computer Science

✗ Minor in Economics

✗ Minor in German

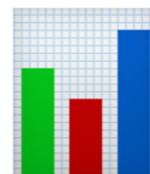


B.S. in Mathematics

✗ Minor in Computer Science

✗ Minor in Economics

✗ Minor in German



First statistics class junior year

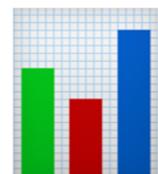


B.S. in Mathematics

✗ Minor in Computer Science

✗ Minor in Economics

✗ Minor in German



First statistics class junior year



M.S. in Statistics

Statistician @ Capital One



Design of Experiments

Regression Modeling



Ph.D. in Statistics

Dissertation Research

Used Bayesian methods to...

... decompose the sources of variability in two fingerprints

... develop a method to quantify certainty in fingerprint examiner's decision, adjusting for the number of fingerprints examined



**Assistant
Professor of @
the Practice**



Research

Statistics Education

Statistics in Criminal
Justice and Forensic
Science

Teaching

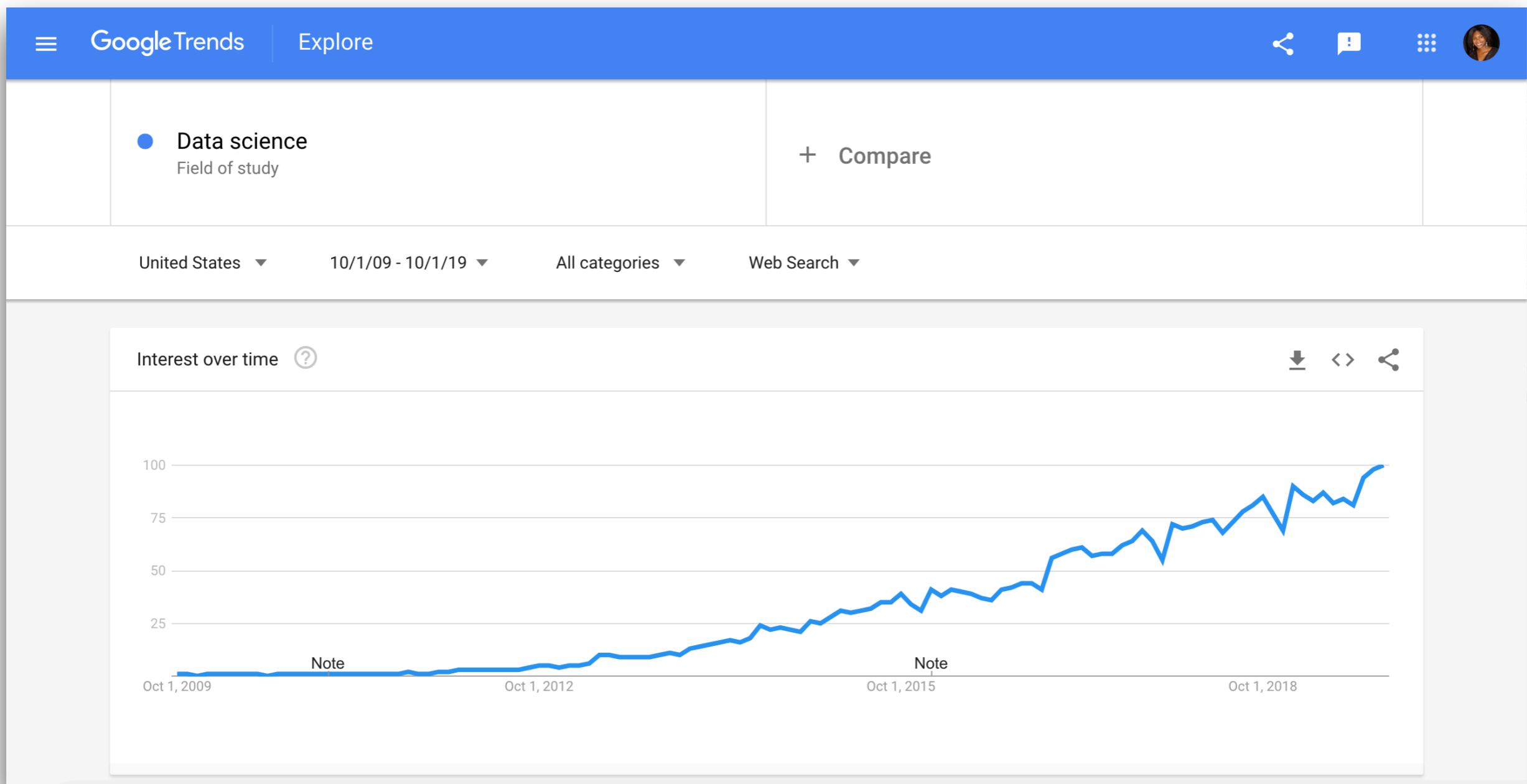
Intro to Data Science

Regression Analysis

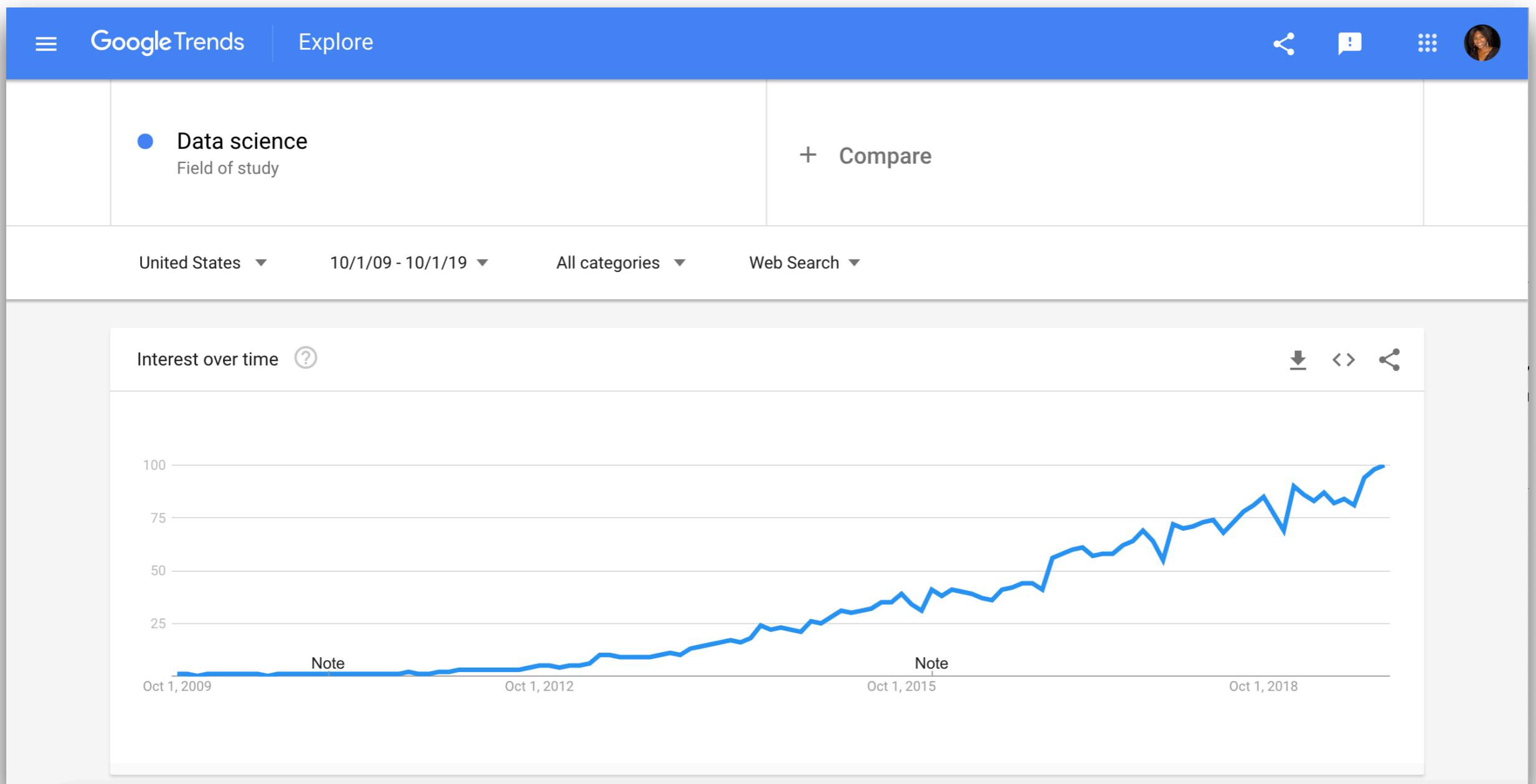


What exactly is
data science?

Data science has become increasingly popular over the past 10 years...

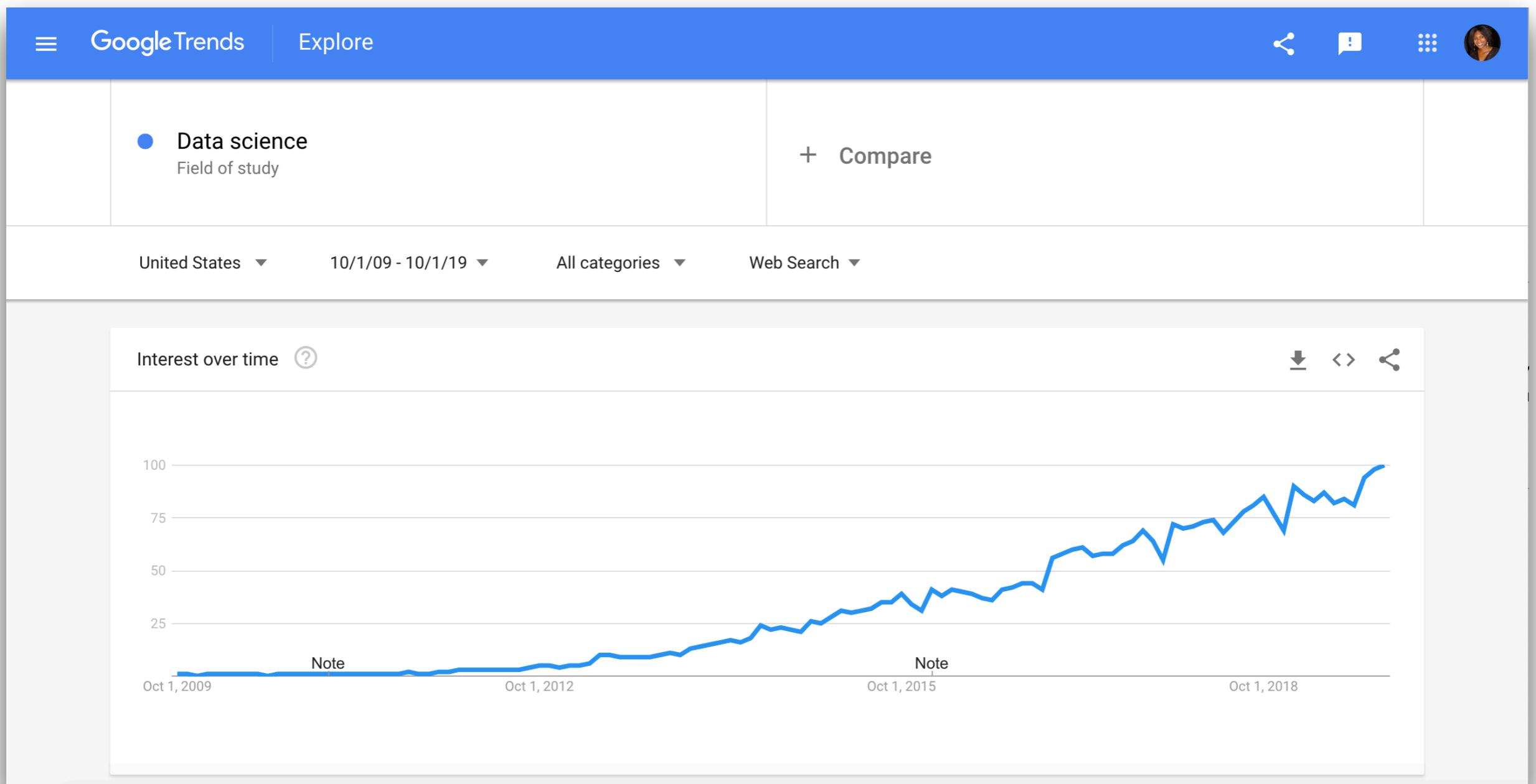


....but what exactly does “data science” mean?



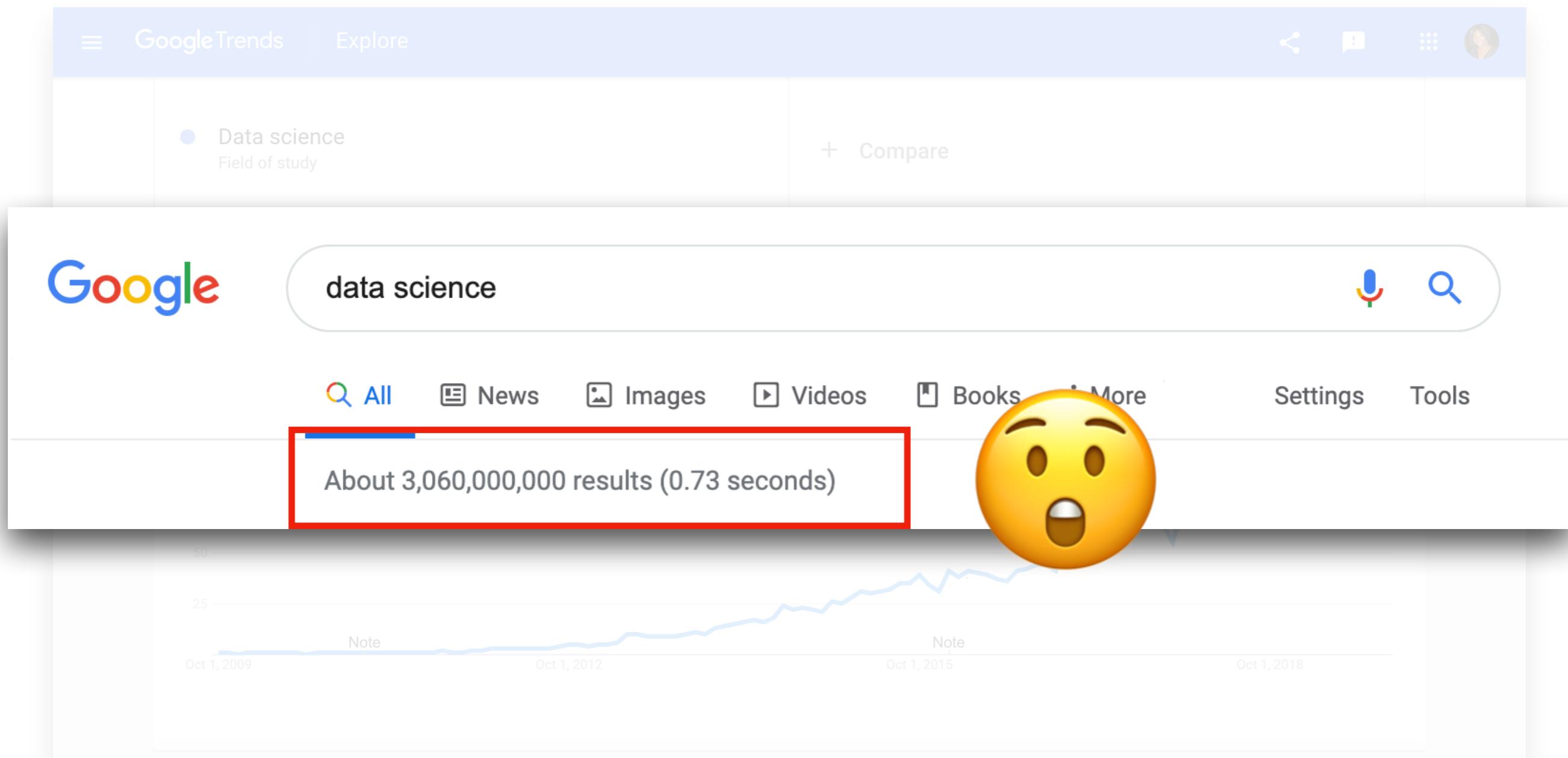
....but what exactly does “data science” mean?

Let's ask Google...



....but what exactly does “data science” mean?

Let's ask Google...



....but what exactly does “data science” mean?

Let's ask Google...

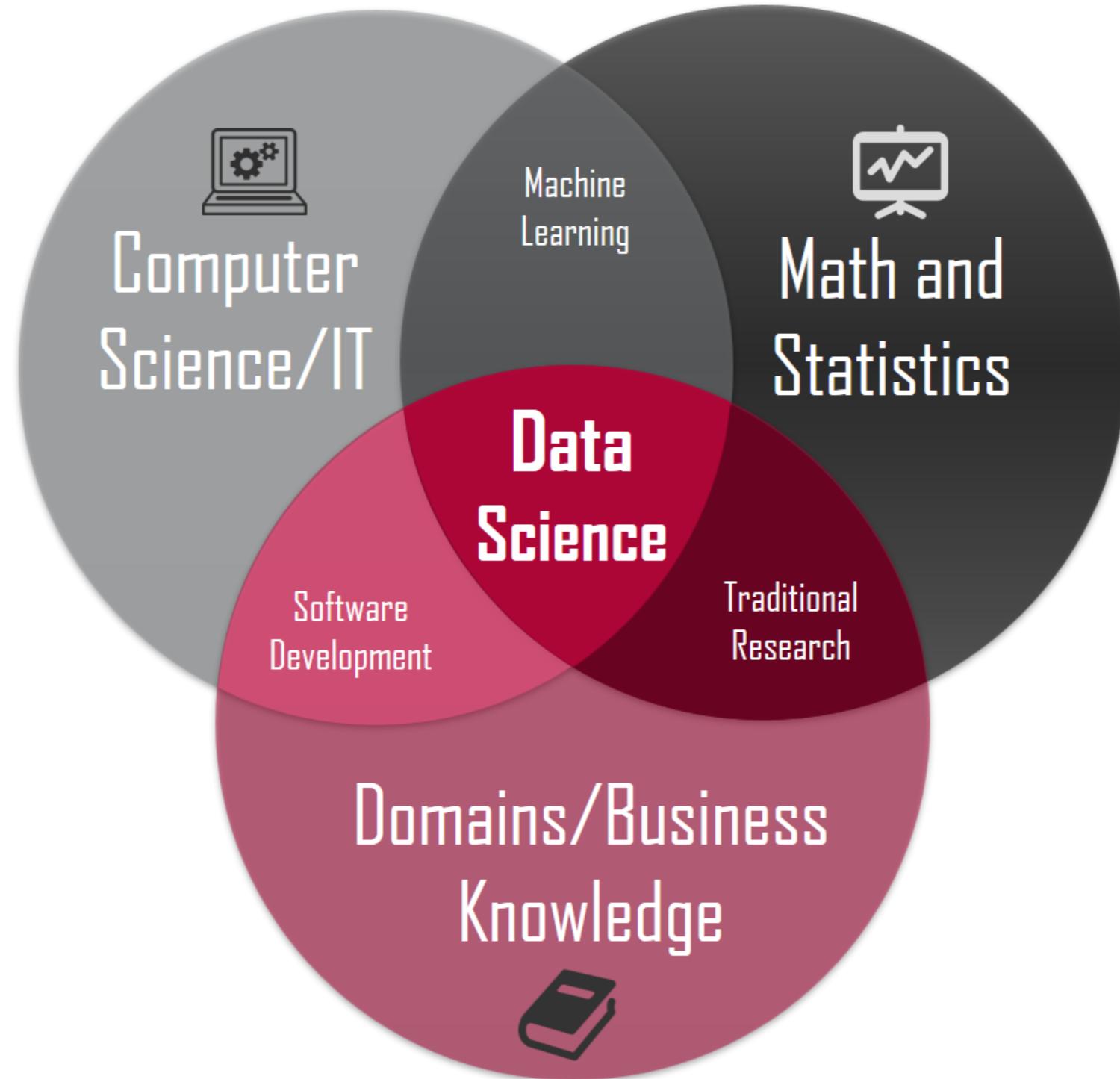
The screenshot shows a Google search result for the query "What is data science in simple words?". At the top, there's a navigation bar with "Google Trends" and "Explore" buttons, and icons for sharing, saving, and profile. Below the search bar, the "People also ask" section is visible, featuring a question "What is data science in simple words?" with a collapse arrow. A detailed answer follows, explaining that Data science is the study of data, involving recording, storing, and analyzing data to extract useful information. The goal is to gain insights and knowledge from structured and unstructured data. The text ends with a timestamp "Aug 17, 2017". Below this, a blue link leads to "Data Science Definition - The Tech Terms Computer Dictionary" with the URL "https://techterms.com › definition › data_science".

People also ask

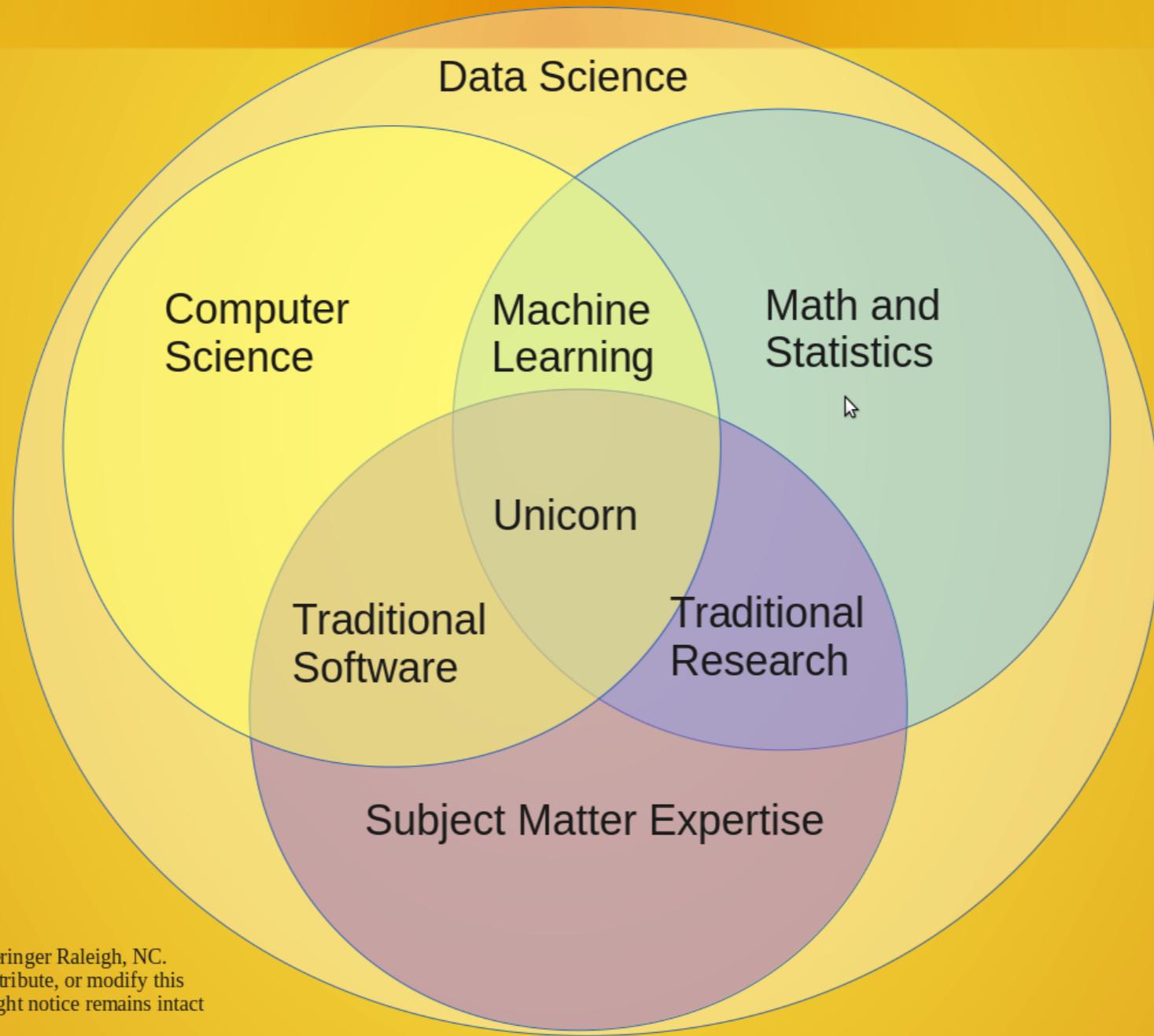
What is data science in simple words? ^

Data science is the study of **data**. It involves developing methods of recording, storing, and analyzing **data** to effectively extract useful information. The goal of **data science** is to gain insights and knowledge from any type of **data** – both structured and unstructured. Aug 17, 2017

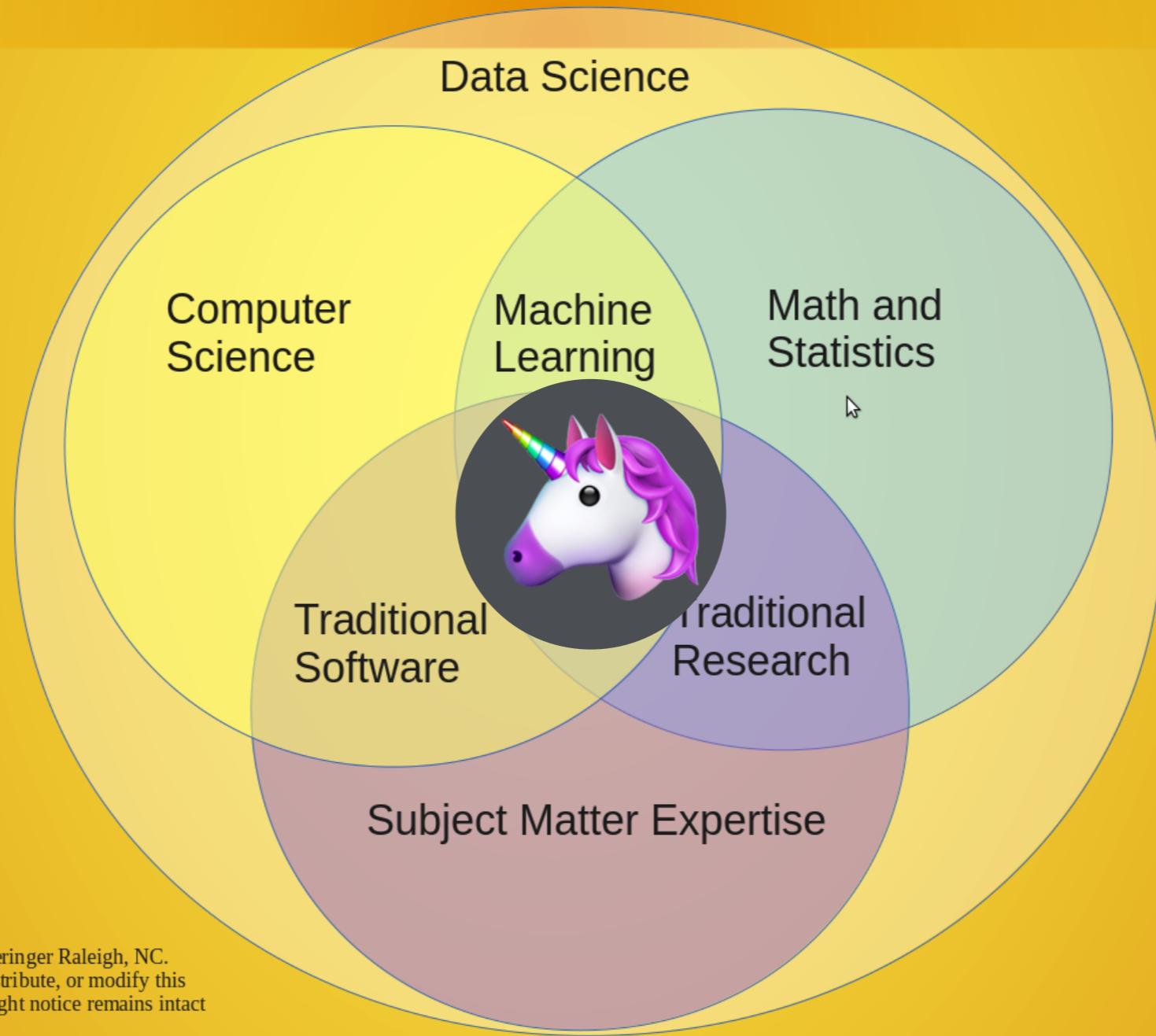
Data Science Definition - The Tech Terms Computer Dictionary
https://techterms.com/definition/data_science



Data Science Venn Diagram v2.0



Data Science Venn Diagram v2.0



What data scientists do...

- ✓ Identify relevant questions
- ✓ Collect data from a variety of data sources
- ✓ Organize information
- ✓ Translate results to solutions
- ✓ Communicate their findings

What data scientists have...

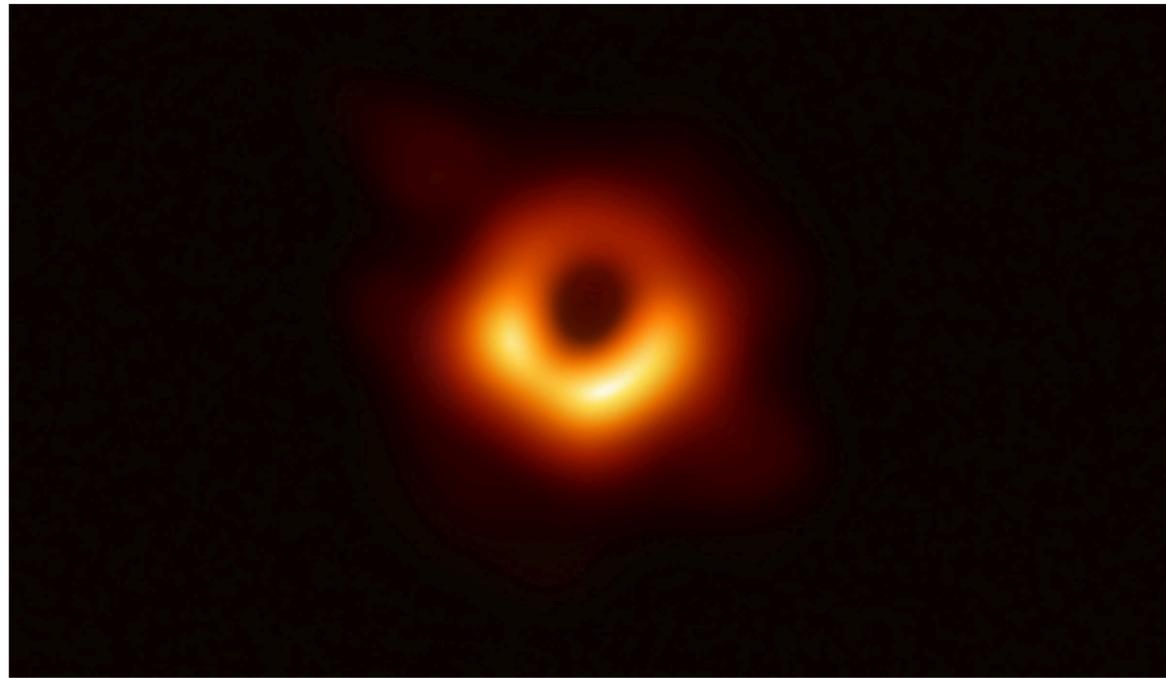
- ✓ Curiosity and results-oriented focus
- ✓ Industry-specific knowledge
- ✓ Ability to explain highly technical results to non-technical peers
- ✓ Background in statistics
- ✓ Background in computing

SOPHIA CHEN

SCIENCE 04.10.2019 09:45 AM

Scientists Reveal the First Picture of a Black Hole

The Event Horizon Telescope has captured a photo of a supermassive black hole at the center of M87, a galaxy 54 million light-years away.



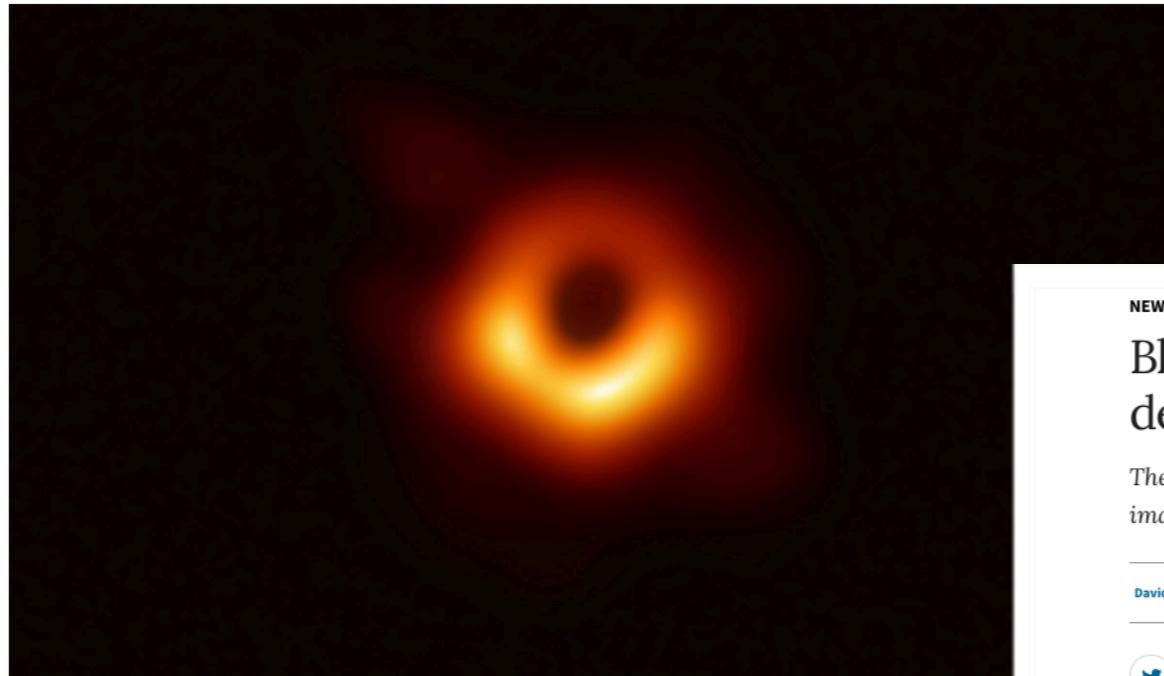
The Event Horizon Telescope has captured a photo of a supermassive black hole at the center of M87, a galaxy 54 million light years away. EVENT HORIZON TELESCOPE COLLABORATION ET AL.

SOPHIA CHEN

SCIENCE 04.10.2019 09:45 AM

Scientists Reveal the First Picture of a Black Hole

The Event Horizon Telescope has captured a photo of a supermassive black hole at the center of M87, a galaxy 54 million light-years away.



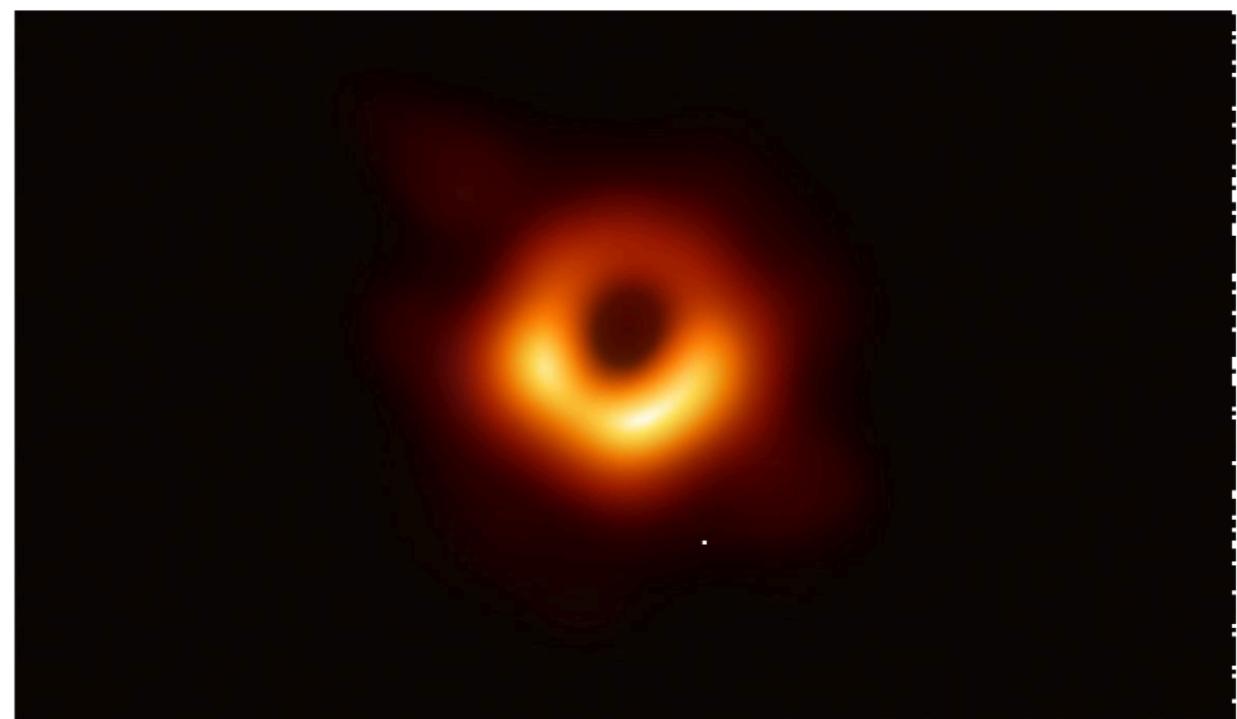
The Event Horizon Telescope has captured a photo of a supermassive black hole at the center of M87, a galaxy 54 million light years away. EVENT HORIZON COLLABORATION ET AL.

NEWS • 10 APRIL 2019

Black hole pictured for first time — in spectacular detail

The Event Horizon Telescope's global network of radio dishes has produced the first-ever direct image of a black hole and its event horizon.

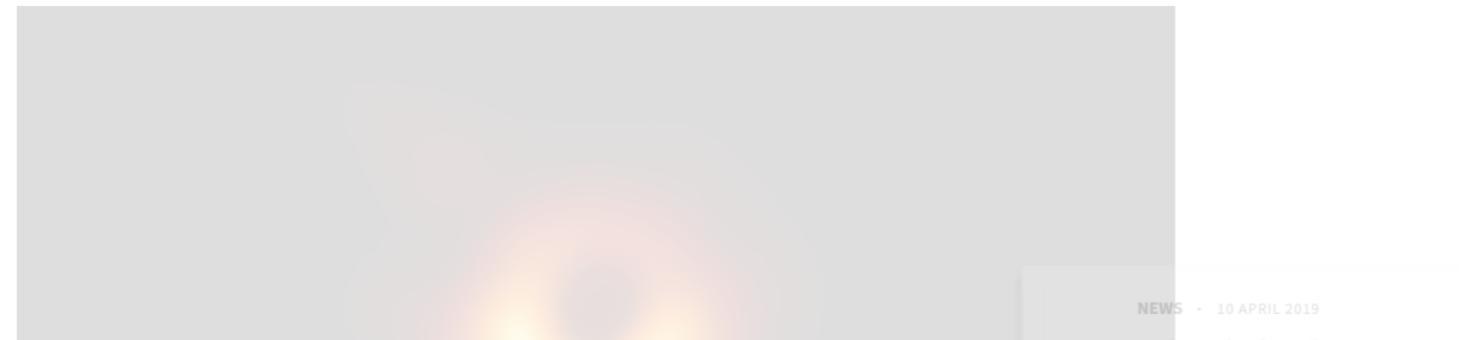
Davide Castelvecchi



The Event Horizon Telescope network has obtained the first image of a black hole — at the centre of the galaxy M87. Credit: EHT Collaboration

Scientists Reveal the First Picture of a Black Hole

The Event Horizon Telescope has captured a photo of a supermassive black hole at the center of M87, a galaxy 54 million light-years away.



NEWS • 10 APRIL 2019

Science

Algorithms gave us the black hole picture. She's the 29-year-old scientist who helped create them.



Katie Bouman was a MIT postdoctoral student when she led a team that designed one of the algorithms that led to capturing the first images of a black hole. (Adriana Usero/The Washington Post)

By Ben Guarino

April 10, 2019 at 3:49 p.m. EDT

d for first time — in spectacular

bal network of radio dishes has produced the first-ever direct horizon.



of a black hole — at the centre of the galaxy M87. Credit: EHT Collaboration

How data science made it possible...

We have telescopes distributed around the world. For every two telescopes in the telescope array, we get a single spatial frequency, which tells you something about like how fast things are changing.

We get this partial information. It's almost like seeing one pixel in an image (but it's in a different kind of domain). We have to come up with methods that take this really sparse, really noisy data, and try to find the image that might have caused those measurements.

• • •

What we have to end up doing is imposing things called “regularizers” or “priors” that allow us to say, “Okay, of all of the images that possibly could fit this data, this set of images is most likely.”

- Katie Bouman

How data science made it possible...

We have telescopes distributed around the world. For every two telescopes in the telescope array, we get a single spatial frequency, which tells you something about like how fast things are changing.

We get this ***partial information***. It's almost like seeing one pixel in an image (but it's in a different kind of domain). We have to come up with methods that take this really sparse, really noisy data, and try to find the image that might have caused those measurements.

• • •

What we have to end up doing is imposing things called “regularizers” or “priors” that allow us to say, “Okay, of all of the images that possibly could fit this data, this set of images is most likely.”

- Katie Bouman

How data science made it possible...

We have telescopes distributed around the world. For every two telescopes in the telescope array, we get a single spatial frequency, which tells you something about like how fast things are changing.

We get this partial information. It's almost like seeing one pixel in an image (but it's in a different kind of domain). We have to come up with ***methods that take this really sparse, really noisy data, and try to find the image that might have caused those measurements.***

• • •

What we have to end up doing is imposing things called “regularizers” or “priors” that allow us to say, “Okay, of all of the images that possibly could fit this data, this set of images is most likely.”

- Katie Bouman

How data science made it possible...

We have telescopes distributed around the world. For every two telescopes in the telescope array, we get a single spatial frequency, which tells you something about like how fast things are changing.

We get this partial information. It's almost like seeing one pixel in an image (but it's in a different kind of domain). We have to come up with methods that take this really sparse, really noisy data, and try to find the image that might have caused those measurements.

• • •

What we have to end up doing is ***imposing things called “regularizers” or “priors”*** that allow us to say, “Okay, of all of the images that possibly could fit this data, this set of images is most likely.”

- Katie Bouman

How data science made it possible...

We have telescopes distributed around the world. For every two telescopes in the telescope array, we get a single spatial frequency, which tells you something about like how fast things are changing.

We get this partial information. It's almost like seeing one pixel in an image (but it's in a different kind of domain). We have to come up with methods that take this really sparse, really noisy data, and try to find the image that might have caused those measurements.

• • •

What we have to end up doing is imposing things called “regularizers” or “priors” that allow us to say, ***“Okay, of all of the images that possibly could fit this data, this set of images is most likely.”***

- Katie Bouman



How can you be a
part of the data science
community?

Be a part of the data science community....

... even before your first internship or job

**Participate in data
science competitions**

Present your work

Share your work online

Participate in data science competitions



This is Statistics Fall Data Challenge
<https://www.causeweb.org/usproc/usresp>

Oct
28

Undergraduate Statistics Class Project Competition
<https://www.causeweb.org/usproc/usclap>

Dec
20

Undergraduate Statistics Research Project Competition
<https://www.causeweb.org/usproc/usresp>

Dec
20

ASA DataFest
<https://ww2.amstat.org/education/datafest/>

Spring
2020



Present your work

Present at student poster fairs

Present at meetings for data science groups (on campus and local)



Electronic Undergraduate Research Conference

Submit abstract by Oct 23

<https://www.causeweb.org/usproc/eusrc/2019>



Watch online
Nov 1

Share your work in an online repository



Maria Tackett
matackett

★ PRO

Edit profile

Assistant Professor of the Practice
Duke University
✉ maria.tackett@duke.edu
🌐 www.mariatackett.net

Organizations



Find a repository... Type: Public Language: All New

12 results for public repositories Clear filter

personal-website
HTML Updated yesterday

cv
Forked from isteves/resume
Resume/CV template using RMarkdown and LaTeX

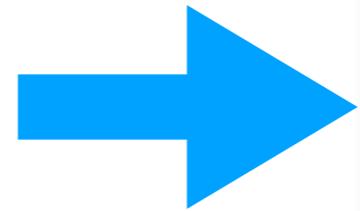
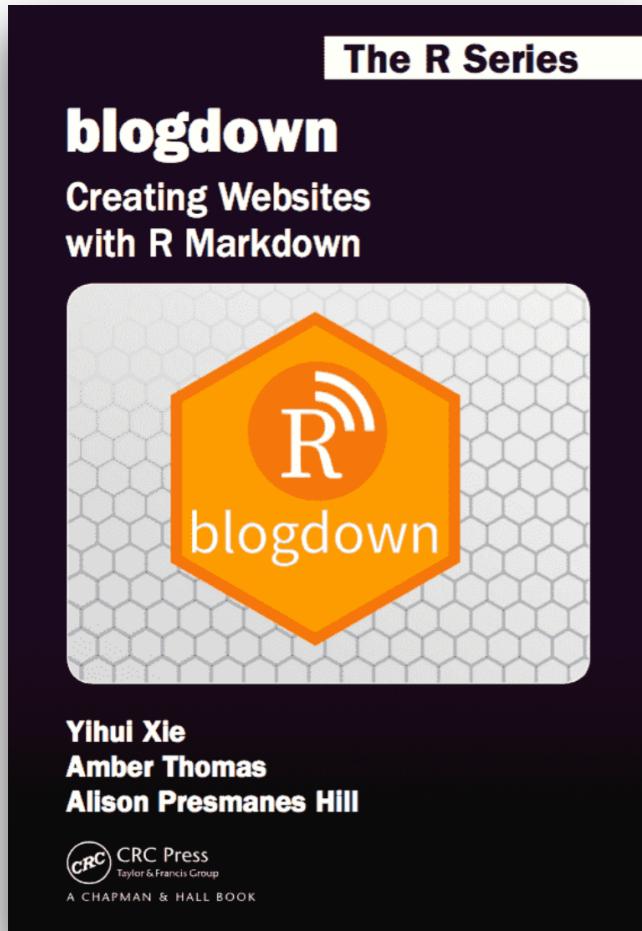
talks
TeX 16 Updated yesterday

intro-regression
Computing Assignments for an Intro Regression Course

sta199-sp19-website
Forked from Sta199-S18/website
Course website for Sta 199 - Intro to Data Science, Spring 18 at Duke University



Create a personal website

The image is a screenshot of a personal website for Maria Tackett. The browser window title is 'Maria Tackett' and the URL is 'www2.stat.duke.edu/~mt324/'. The page has a blue header with the name 'Maria Tackett' and navigation links for 'Home', 'Teaching', 'Talks', 'Media', and 'CV'. The main content area features a circular profile picture of Maria Tackett, a short bio, and sections for 'Interests' and 'Education'.

Maria Tackett

I am an Assistant Professor of the Practice in the Department of Statistical Science at Duke University. My current work involves applications in forensic science with a focus on latent fingerprint evidence. I am also interested in innovations in statistics education, specifically in using technology to enhance student learning.

Maria Tackett
Assistant Professor of the Practice
Duke University

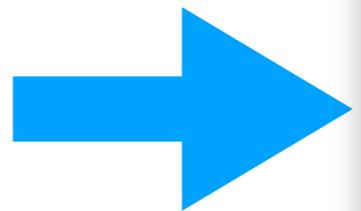
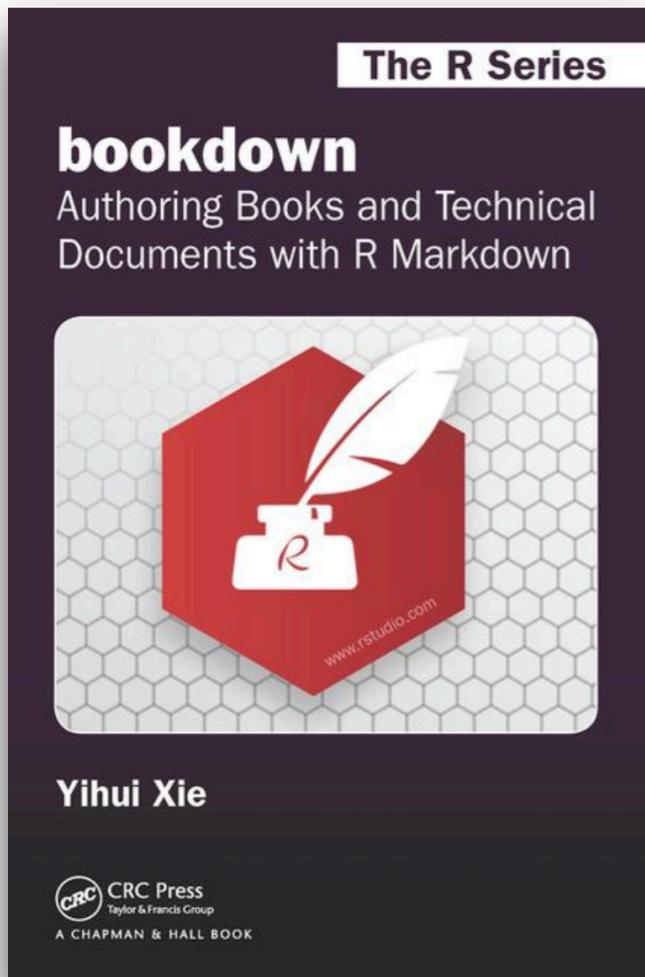
Interests

- Statistics Education
- Statistics in Forensic Science

Education

- Ph.D. in Statistics, 2018
University of Virginia
- M.S. in Statistics, 2010
University of Tennessee - Knoxville
- B.S. in Mathematics, 2009
University of Tennessee - Knoxville

Write about your work



The screenshot shows a web browser window titled 'Intro Regression' with the URL 'introregression.org'. The browser's toolbar includes icons for Apps, Gmail, Duke Email, Duke Sakai, fall-19, Piazza, R Studio, GitHub, Duke Box, previous-semesters, Resources, and Other Bookmarks. The main content area displays the 'Intro Regression' page. On the left is a sidebar with a navigation menu:

- Welcome to Intro Regression!
- 1 Getting Started
- 2 Notes for Instructors
- 3 Intro to R
- 4 Simple Linear Regression
- 5 Analysis of Variance
- 6 Multiple Linear Regression
- 7 Model Selection
- 8 Logistic Regression
- 9 Multinomial Logistic Regression
- 10 Special Topics
- 11 Data Sets
- References

Below the sidebar, a note states 'Published with bookdown'.

The main content area features a heading 'Intro Regression' and a sub-heading 'Welcome to Intro Regression!'. It is attributed to 'Maria Tackett' and updated on 'Latest update: 2019-07-10'. A decorative graphic of a scatter plot with a regression line is shown within a hexagonal frame. A note at the bottom explains the book's origin and provides contact information for updates.

The content in this book was originally developed for STA 210: Regression Analysis at Duke University. The computing aspects of the assignments are written using the `tidyverse` syntax in R; however, the assignments can be adapted to fit the computing language of your choice. All of the files are available in the [Intro Regression GitHub repo](#).

This book is under development and will be periodically updated with new material. Please email me (maria.tackett@duke.edu) if you have any questions, feedback, or suggestions. I would also love to hear about your experience if you use any of the content in your course.

Protect your major contributions

Branch: master [intro-regression / LICENSE.md](#)

matackett/intro-regression is licensed under the [Creative Commons Attribution Share Alike 4.0 International](#)

Similar to CC-BY-4.0 but requires derivatives be distributed under the same or a similar, compatible license. Frequently used for media assets and educational materials. A previous version is the default license for Wikipedia and other Wikimedia projects. Not recommended for software.

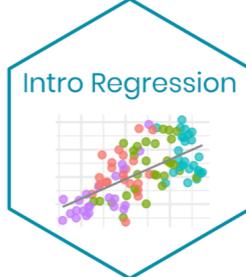
This is not legal advice. [Learn more about repository licenses.](#)

Permissions	Limitations	Conditions
✓ Commercial use ✓ Modification ✓ Distribution ✓ Private use	✗ Liability ✗ Trademark use ✗ Patent use ✗ Warranty	License and copyright notice State changes Same license

- Use a creative commons license to protect your work
- Make the license visible on the first page of your project or the top of your online repository

<https://creativecommons.org/licenses/>

Welcome to Intro Regression!



The content in this book was originally developed for STA 210: Regression Analysis at Duke University. The computing aspects of the assignments are written using the `tidyverse` syntax in R; however, the assignments can be adapted to fit the computing language of your choice. All of the files are available in the [Intro Regression GitHub repo](#).

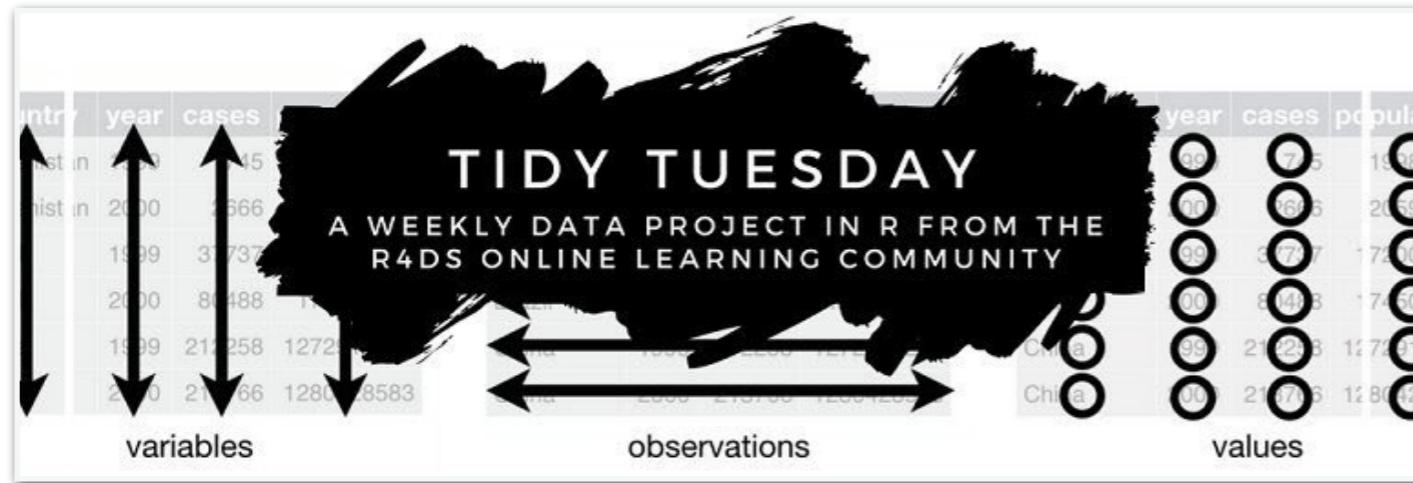
This book is under development and will be periodically updated with new material. Please email me (maria.tackett@duke.edu) if you have any questions, feedback, or suggestions. I would also love to hear about your experience if you use any of the content in your course.

License

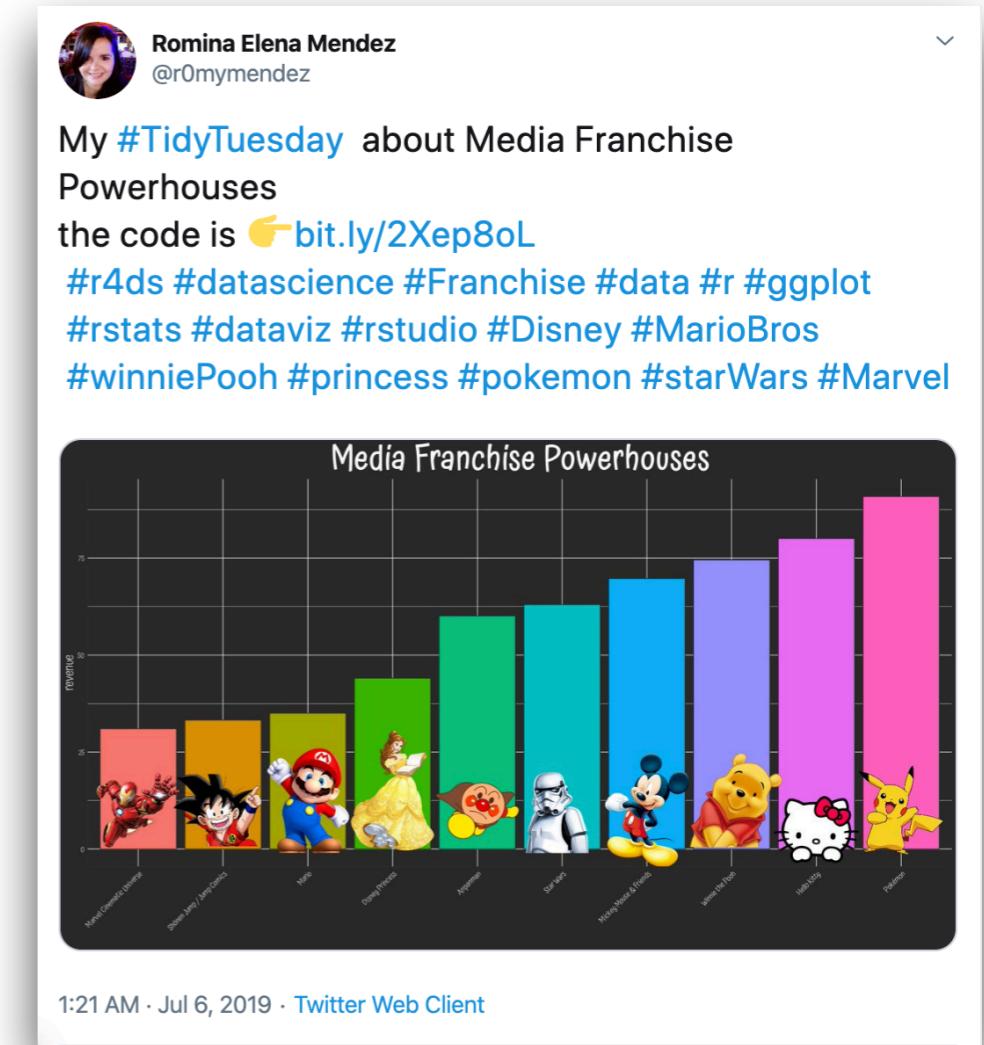
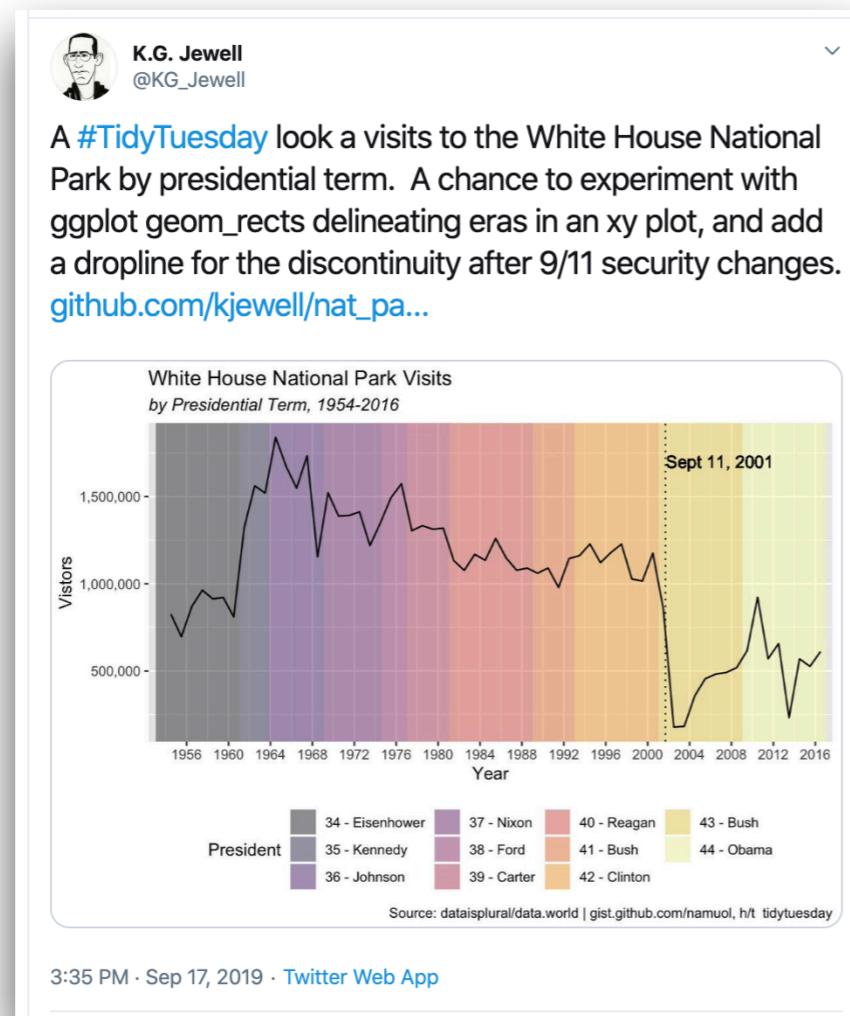


This work is licensed under the [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).

Try new things and share your experience



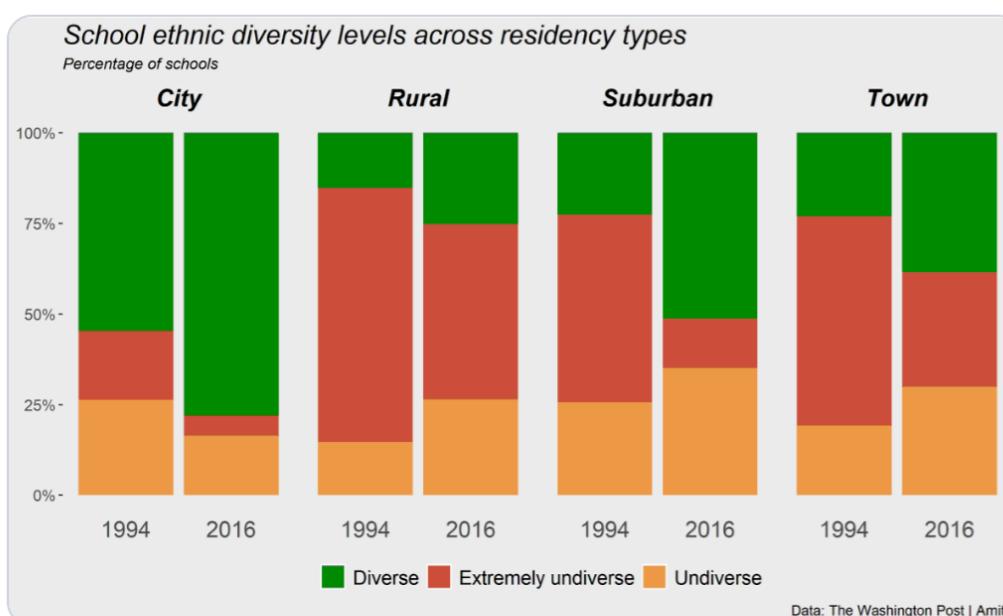
<https://github.com/rfordatascience/tidytuesday>





Amit Levinson
@Amit_Levinson

#TidyTuesday plot of school ethnic diversity levels across residency types for 1994 & 2016.
I love seeing how it fits with some sociological thoughts on minorities' residency: schools are more diverse in cities > suburb/town > rural.
code: bit.ly/2nuiv51



3:44 PM · Sep 26, 2019 · Twitter Web App



Amit Levinson @Amit_Levinson · Sep 26

Replying to @Amit_Levinson

I learned a lot this week!
Analysis: using stringr package instead of base r to change strings.
Visualization: used a stacked bar with 3 variables and various levels. I initially tried using facet of year ~ residency with regular bar plot, but i think stacked is neater.



1



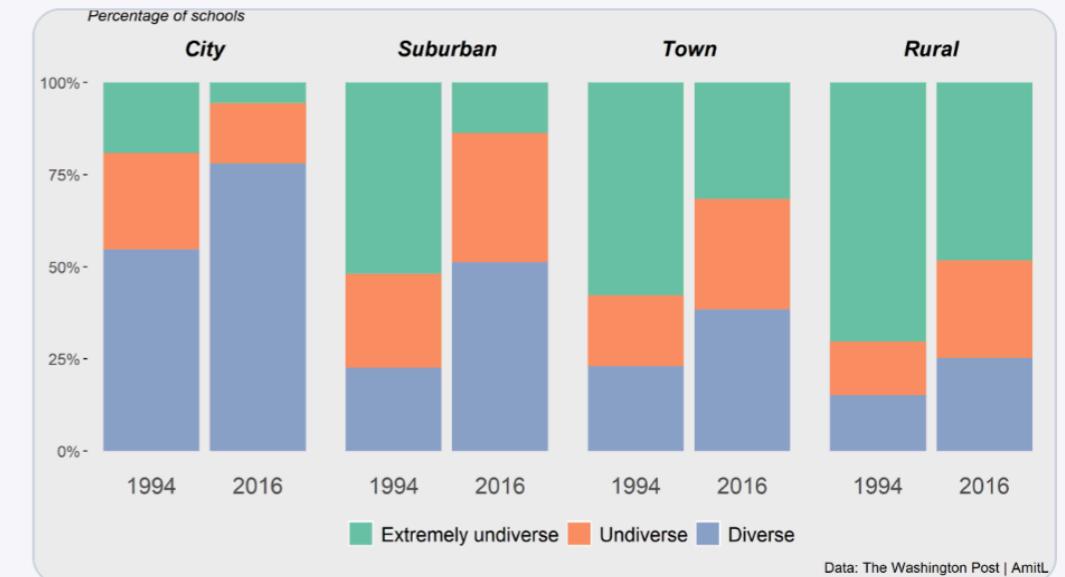
1



Amit Levinson @Amit_Levinson · Sep 28

Replying to @Amit_Levinson

some minor changes after some suggestions :)

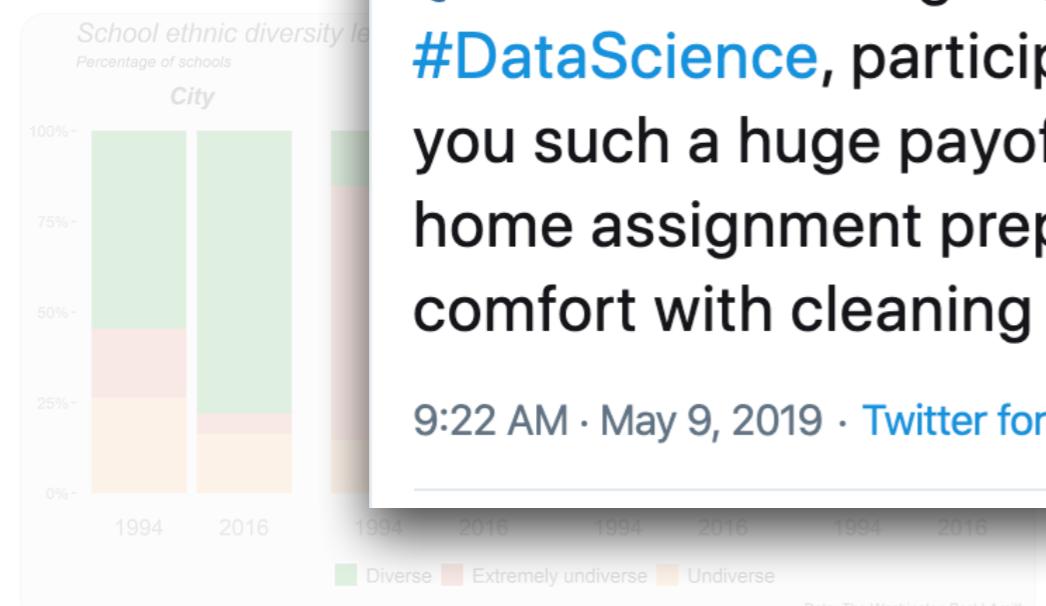




Amit Levinson
@Amit_Levinson

#TidyTuesday plot of school ethnic diversity levels across residency types for 1994 & 2016

I love seeing how it thoughts on minority diverse in cities > schools code: bit.ly/2nuiv5



Amit Levinson @Amit_Levinson · Sep 26

Replying to @Amit_Levinson

I learned a lot this week!

Analysis: using stringr package instead of base r to change strings.

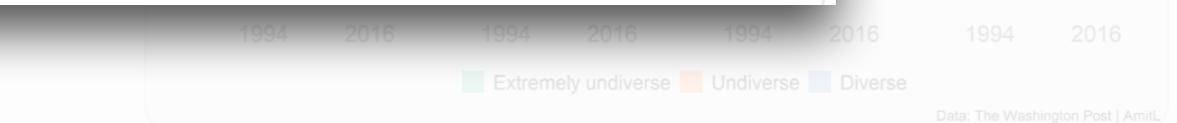
Visualization: used a stacked bar with 3 variables and various levels. I initially plot, but i think stacked



Rika Gorn 🎃
@RikaGorn

🗣 Friends looking to get into analytics and #DataScience, participating in #TidyTuesday will give you such a huge payoff including: a portfolio, take home assignment prep, #rstats & #dataviz learning, comfort with cleaning messy datasets, a community!

9:22 AM · May 9, 2019 · Twitter for Android



3:44 PM · Sep 26, 2019 · Twitter Web App

1

Final thoughts

- * Find what interests you and “follow” the people doing that work
- * Find opportunities outside of class to continue developing your skills
- * Stay current with what's happening in data science

Final thoughts

- * Find what interests you and “follow” the people doing that work
- * Find opportunities outside of class to continue developing your skills
- * Stay current with what's happening in data science

Have fun! 

Thank You!



maria.tackett@duke.edu



@MT_statistics



bit.ly/beyond-the-buzzword

Maria Tackett
Duke University

References

"What is Data Science"

<https://datascience.berkeley.edu/about/what-is-data-science/>

"Algorithms Gave Us the Black Hole Pictures. She's the 29-year-old Scientist Who Created Them",

<https://www.washingtonpost.com/science/2019/04/10/algorithms-gave-us-black-hole-pic-this-scientist-helped-create-them/>

"Share, Contribute, Collaborate, Broadcast".

<https://speakerdeck.com/minecr/share-contribute-collaborate-broadcast>