

Teaching intro data science

Maria Tackett
Duke University

Preparing to Teach
August 6, 2022

Courses I teach



Introduction to
Data Science

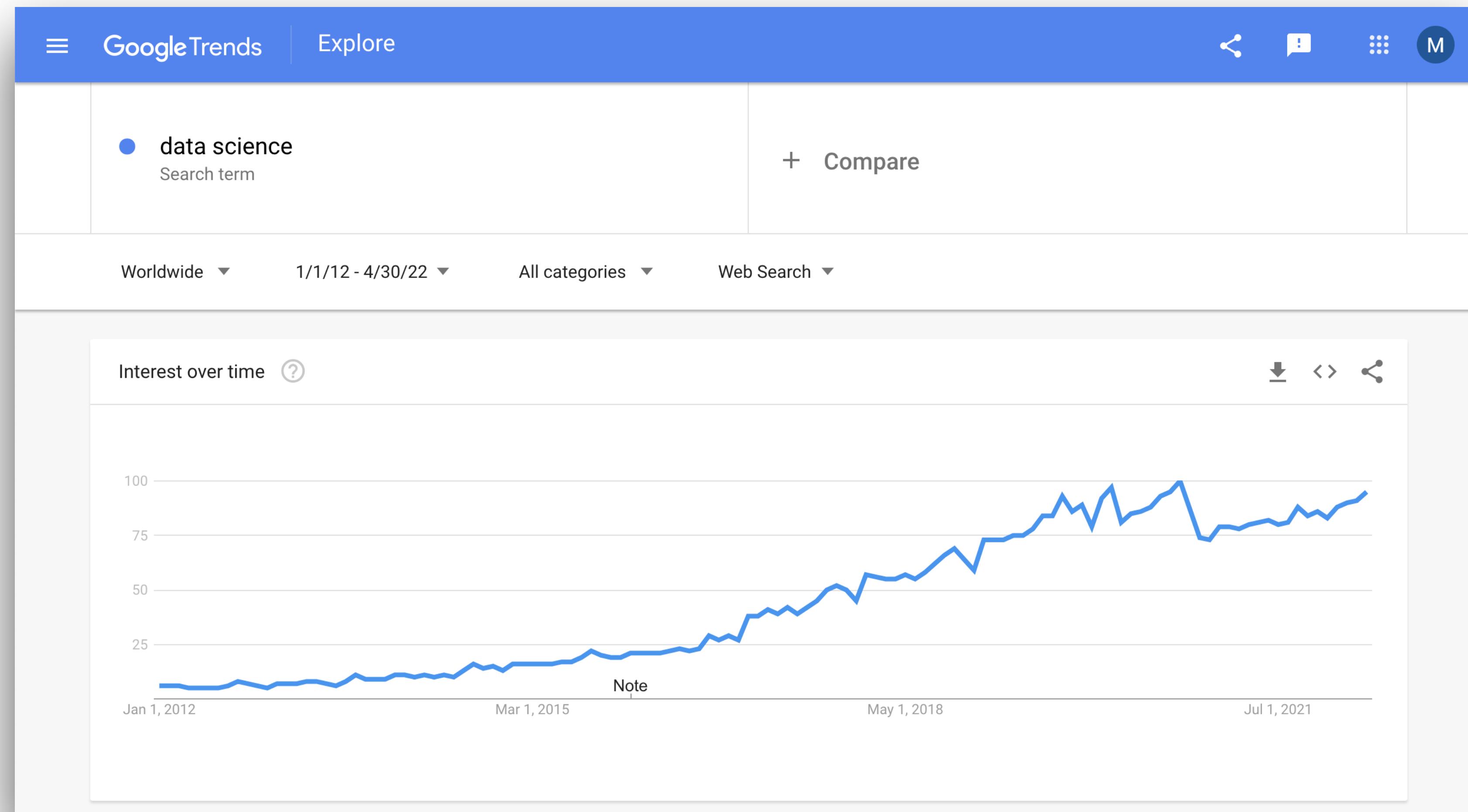


Regression
Analysis

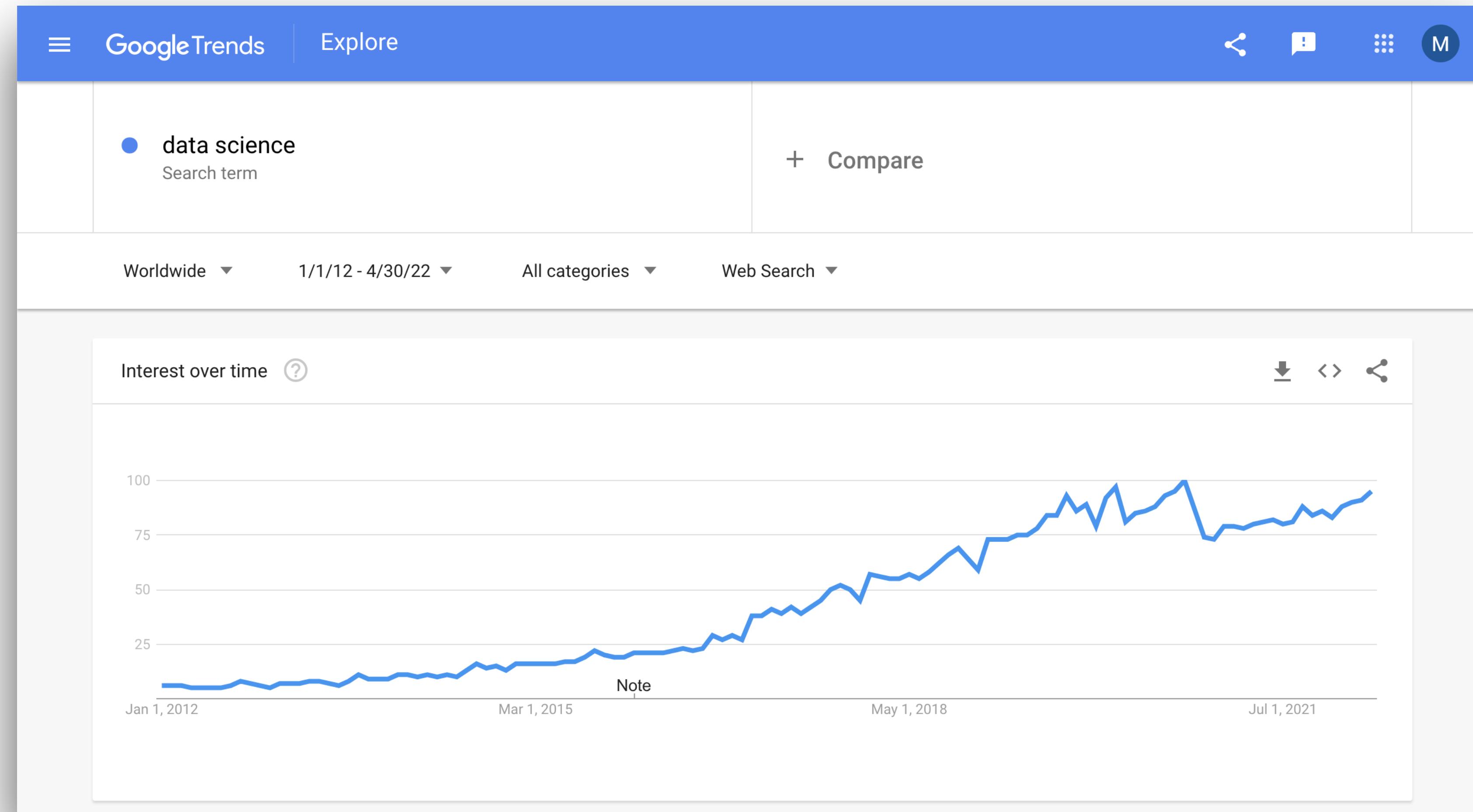


Generalized
Linear Models

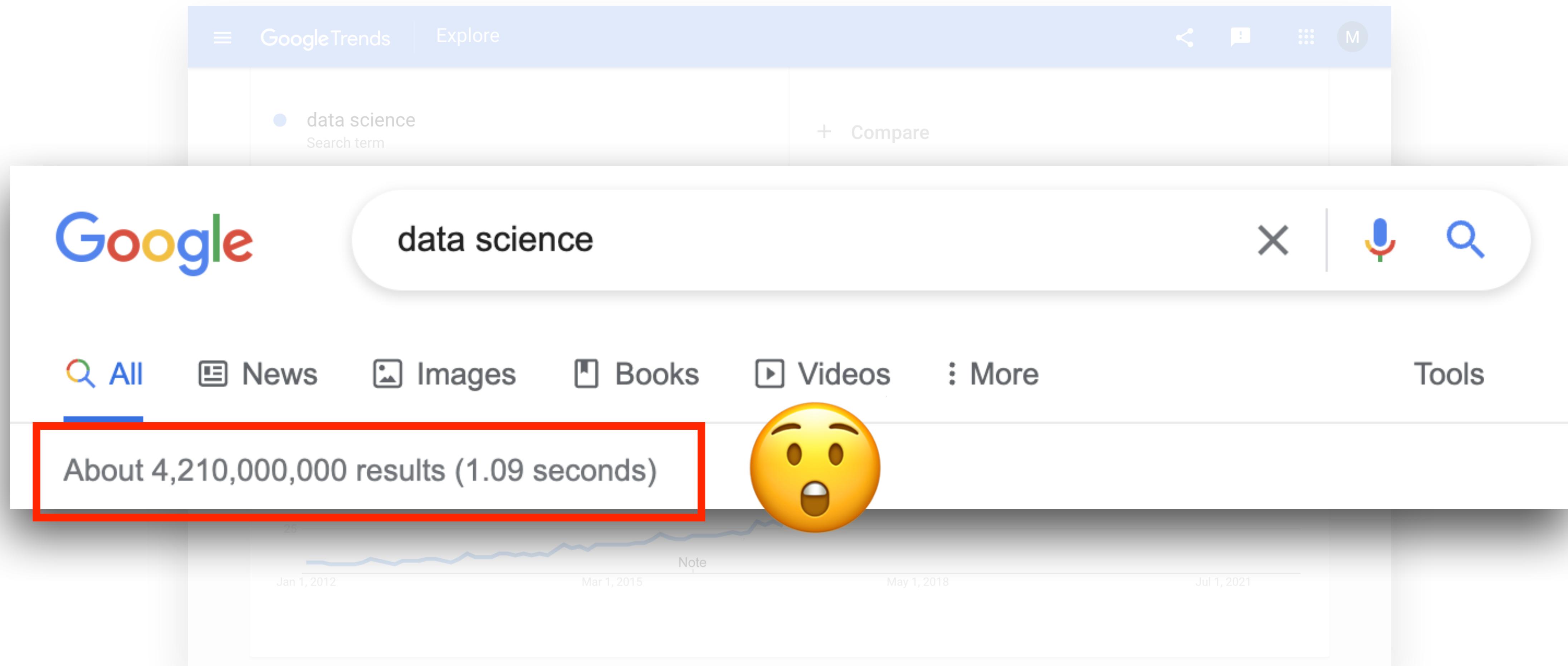
Data science has become increasingly popular over the past 10 years...



...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 results page with a light blue header bar containing the Google Trends logo, an 'Explore' button, and several icons. Below the header, a white box is titled 'People also ask'. Inside this box, the question 'What is data science in simple words?' is listed. A detailed answer follows, starting with the definition of data science as the study of data, involving recording, storing, and analyzing data to extract useful information. The text continues to explain that the goal of data science is to gain insights and knowledge from both structured and unstructured data. At the bottom of this section, there is a timestamp 'Aug 17, 2017'. Below the 'People also ask' box, a blue link reads 'Data Science Definition - The Tech Terms Computer Dictionary' and a green URL 'https://techterms.com › definition › data_science'.

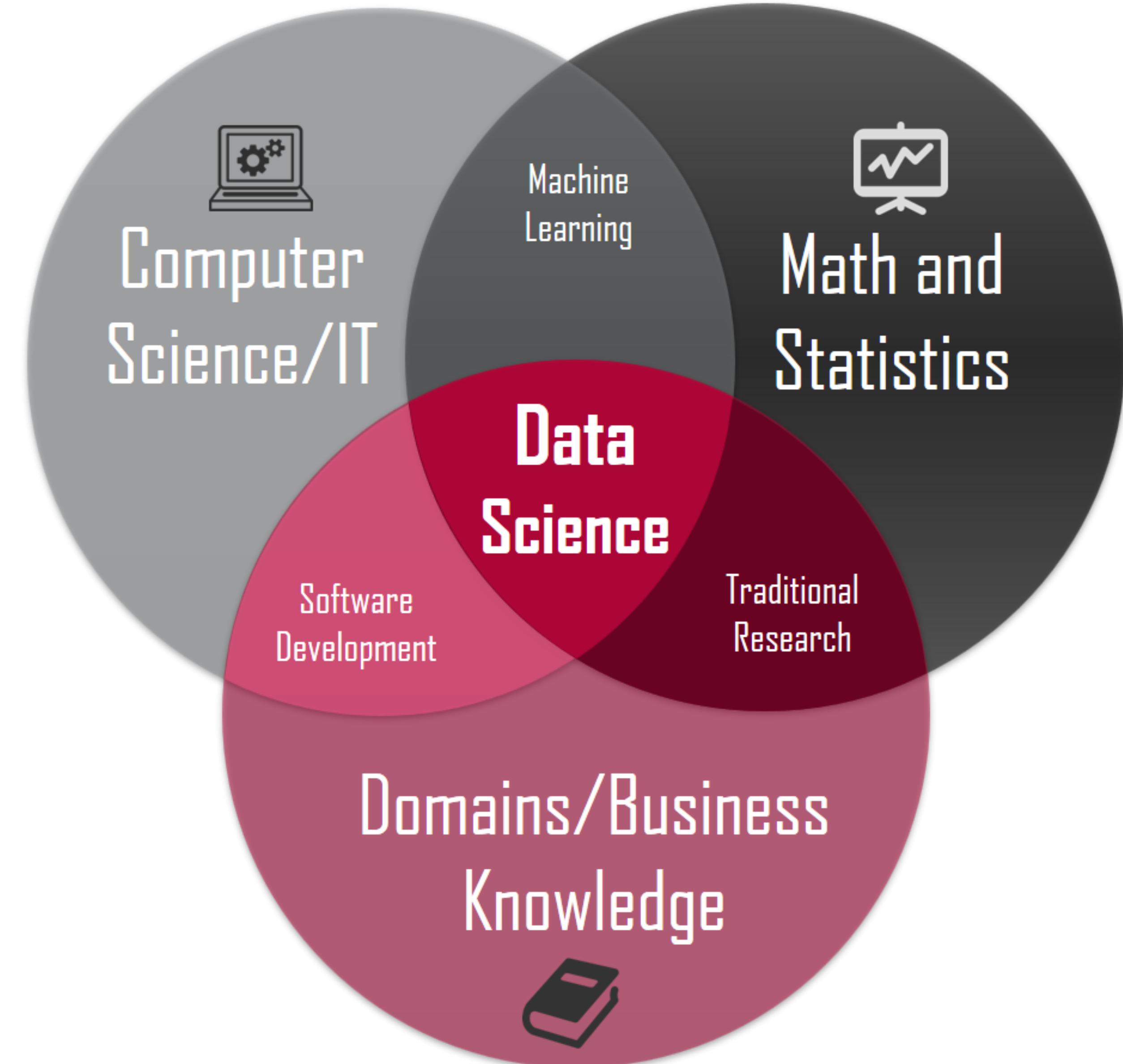
Google Trends Explore

People also ask

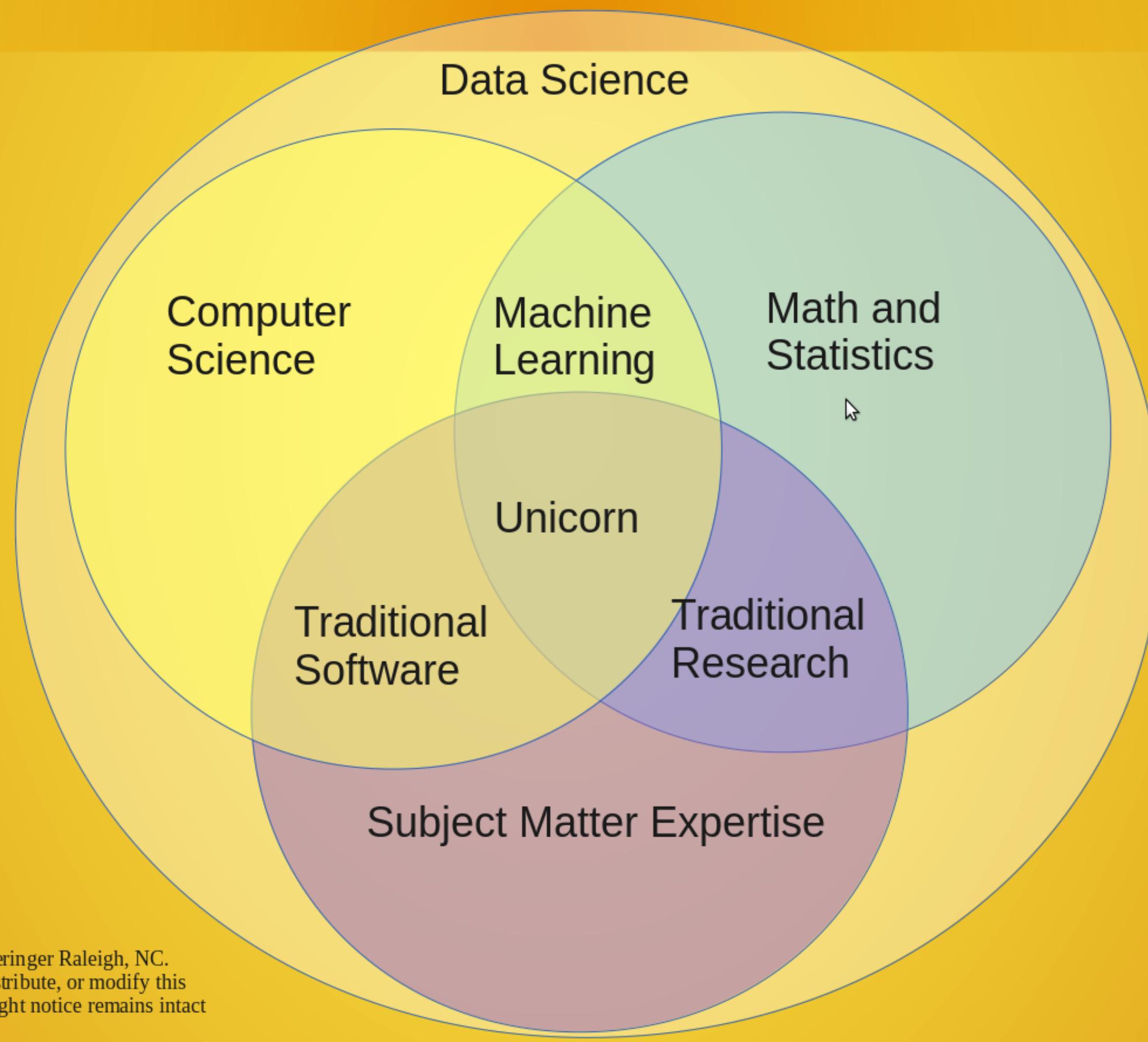
What is data science in simple words? ^

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](https://techterms.com/definition/data_science)



Data Science Venn Diagram v2.0



2016 GAISE Report

1. Teach statistical thinking.

- Teach statistics as an investigative process of problem-solving and decision-making.
- **Give students experience with multivariable thinking.**

2. Focus on conceptual understanding.

3. Integrate real data with a context and purpose.

4. Foster active learning.

5. Use technology to explore concepts and analyze data.

6. Use assessments to improve and evaluate student learning.

Course Learning Objectives

By the end of the semester, you will...

- ✓ learn to explore, visualize, and analyze data in a reproducible and shareable manner
- ✓ gain experience in data wrangling, exploratory data analysis, predictive modeling, and data visualization
- ✓ work on problems and case studies inspired by and based on real-world questions and data
- ✓ learn to effectively communicate results through written assignments and final project presentation

Traditional Intro Statistics vs. Intro Data Science

	Traditional Intro Statistics	Intro Data Science
Data	Structured, sometimes smaller data sets	Large data sets, structured and unstructured
Analysis purpose	Description, inference, interpretation	Description, inference, interpretation, prediction
Inference	Central Limit Theorem-based, more emphasis on equations	Simulation-based, more emphasis on conceptual understanding
Ethics	Sampling bias, misleading graphs	Sampling bias, misleading graphs, algorithmic bias, data privacy
Workflow	Focus on analysis workflow: exploration, inference / modeling, conclusion	“Start-to-finish” reproducible workflow
Computing	Range of technology from calculators to statistical programming	Technology for large data sets and reproducibility, primarily statistical programming

Data

Large data sets
structured & unstructured

	A	B	C	D	E	F	G
	year	id	county	trauma	totalvisits	admit	
1	2013	106014233	Alameda	0	43286	6186	
2	2013	106190758	Los Angeles	0	61915	8648	
3	2013	106300032	Orange	0	54217	5670	
4	2013	106301205	Orange	0	69558	10596	
5	2013	106301262	Orange	LEVEL II	42994	9796	
6	2013	106304113	Orange	0	14529	921	
7	2013	106380964	San Francisco	0	24783	2068	
8	2014	106014233	Alameda	LEVEL II	44068	4820	
9	2014	106190758	Los Angeles	0	50818	7370	
10	2014	106301262	Orange	LEVEL II	43440	10337	
11	2014	106304113	Orange	0	13632	1088	
12	2014	106380964	San Francisco	0	25213	2015	

Data

Large data sets
structured & unstructured



Data

Large data sets
structured & unstructured



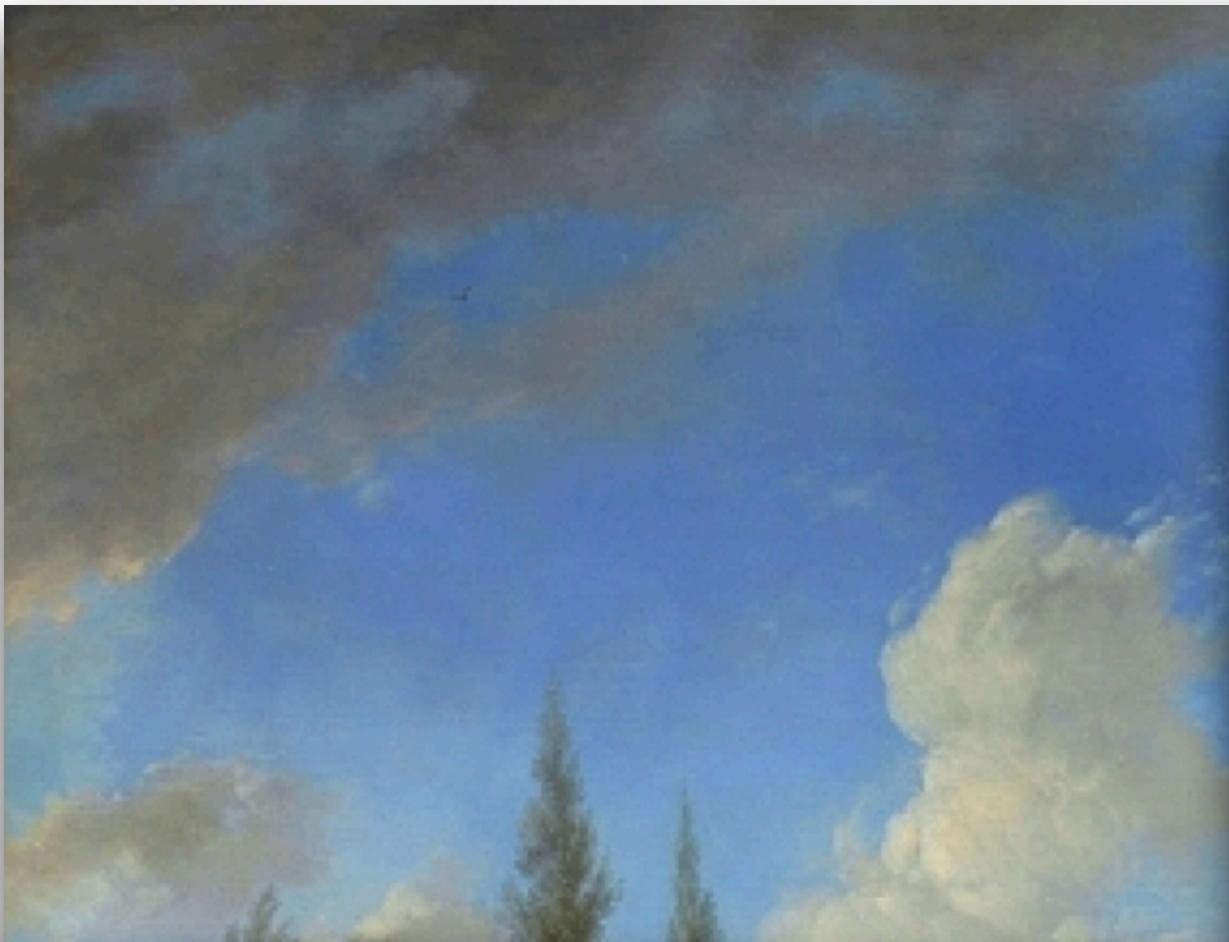
89 Deux tableaux très riches de composition , d'une belle exécution , & dont le mérite est très remarquable , chacun de 17 pouces 3 lignes de haut , sur 23 pouces de large ; le premier , peint sur bois , vient du Cabinet de Madame la Comtesse de Verrue ; il représente un départ pour la chasse : on y voit sur le devant un enfant sur un cheval blanc , un homme qui donne de la trompe pour rassembler les chiens , un Fauconnier & d'autres figures distribuées agréablement dans toute la largeur du tableau ; deux chevaux qui boivent à une fontaine ; à droite dans le coin une jolie maison de campagne surmontée d'une terrasse , & sur laquelle sont des gens à table , d'autres qui jouent des instru-

48 Tableaux .
ments ; des arbres & des fabriques enrichissent agréablement le fond . Le second tableau , qui est sur toile , fait voir un terrain d'une grande étendue , près la mer qui est à gauche , & sur laquelle sont des vaisseaux : on y voit aussi des bagages que l'on décharge d'un chariot , des hommes , des femmes , des enfants , deux chevaux qui mangent , & des mulets chargés de bagages .

Data

Large data sets

structured & unstructured



89 Deux tableaux très riches de composition, d'une belle exécution, & dont le mérite est très remarquable, chacun de 17 pouces 3 lignes de haut, sur 23 pouces de large ; le premier, peint sur bois, vient du Cabinet de Madame la Comtesse de Verrue ; il représente un départ pour la chasse : on y voit sur le devant un enfant sur un cheval blanc, un homme qui donne de la trompe pour rassembler les chiens, un Fauconnier & d'autres figures distribuées agréablement dans toute la largeur.

name	sale	lot	position	dealer	year	origin_author	origin_cat	school_pntg	diff_origin	logprice	price	count	subject
R1777-59	R1777	59	0.248945148	R	1777	D/FL	D/FL	D/FL	0	9.210340372	10000	1	Fte flamande
L1768-109a	L1768	109	0.534313725	L	1768	F	O	F	1	5.703782475	300	1	Paysage et fig
L1778b-76	L1778b	76	0.503311258	L	1778	D/FL	D/FL	D/FL	0	5.123963979	168	1	L'interieur d'u
R1777-131a	R1777	131	0.552742616	R	1777	D/FL	D/FL	D/FL	0	6.748173209	852.5	1	Matelot tenant
R1773-24a	R1773	24	0.421052632	R	1773	D/FL	D/FL	D/FL	0	6.684611728	800	1	(2) Deux marin
R1764-139	R1764	139	0.27689243	R	1764	F	F	F	0	4.394449155	81	1	Un lapin, une
R1777-211a	R1777	211	0.890295359	R	1777	F	F	F	0	7.094234846	1205	1	Buste de jeun
R1767-196	R1767	196	0.616352201	R	1767	D/FL	D/FL	D/FL	0	7.787382026	2410	1	Une Tabagie c
R1776-103	R1776	103	0.311178248	R	1776	D/FL	D/FL	D/FL	0	8.006367568	3000	1	Joueurs de tri
R1767-157b	R1767	157	0.493710692	R	1767	D/FL	D/FL	D/FL	0	7.150701458	1275	1	(2) Deux Table
P1775-13	P1775	13	0.173333333	P	1775	D/FL	D/FL	D/FL	0	6.173786104	480	1	Une jeune fille
R1777-130	R1777	130	0.548523207	R	1777	D/FL	D/FL	D/FL	0	6.29156914	540	1	Autoportrait
L1764-16b	L1764	16	0.262295082	L	1764	I	O	I	1	2.48490665	12	1	(2) Fleurs & Fr
R1771-65b	R1771	65	0.380116959	R	1771	F	F	F	0	3.871201011	48	1	(2) Filles de Jž
J1775-16	J1775	16	0.8	J	1775	F	F	F	0	6.802394763	900	1	Jupiter, sous l
R1765-40b	R1765	40	0.655737705	R	1765	F	F	F	0	6.416732283	612	1	(2) une SoirŽe
J1779-23	J1779	23	1	J	1779	F	F	F	0	1.098612289	3	1	Paysage ornŽ

Examples from Data Science in a Box

Data

Large data sets
structured & unstructured

Examples from Data Science in a Box

imdb.com/chart/top/

IMDb Charts

IMDb Top 250 Movies

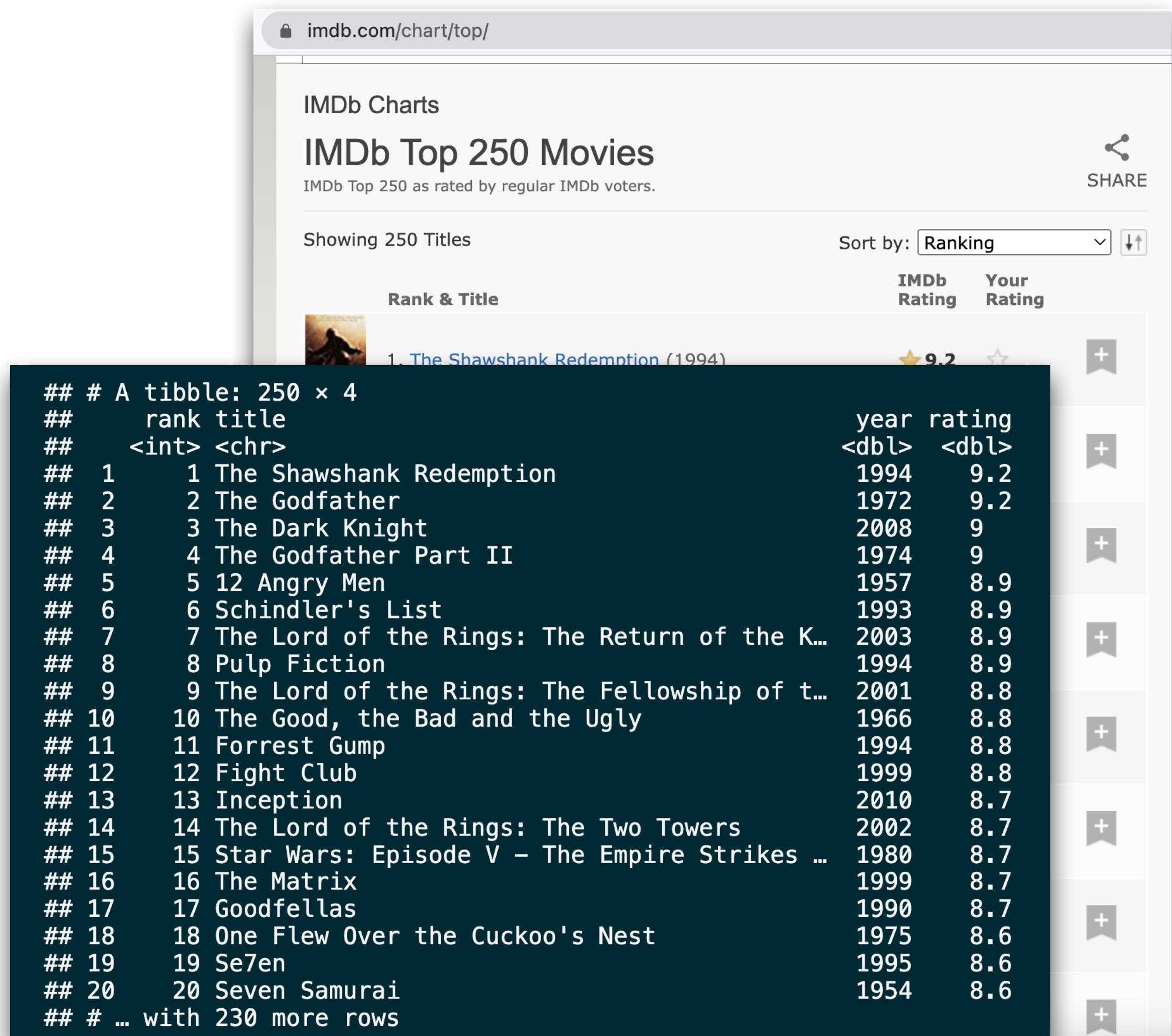
IMDb Top 250 as rated by regular IMDb voters.

Showing 250 Titles Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating	
1. The Shawshank Redemption (1994)	★ 9.2	☆	
2. The Godfather (1972)	★ 9.2	☆	
3. The Dark Knight (2008)	★ 9.0	☆	
4. The Godfather: Part II (1974)	★ 9.0	☆	
5. 12 Angry Men (1957)	★ 8.9	☆	
6. Schindler's List (1993)	★ 8.9	☆	
7. The Lord of the Rings: The Return of the King (2003)	★ 8.9	☆	
8. Pulp Fiction (1994)	★ 8.9	☆	

Data

Large data sets
structured & unstructured



The screenshot shows the IMDb Top 250 Movies chart page. The page title is "IMDb Charts" and the specific chart is "IMDb Top 250 Movies". It states "IMDb Top 250 as rated by regular IMDb voters." and "Showing 250 Titles". The sorting is set to "Ranking". The table has columns for "Rank & Title", "IMDb Rating", and "Your Rating". The first entry is "The Shawshank Redemption (1994)" with a rating of 9.2. The data is presented in a tibble format:

```
## # A tibble: 250 x 4
##   rank title      year rating
##   <int> <chr>    <dbl>  <dbl>
## 1     1 The Shawshank Redemption 1994   9.2
## 2     2 The Godfather           1972   9.2
## 3     3 The Dark Knight        2008   9.0
## 4     4 The Godfather Part II  1974   9.0
## 5     5 12 Angry Men          1957   8.9
## 6     6 Schindler's List       1993   8.9
## 7     7 The Lord of the Rings: The Return of the King 2003   8.9
## 8     8 Pulp Fiction          1994   8.9
## 9     9 The Lord of the Rings: The Fellowship of the Ring 2001   8.8
## 10    10 The Good, the Bad and the Ugly 1966   8.8
## 11    11 Forrest Gump           1994   8.8
## 12    12 Fight Club             1999   8.8
## 13    13 Inception              2010   8.7
## 14    14 The Lord of the Rings: The Two Towers 2002   8.7
## 15    15 Star Wars: Episode V – The Empire Strikes Back 1980   8.7
## 16    16 The Matrix              1999   8.7
## 17    17 Goodfellas            1990   8.7
## 18    18 One Flew Over the Cuckoo's Nest 1975   8.6
## 19    19 Se7en                  1995   8.6
## 20    20 Seven Samurai         1954   8.6
## # ... with 230 more rows
```

Analysis purpose

Linear Regression

- Estimation
- Interpretation
- Inference on slope
- Predicted values

Description, inference,
interpretation, prediction

Analysis purpose

Linear Regression

- Estimation
- Interpretation
- Inference on slope
- Predicted values
- **Model selection**
- **Prediction intervals**
- **Cross-validation**

Description, inference,
interpretation, prediction

Inference

Simulation-based, more
emphasis on conceptual
understanding

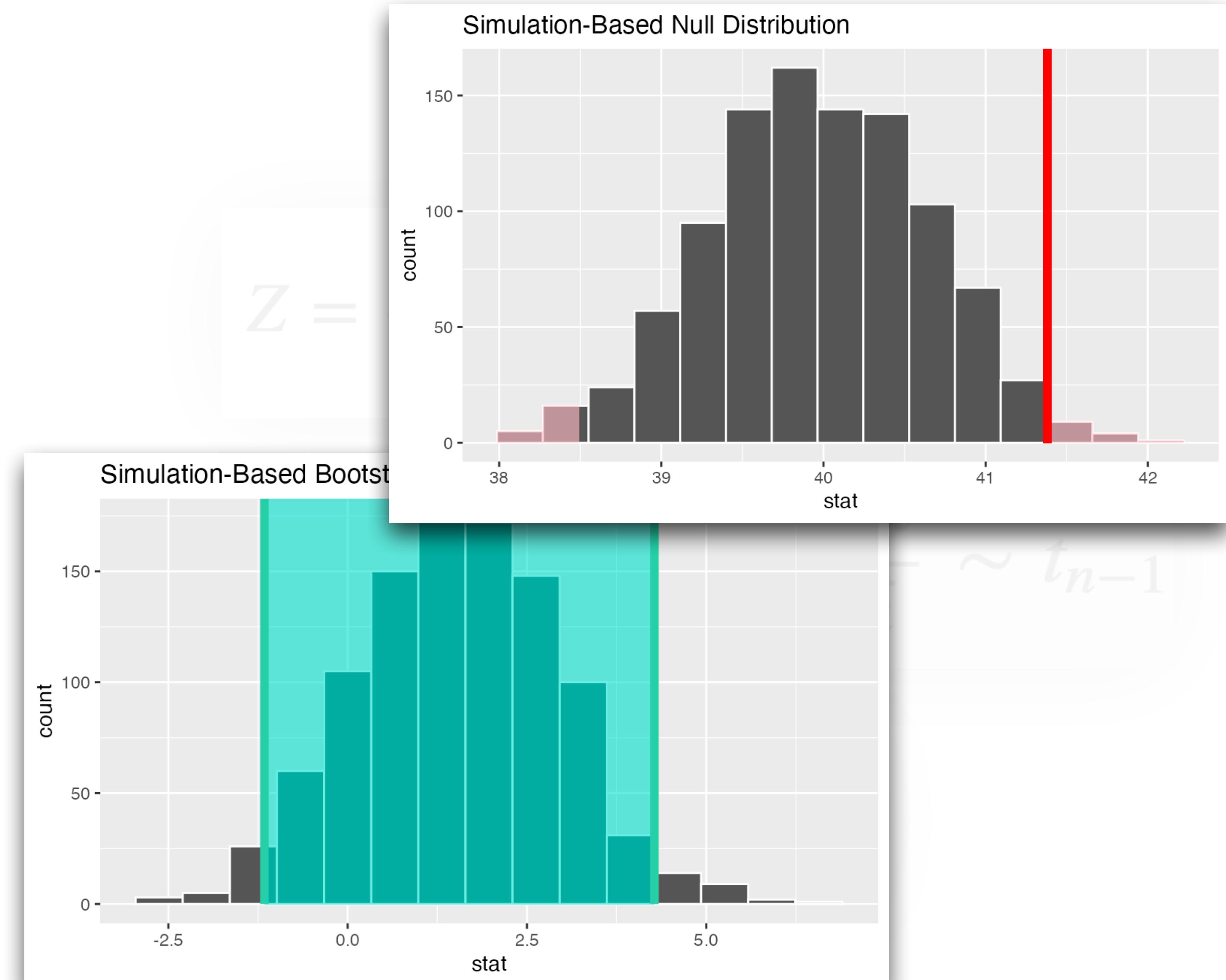
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

$$\bar{x} \pm t_{n-1}^* \times \frac{s}{\sqrt{n}}$$

Inference

Simulation-based, more
emphasis on conceptual
understanding



Graphs from [infer website](#)

Ethics

Sampling bias, misleading graphs, algorithmic bias, data privacy



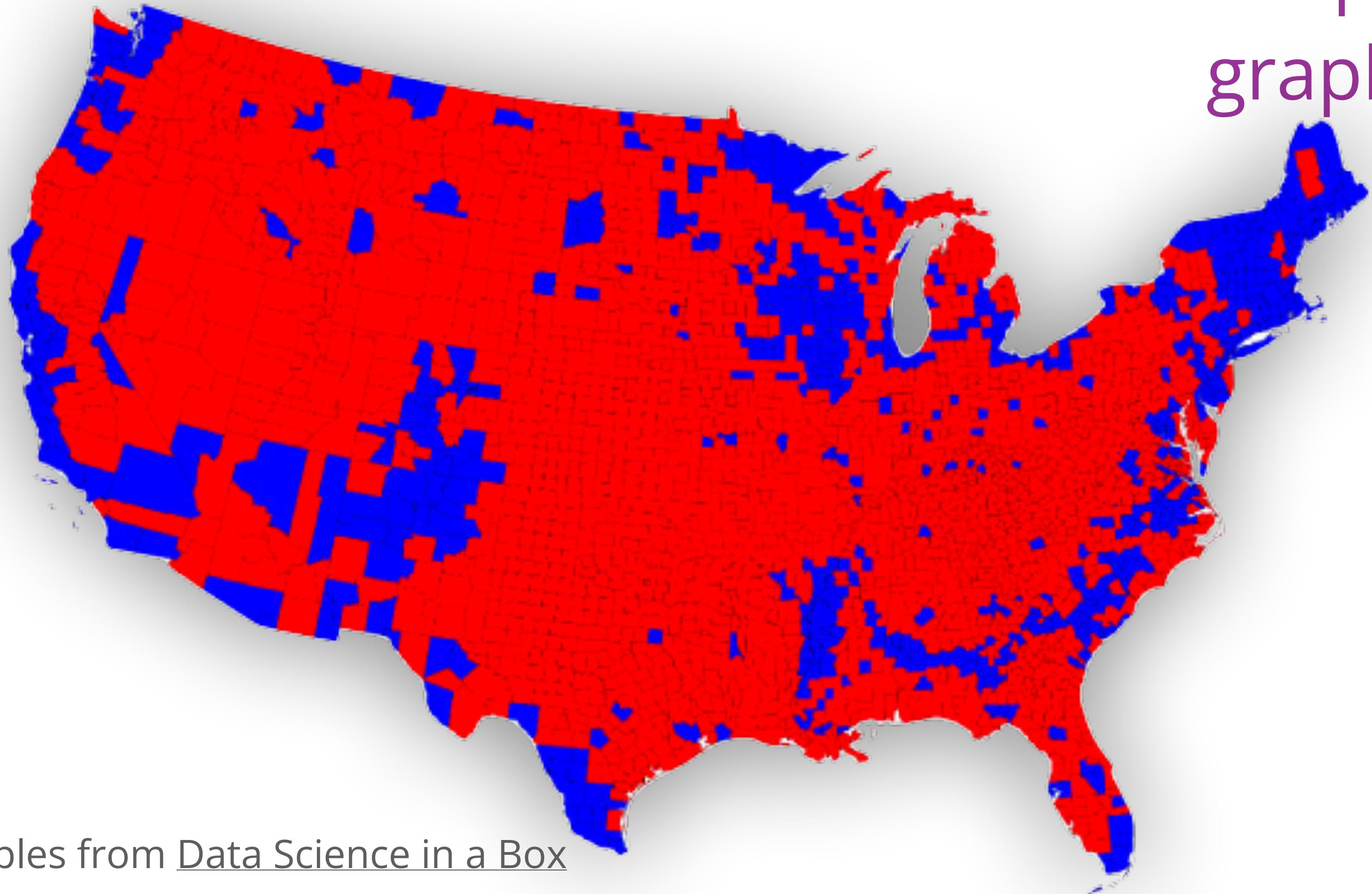
Elon Musk  @elonmusk · May 13

To find out, my team will do a random sample of 100 followers of [@twitter](#).

I invite others to repeat the same process and see what they discover ...

13K 13.1K 124.2K

Ethics



Sampling bias, misleading
graphs, algorithmic bias,
data privacy

Ethics

Sampling bias, misleading graphs, algorithmic bias, data privacy

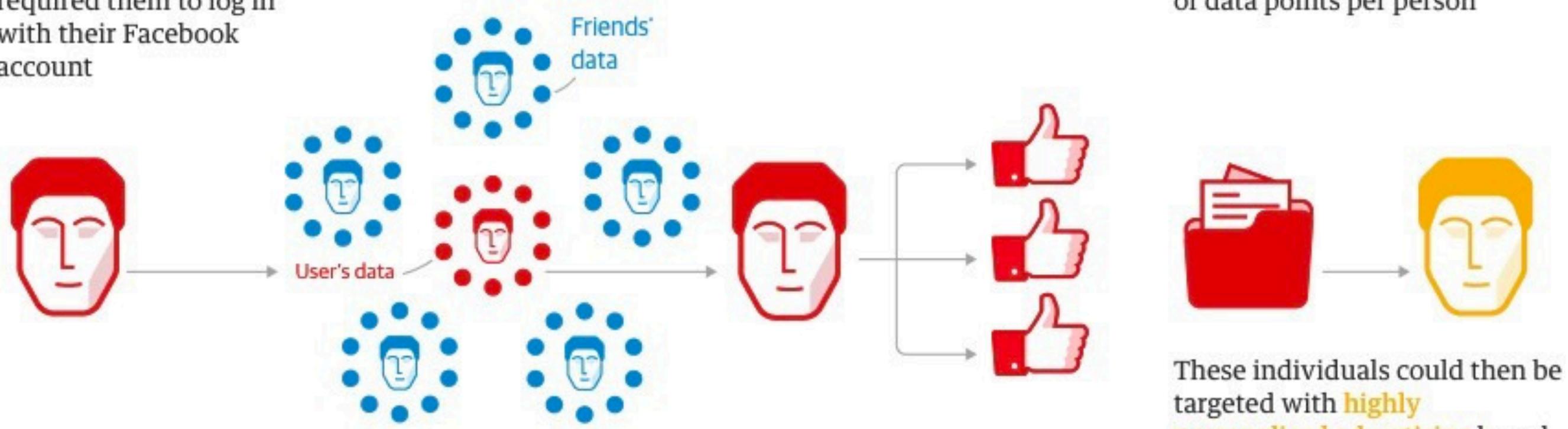
Cambridge Analytica: how 50m Facebook records were hijacked

1 Approx. 320,000 US voters ('seeders') were paid \$2-5 to take a detailed personality/political test that required them to log in with their Facebook account

2 The app also collected data such as likes and personal information from the test-taker's Facebook account ...

3 The personality quiz results were paired with their Facebook data - such as likes - to seek out psychological patterns

4 Algorithms combined the data with other sources such as voter records to create a superior set of records (initially 2m people in 11 key states*), with hundreds of data points per person



Guardian graphic. *Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

Ethics



Interview

'A white mask worked better': why algorithms are not colour blind

Ian Tucker

When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing

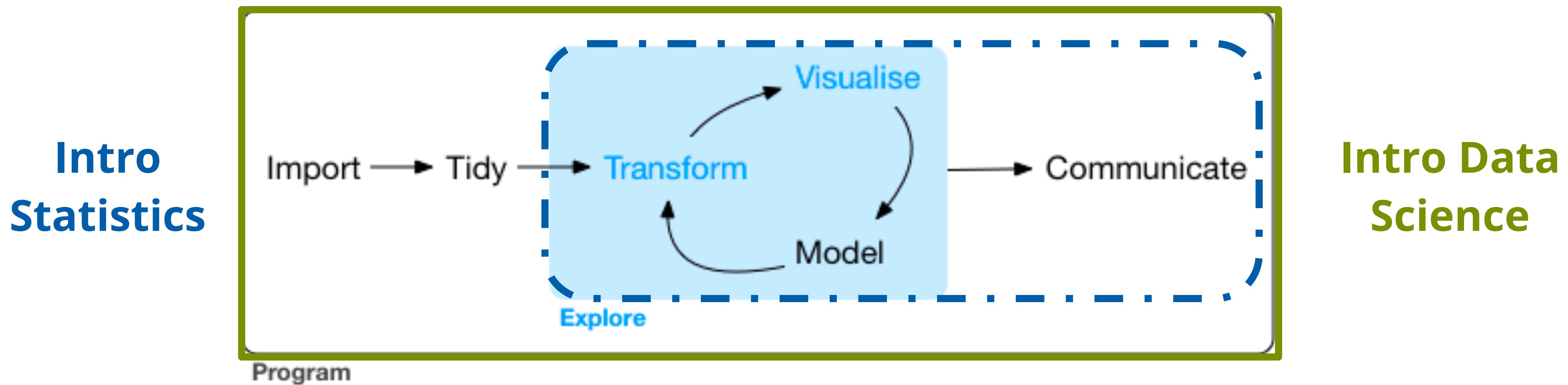
Sun 28 May 2017 13.27 BST

Joy Buolamwini is a graduate researcher at the MIT Media Lab and founder of the Algorithmic Justice League – an organisation that aims to challenge the biases in decision-making software. She grew up in Mississippi, gained a Rhodes scholarship, and she is also a Fulbright fellow, an Astronaut scholar and a Google Anita Borg scholar. Earlier this year she won a \$50,000 scholarship funded by the makers of the film *Hidden Figures* for her work fighting coded discrimination.

Sampling bias, misleading graphs, algorithmic bias, data privacy

Workflow

“Start-to-finish” reproducible workflow



Computing

Technology for large data sets
and reproducibility, primarily
statistical programming

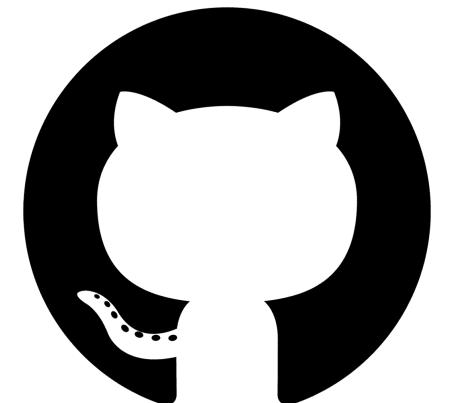


Computing toolkit in STA 199



Studio®

- R Markdown / Quarto for write up
- Run Git commands using point-and-click interface
- Server-based RStudio*
 - Git already configured
 - Same set up for all students



- Assign and submit assignments
- Facilitates collaboration on group assignments
- Course management using **ghclass** R package (or GitHub Classroom**)

*Çetinkaya-Rundel, M., and Rundel, C. (2018), "Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum," *The American Statistician*, 72, 58–65,

**Fiksel, J., Jager, L. R., Hardin, J. S., and Taub, M. A. (2019), "Using GitHub Classroom to Teach Statistics," *Journal of Statistics Education*, 27, 100–119.

Assessing student learning

Types of assessments

- In-class exercises, computing labs, homework
- Exams, final project

Tips

- Design assessments to emphasize skills students will use in practice
- Provide scaffolding early on, particularly for code
- Give opportunities for practice before graded assignments

Getting started

Consider the course learning objectives + the student population

- What statistics and computational skills do they have coming into the course? Are there prerequisites?
- Are students preparing for the workplace? Subsequent statistics courses? Both?
- What skills do they need to prepare for the next step?

Traditional
Intro Statistics

Intro
Data Science

Data Science in a Box

Collection of intro data science slides, assignments, and other resources
by Mine Çetinkaya-Rundel

Overview Hello #dsbox! Content Infrastructure Design

Q Cloud ⌂ Q

Welcome

How can we effectively and efficiently teach data science to students with little to no background in computing and statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and leave them wanting to learn more? This introductory data science course is our (working) answer to this question.

The source code for everything you see here can be found [on GitHub](#).



On this page

License

Acknowledgements

Edit this page

Report an issue

Designing the Data Science Classroom

rstudio::conf(2022) workshop on teaching data science using R with
Mine Çetinkaya-Rundel

Designing the data science classroom

July 25 and 26, 2022

09:00 - 17:00

Maryland 3

[Click here to register](#)

On this page

- Overview
- Learning objectives
- Is this course for me?
- Prework
- RStudio Cloud
- Instructors

[Edit this page](#)

[Report an issue](#)

rstd.io/teach-ds-conf22

Resources

Websites

- [Data Science in a Box](#) by Mine Çetinkaya-Rundel
- [Designing the Data Science Classroom](#) rstudio::conf(2022) workshop by Mine Çetinkaya-Rundel and Maria Tackett

Textbooks (free online)

- [Modern Data Science with R](#) by Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton
- [Introduction to Modern Statistics](#) by Mine Çetinkaya-Rundel and Johanna Hardin
- [Statistical Inference Via Data Science \(Modern Dive\)](#) by Chester Ismay and Albert Y. Kim

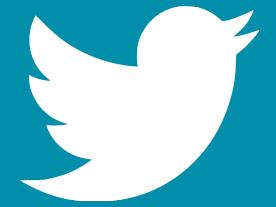
Reports

- [Data Science for Undergraduates: Opportunities and Options](#)
- [Computing Competencies for Undergraduate Data Science Curricula](#)
- [The Two-Year College Data Science Summit](#)

Thank You!



maria.tackett@duke.edu



@MT_statistics