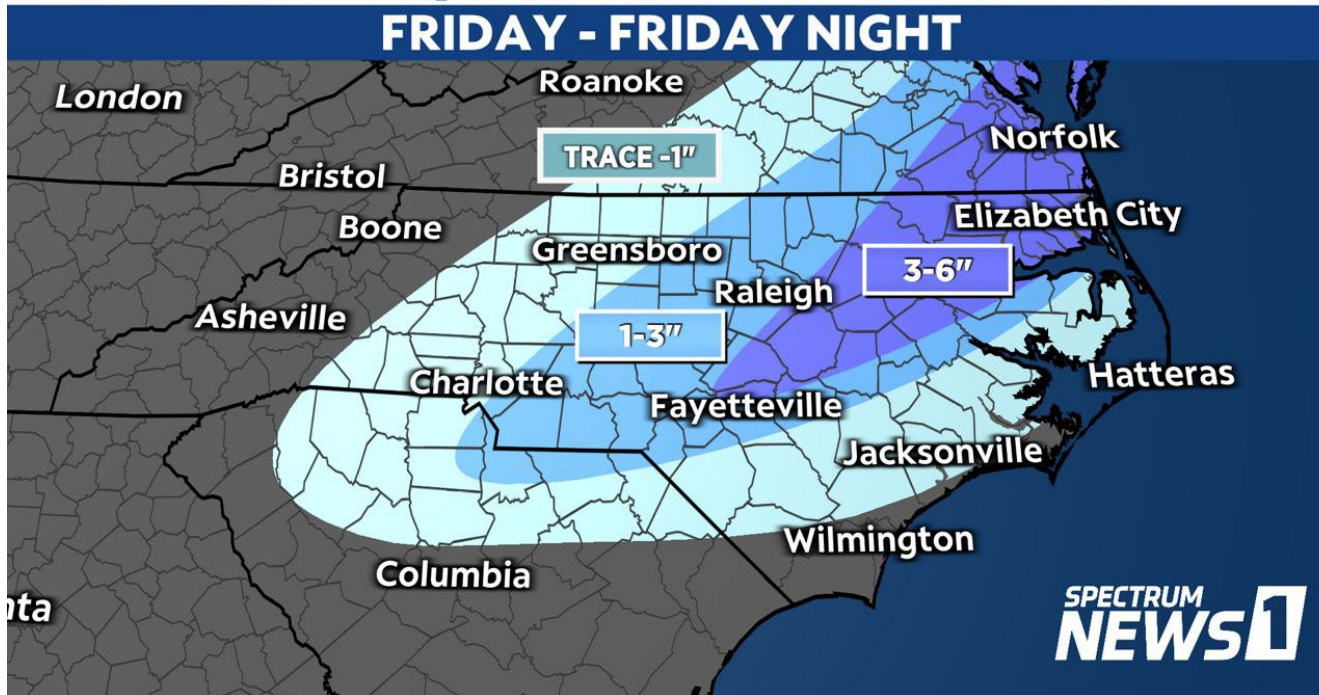


Introduction to Data Science

CSC 405/605/705

Can weather data predict whether school will close?

SNOW/SLEET ACCUMULATION FRIDAY - FRIDAY NIGHT



WINTER STORM EXPECTED TOMORROW

Starting: 8:00 AM

Expected snowfall: 6–10 inches

Temperature: 25°F

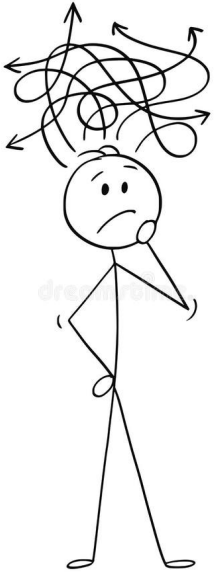
Wind: 15–25 mph

Conditions: Heavy snow / Icy roads



Do I need to go to
school tomorrow?

What information would you need to make this decision?

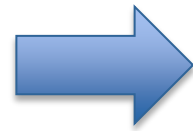


- How much snow?
- When will the snow start?
- How long will it last?
- Temperature?
- Ice?
- Wind?
- Road conditions?
- How did the school respond in similar situations?

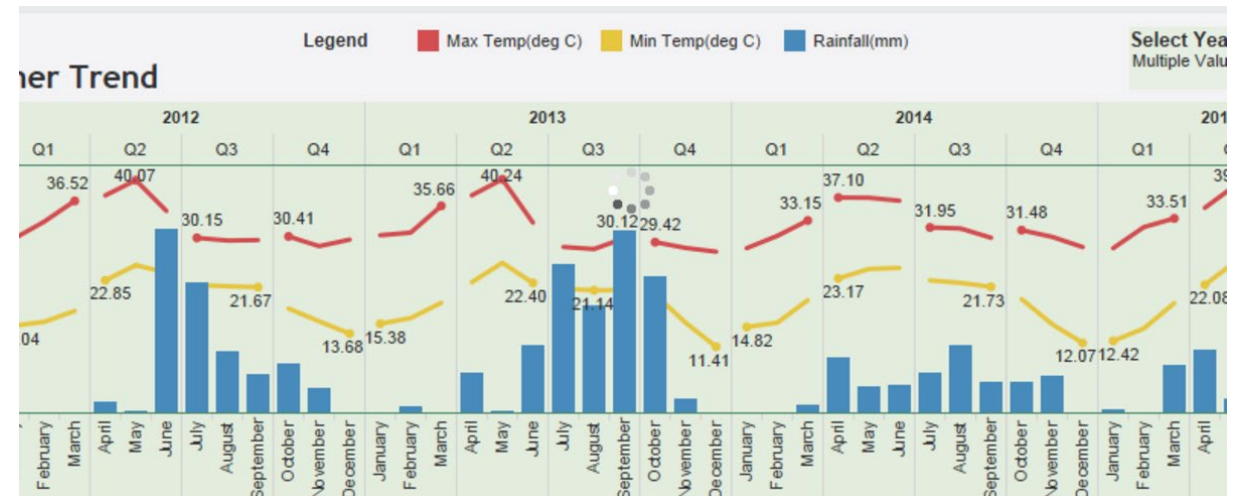
What would you do to make this decision?

- Gather past data: snowfall, temperature, ice, wind, past closures.

date	nearestptn_1	nearestptn_2	nearestptn_3	nearestptn_4	nearestptn_5	nearestptn_6	nearestptn_7	nearestptn_8	nearestptn_9
2015-01-04	-4.0	-4.0	-5.0	-11.0	-8.0	-12.0	1016.0	1022.0	
2015-01-05	-14.0	-4.0	-5.0	-10.0	-11.0	-8.0	1030.0	1016.0	
2015-01-06	-9.0	-14.0	-4.0	-14.0	-19.0	-11.0	1032.0	1033.0	
2015-01-07	-10.0	-8.0	-14.0	-15.0	-14.0	-19.0	1036.0	1032.0	
2015-01-08	-10.0	-10.0	-9.0	-22.0	-15.0	-14.0	1035.0	1036.0	
2015-01-09	-7.0	-10.0	-10.0	-12.0	-22.0	-15.0	1034.0	1035.0	
2015-01-10	-11.0	-7.0	-10.0	-10.0	-12.0	-22.0	1035.0	1024.0	
2015-01-11	-6.0	-11.0	-7.0	-12.0	-19.0	-12.0	1023.0	1035.0	
2015-01-12	-5.0	-6.0	-11.0	-11.0	-12.0	-19.0	1024.0	1023.0	
2015-01-13	-13.0	-8.0	-9.0	-17.0	-15.0	-12.0	1040.0	1024.0	
2015-01-14	-12.0	-13.0	-5.0	-18.0	-17.0	-11.0	1017.0	1040.0	
2015-01-15	-2.0	-12.0	-13.0	-8.0	-10.0	-17.0	1026.0	1037.0	
2015-01-16	1.0	-3.0	-12.0	-6.0	-8.0	-19.0	1023.0	1026.0	
2017-06-14	21.0	22.0	21.0	15.0	14.0	17.0	1030.0	1014.0	
2017-06-15	27.0	23.0	22.0	13.0	13.0	14.0	1035.0	1030.0	
2017-06-16	27.0	27.0	23.0	15.0	12.0	13.0	1036.0	1030.0	
2017-06-17	19.0	27.0	27.0	17.0	15.0	13.0	1012.0	1009.0	
2017-06-18	15.0	19.0	27.0	9.0	17.0	15.0	1021.0	1011.0	
2017-06-19	21.0	16.0	19.0	10.0	8.0	12.0	1014.0	1021.0	
2017-06-20	25.0	21.0	16.0	10.0	10.0	8.0	1025.0	1014.0	
2017-06-21	22.0	26.0	21.0	13.0	10.0	10.0	1017.0	1016.0	
2017-06-22	25.0	22.0	20.0	15.0	10.0	10.0	1030.0	1027.0	
2017-06-23	30.0	25.0	20.0	20.0	10.0	13.0	1030.0	1030.0	
2017-06-24	29.0	30.0	20.0	10.0	20.0	10.0	1011.0	1030.0	
2017-06-25	24.0	20.0	30.0	10.0	10.0	20.0	1012.0	1011.0	
2017-06-26	10.0	24.0	20.0	15.0	10.0	10.0	1014.0	1012.0	
2017-06-27	10.0	10.0	24.0	11.0	10.0	10.0	1010.0	1014.0	



- Look for patterns with simple charts.



What would you do to make this decision?

- Make a prediction rule or basic model.

date	meansepm_1	meansepm_2	meansepm_3	meansepm_4	meansepm_5	meansepm_6	meansepm_7	meansepm_8
2015-01-04	-4.0	-8.0	-9.0	-11.0	-8.0	-12.0	1016.0	1022.0
2015-01-05	-14.0	-4.0	-9.0	-10.0	-11.0	-8.0	1033.0	1018.0
2015-01-06	-9.0	-14.0	-4.0	-14.0	-19.0	-11.0	1032.0	1033.0
2015-01-07	-10.0	-8.0	-14.0	-15.0	-14.0	-10.0	1034.0	1032.0
2015-01-08	-15.0	-10.0	-9.0	-22.0	-15.0	-14.0	1035.0	1030.0
2015-01-09	-7.0	-10.0	-13.0	-12.0	-22.0	-15.0	1034.0	1035.0
2015-01-10	-11.0	-7.0	-10.0	-10.0	-12.0	-22.0	1035.0	1024.0
2015-01-11	-6.0	-11.0	-7.0	-12.0	-10.0	-12.0	1023.0	1035.0
2015-01-12	-5.0	-6.0	-11.0	-11.0	-12.0	-10.0	1024.0	1023.0
2015-01-13	-13.0	-5.0	-9.0	-17.0	-11.0	-12.0	1040.0	1024.0
2015-01-14	-12.0	-13.0	-5.0	-10.0	-17.0	-11.0	1037.0	1040.0
2015-01-15	-2.0	-12.0	-13.0	-8.0	-10.0	-17.0	1026.0	1037.0
2015-01-16	1.0	-2.0	-12.0	-6.0	-8.0	-10.0	1023.0	1026.0
2017-09-14	21.0	22.0	21.0	13.0	14.0	17.0	1030.0	1014.0
2017-09-15	27.0	27.0	22.0	13.0	13.0	14.0	1035.0	1030.0
2017-09-16	27.0	27.0	22.0	10.0	11.0	13.0	1036.0	1035.0
2017-09-17	19.0	27.0	27.0	17.0	19.0	13.0	1012.0	1009.0
2017-09-18	15.0	19.0	27.0	9.0	17.0	15.0	1021.0	1012.0
2017-09-19	21.0	19.0	19.0	10.0	8.0	17.0	1014.0	1021.0
2017-09-20	25.0	24.0	16.0	10.0	10.0	8.0	1035.0	1014.0
2017-09-21	22.0	26.0	21.0	13.0	10.0	10.0	1037.0	1036.0
2017-09-22	25.0	22.0	20.0	10.0	11.0	10.0	1030.0	1027.0
2017-09-23	30.0	25.0	20.0	20.0	10.0	13.0	1030.0	1009.0
2017-09-24	30.0	30.0	20.0	13.0	20.0	10.0	1011.0	1009.0
2017-09-25	24.0	20.0	30.0	10.0	10.0	20.0	1012.0	1011.0
2017-09-26	10.0	24.0	20.0	15.0	10.0	10.0	1014.0	1012.0
2017-09-27	10.0	10.0	24.0	11.0	10.0	10.0	1010.0	1014.0

WINTER STORM EXPECTED TOMORROW

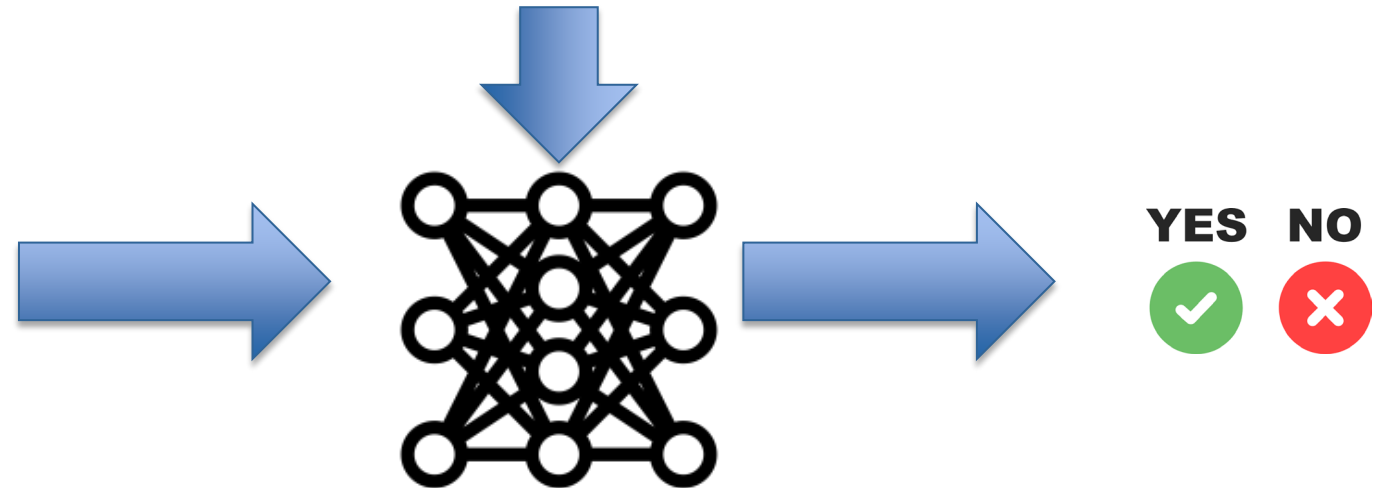
Starting: 8:00 AM

Expected snowfall: 6–10 inches

Temperature: 25°F

Wind: 15–25 mph

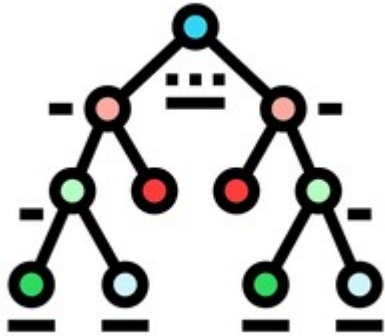
Conditions: Heavy snow / Icy roads



- Test it on new days

What would you do to make this decision?

- Something could go wrong...



Snowfall: 10 inches
Temperature: 25°F
Strong wind
Icy roads



School should be closed

The school was open.

"When snowfall > 8 inches, temperature < 30°F, and roads are icy, the school usually closes."



What would you do to make this decision?

- Something could go wrong...
 - The snow happened during the weekend, so roads were already cleared.
 - The storm arrived later than expected.
 - The university had fewer safety concerns than usual.
 - The decision maker used additional information.
 - The historical data was incomplete.

What would you do to make this decision?

- Most winter storms with more than 8 inches of snow lead to school closure. But this year, many similar storms did not.

Null hypothesis (H_0):

The school closure policy has not changed.

Alternative hypothesis (H_1):

The school uses a different decision rule now.

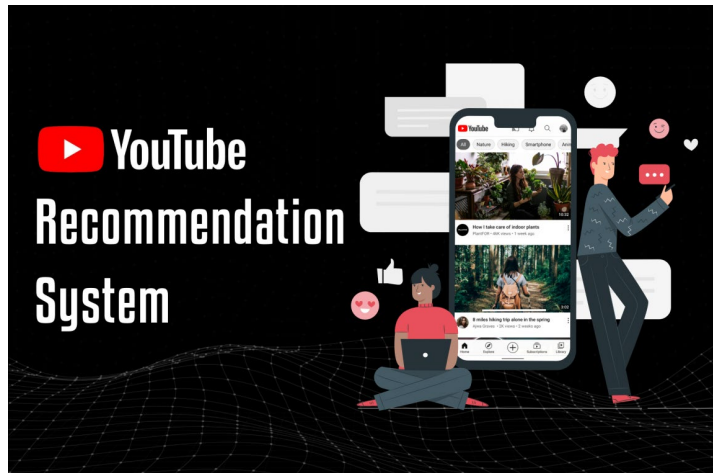
Is the difference large enough that we believe something has changed?

That whole process—asking a question, using data, finding patterns, making and testing predictions, and considering limitations—is data science.

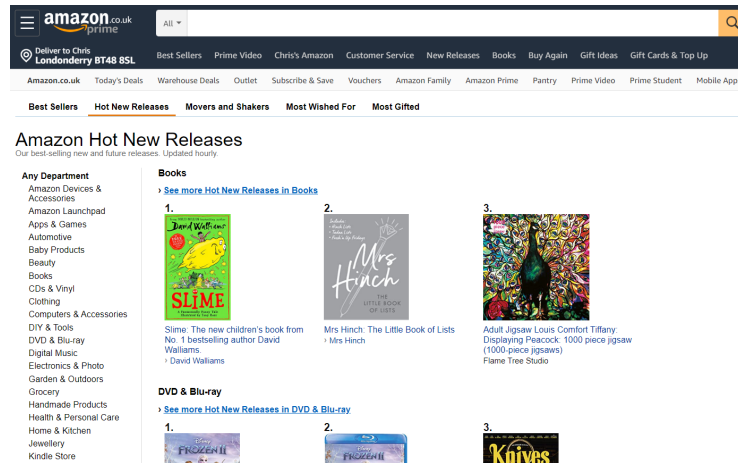
Using Evidence to Answer a Mystery

- “Is our school cafeteria line predictable?”
- “Can weather data predict whether school will close?”
- “Which movie features predict box-office success?”
- “Do sports statistics really identify the best player?”
- “Can we detect misleading graphs on social media?”

Data Science Is Everywhere



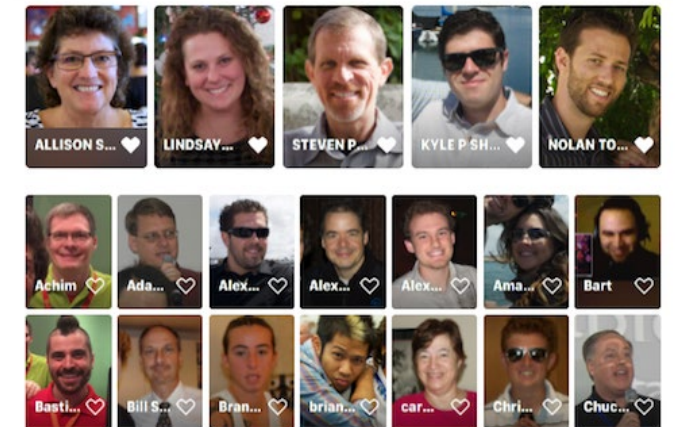
How does YouTube know what you want to watch



How does Amazon know what are hot products

People

111 People



How does Apple group the photos on your phone

Why Do We Learn Data Science?



How does ChatGpt understand and answer your question?

Definition of Data Science

What is Data Science?

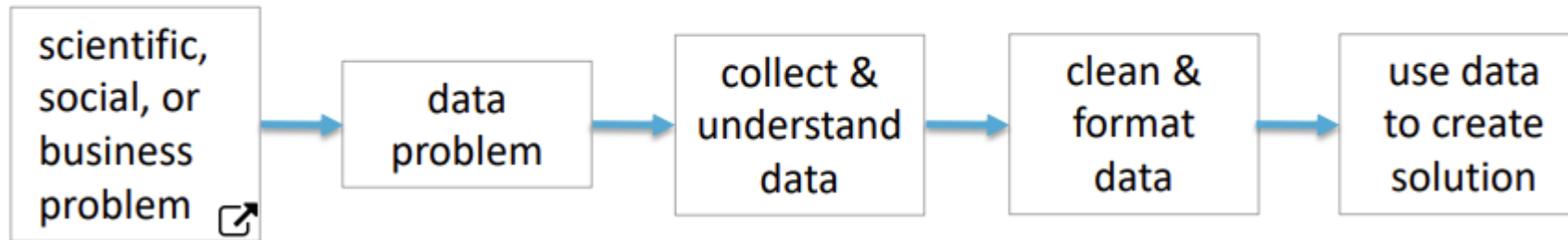
- Data science is an interdisciplinary field that uses algorithms, procedures, and processes to examine large amounts of data to uncover hidden patterns, generate insights, and direct decision making.
- Big data is a broad term for *data-sets* that are large and/or complex.
- Challenges include analysis, capture, data-curation, search, machine learning, statistics, sharing, storage, transfer, visualization, and information privacy.

How is it different from Data Analytics?

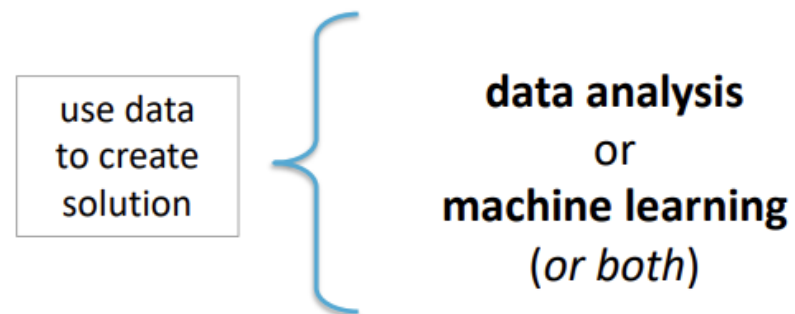
- Data Analytics is a sub-set of Data Science
- Data Analytics is mostly about using statistical and enterprise software to gain insights in the data
- Does not solve all components of data problems

Definition of Data Science

...solving problems with data...



...which step is most challenging?



Definition of Data Science

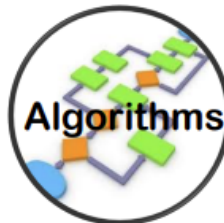
➤ **Data Analysis:** ...using data to discover useful information...



- **data:** anything you can *measure* or *record*



- **statistics:** summarize (and visualize) *main characteristics* of the data



- **algorithms:** apply algorithms to find *patterns* in the data

Definition of Data Science

➤ **Machine Learning:** ...creating and using models that learn from data...



- **data:** anything you can *measure* or *record*



- **model:** specification of a (mathematical) *relationship* between different variables



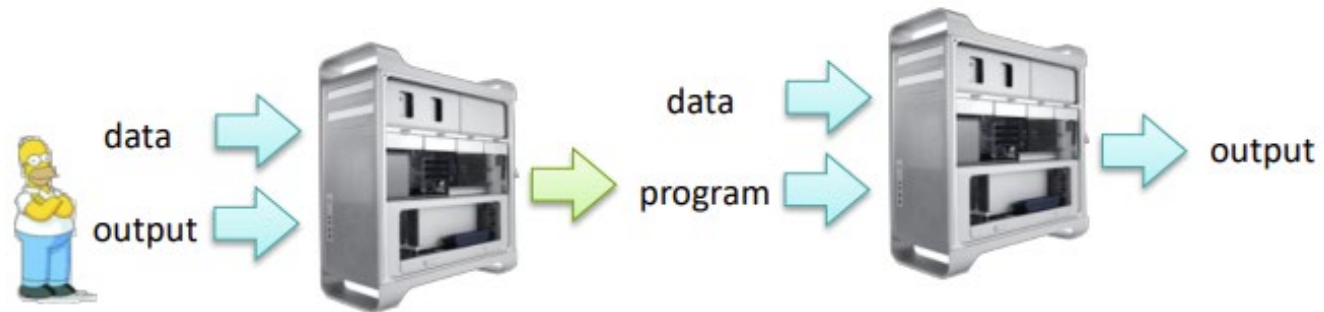
- **evaluation:** how well does the model *work?*

Definition of Data Science

➤ **Machine Learning:** ...creating and using models that learn from data...



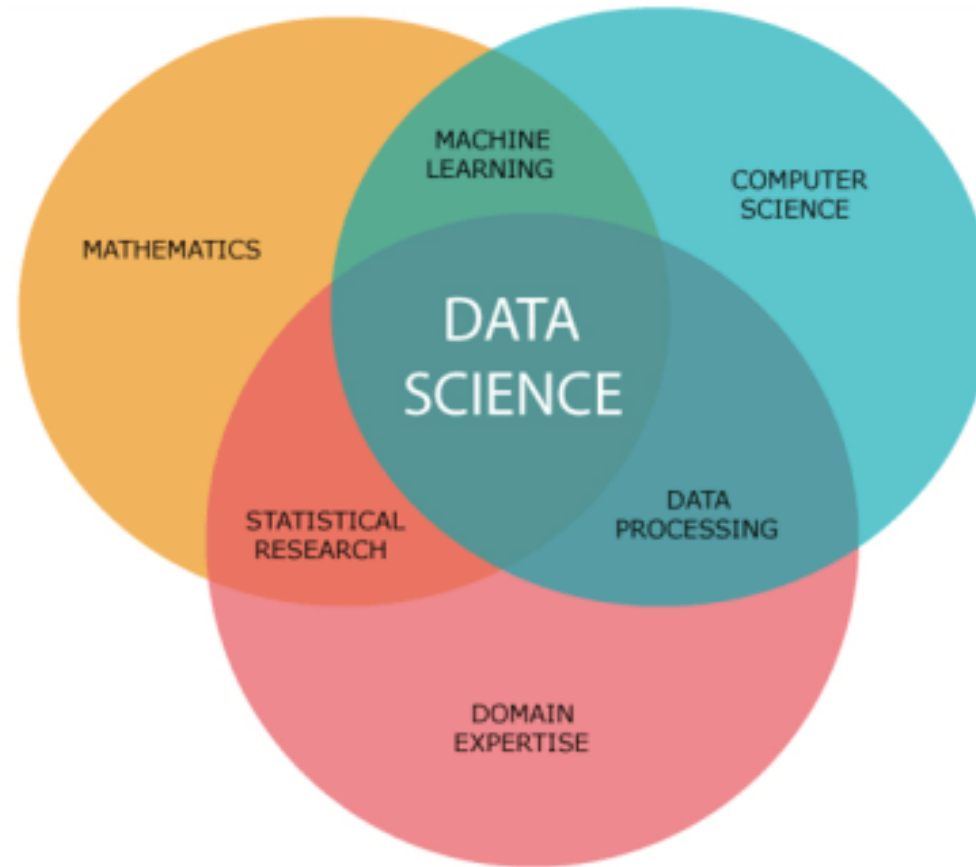
Traditional CS



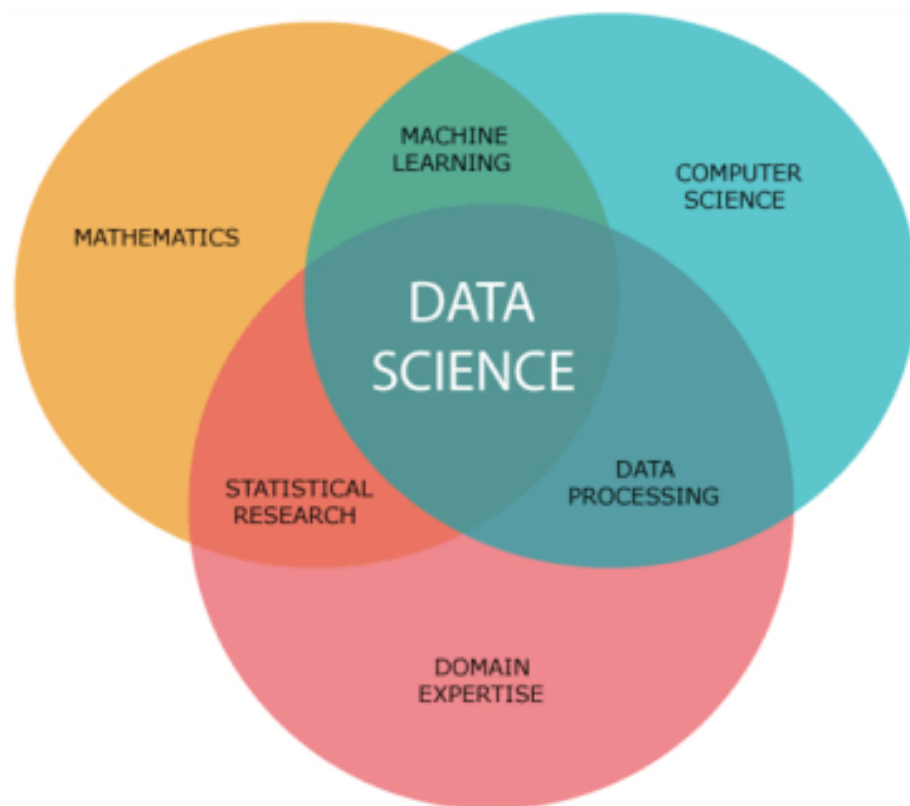
Machine Learning

Definition of Data Science

➤ **Goal of Data Science:** “Turn Data into Data Products”



Career Opportunities



UNC
GREENSBORO

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



Career Opportunities

Often responds within 1 day

Senior Analyst - Data Scientist (Hybrid - Washington, D.C.)

Risant Health
Washington, DC

\$142,917 - \$188,650 a year

[Easily apply](#)

Data Scientist / Physical Scientist

Bright Star Solutions LLC
Hybrid work in Washington, DC 20004
🚗 Federal Triangle

\$100,000 - \$140,000 a year

Full-time 401(k) Health insurance 401(k) matching

Paid time off Vision insurance Dental insurance

[Easily apply](#)

Data Scientist

Kaiser Aluminum
Spokane, WA 99216

\$85,000 - \$150,000 a year

Tuition reimbursement Health insurance 401(k) matching

Vision insurance Dental insurance Life insurance

[Easily apply](#)

Data Analyst

Delphinus Engineering
Bethesda, MD 20817

\$80,000 - \$85,000 a year

[Easily apply](#)

Data Scientist, SME

CGI Group, Inc.
Arlington, VA 22201 (Bluemont area)
🚗 Ballston Mu Metrorail Station

\$160,000 - \$180,000 a year

Tuition reimbursement Paid parental leave Parental leave

401(k) matching Paid time off Paid holidays

[View similar jobs with this employer →](#)

Data Analyst

SteerBridge
Cherry Point, NC

\$116,000 - \$126,000 a year

Full-time Health insurance 401(k) matching Paid time off

Vision insurance Dental insurance Life insurance

[Easily apply](#)



Indeed.com

Career Opportunities

- "The rise of Data Science needs will create roughly 11.5 million job openings by 2026"
 - *US Bureau of Labor Statistics*
- "By 2026, Data Scientists and Analysts will become the number one emerging role in the world."
 - *World Economic Forum*

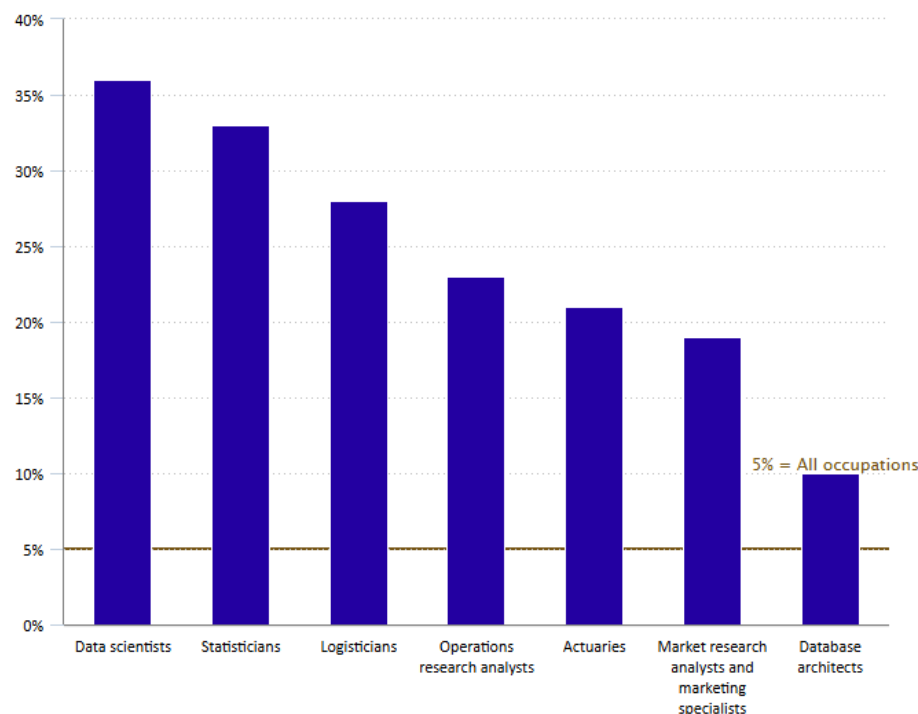
Data Science and Artificial Intelligence are amongst the **hottest fields** of the 21st century that will impact all segments of daily life by 2026, from transport and logistics to healthcare and customer service.

Career Opportunities

Employment growth in selected data-related occupations, projected 2021–31



U.S. BUREAU OF LABOR STATISTICS



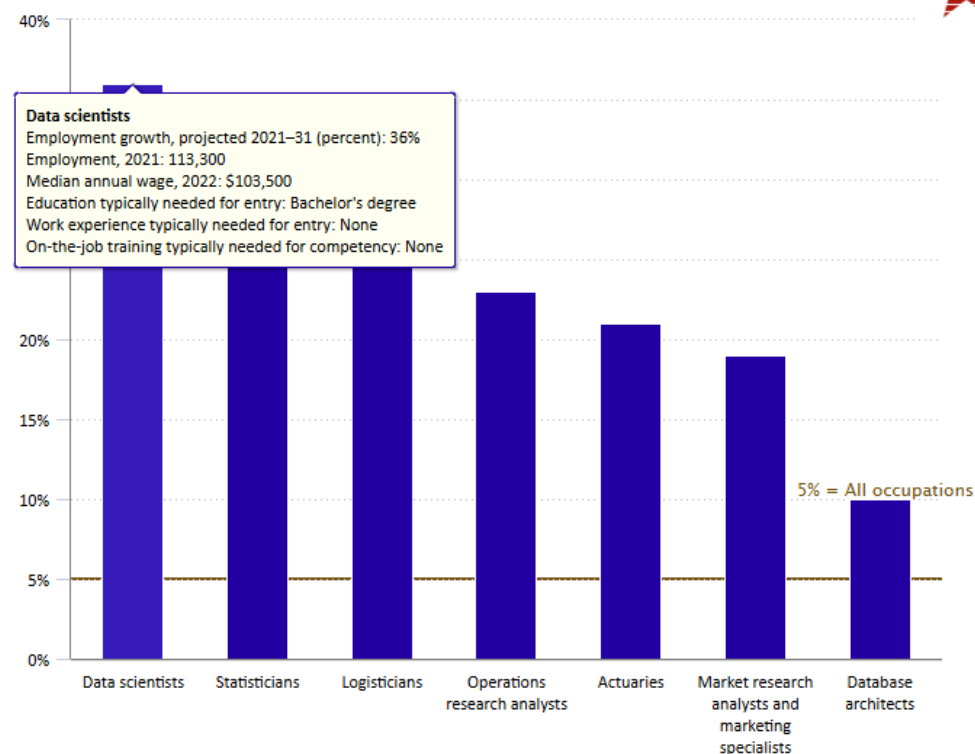
The occupation of data scientists is projected to have the fastest employment growth (grow 36% between 2021 and 2031) over the decade.

Career Opportunities

Employment growth in selected data-related occupations, projected 2021–31

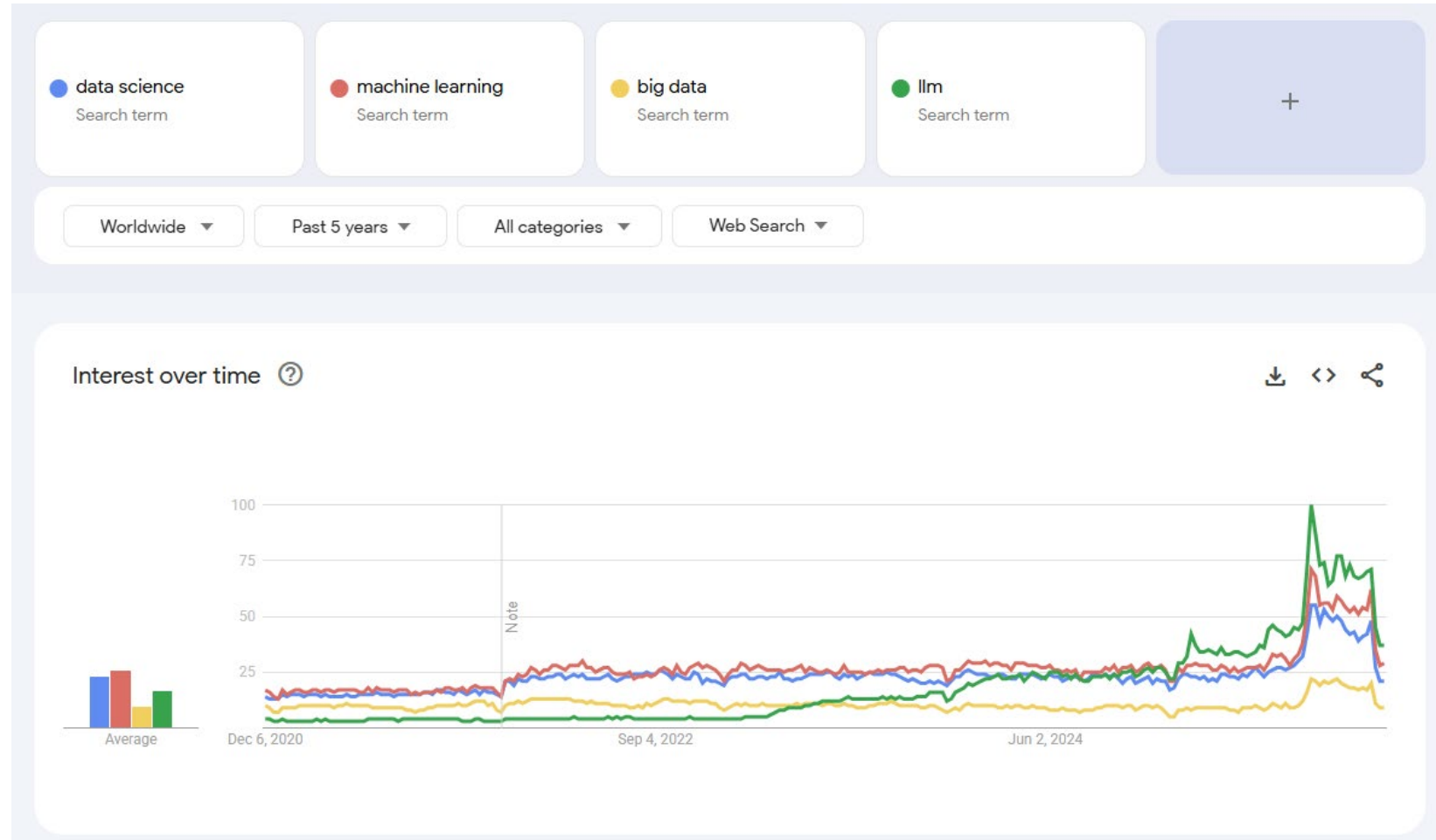


U.S. BUREAU OF LABOR STATISTICS



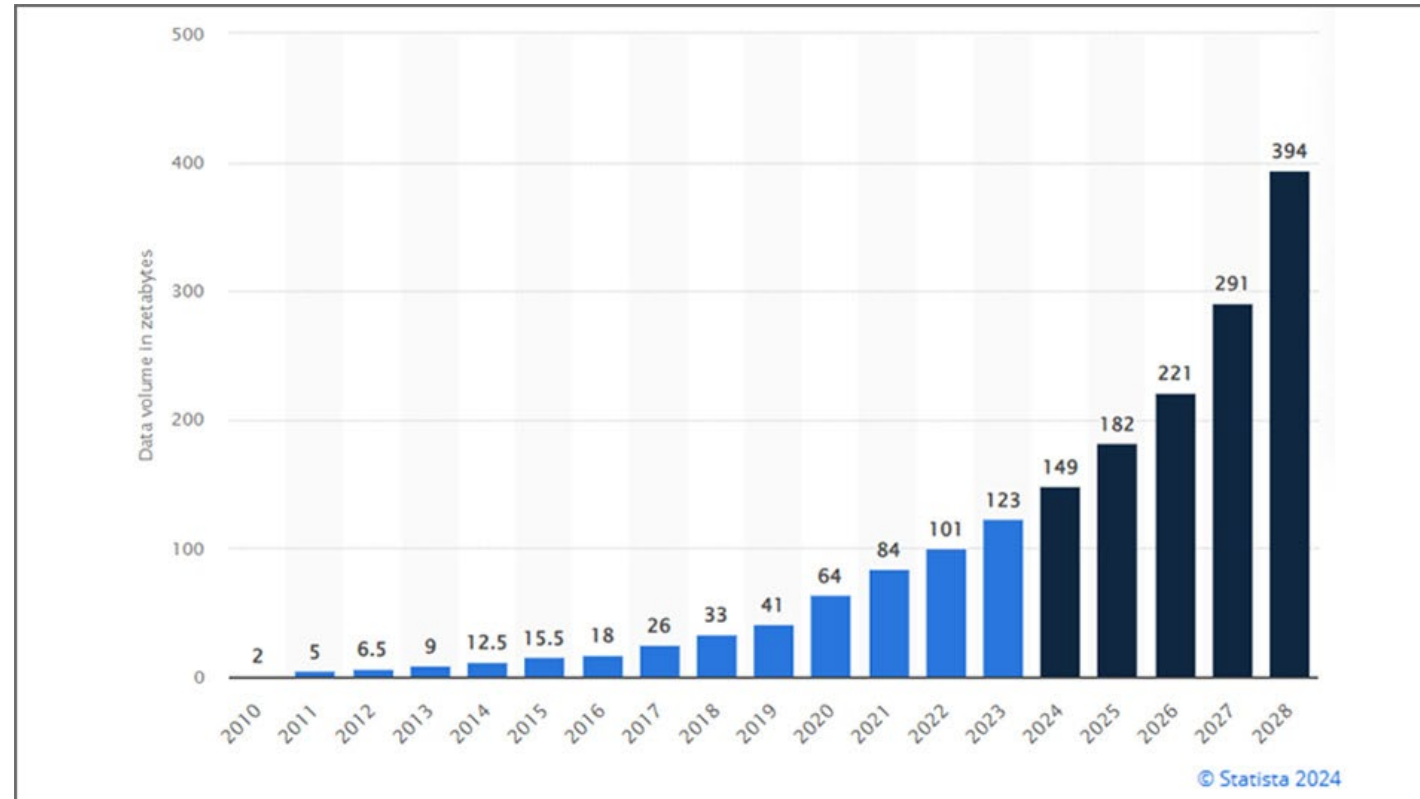
The occupation of data scientists is projected to have the fastest employment growth (grow 36% between 2021 and 2031) over the decade.

Career Opportunities



Google trends

Career Opportunities



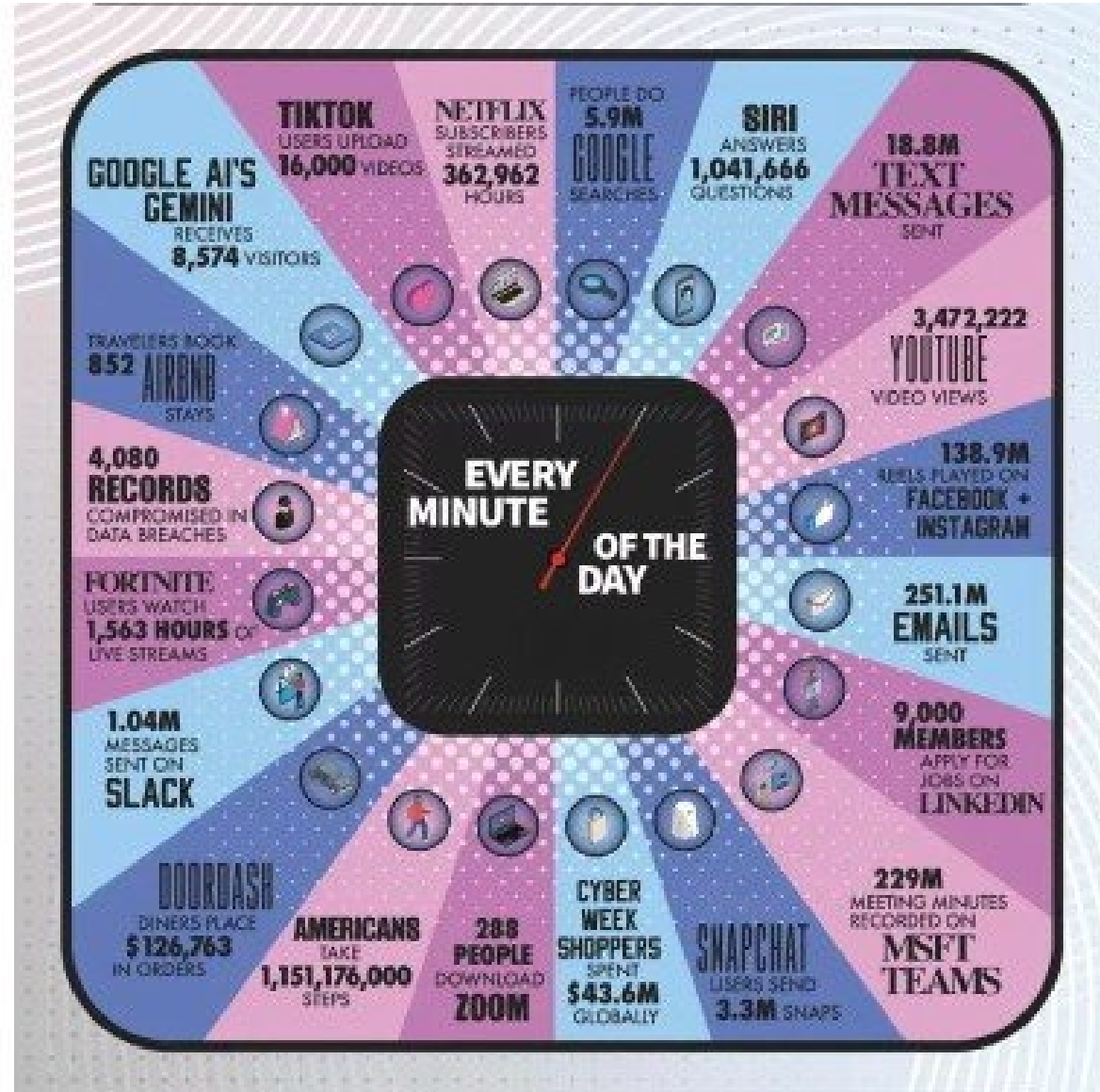
Global data volume is doubling every few years—
and there's no indication that trend is slowing



<https://www.studionetworksolutions.com/how-much-data-is-used-and-stored-in-the-world/>

Career Opportunities

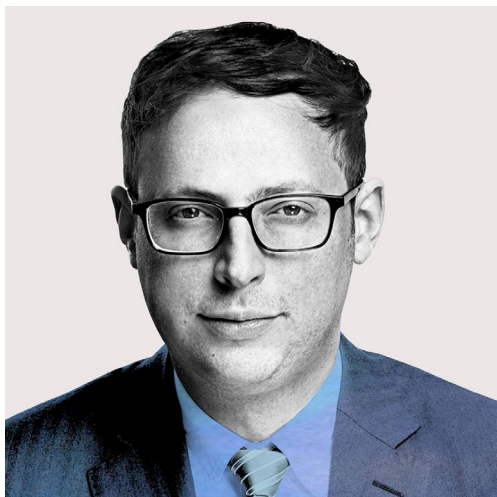
Every Minute on the Internet in 2025



Examples

FiveThirtyEight: FiveThirtyEight, also rendered as 538, was an American website that focused on opinion poll analysis, politics, economics, and sports blogging in the United States.

- Wikipedia



Nate Silver



FiveThirtyEight



Politics Sports Science & Health Economics Culture

Our Data

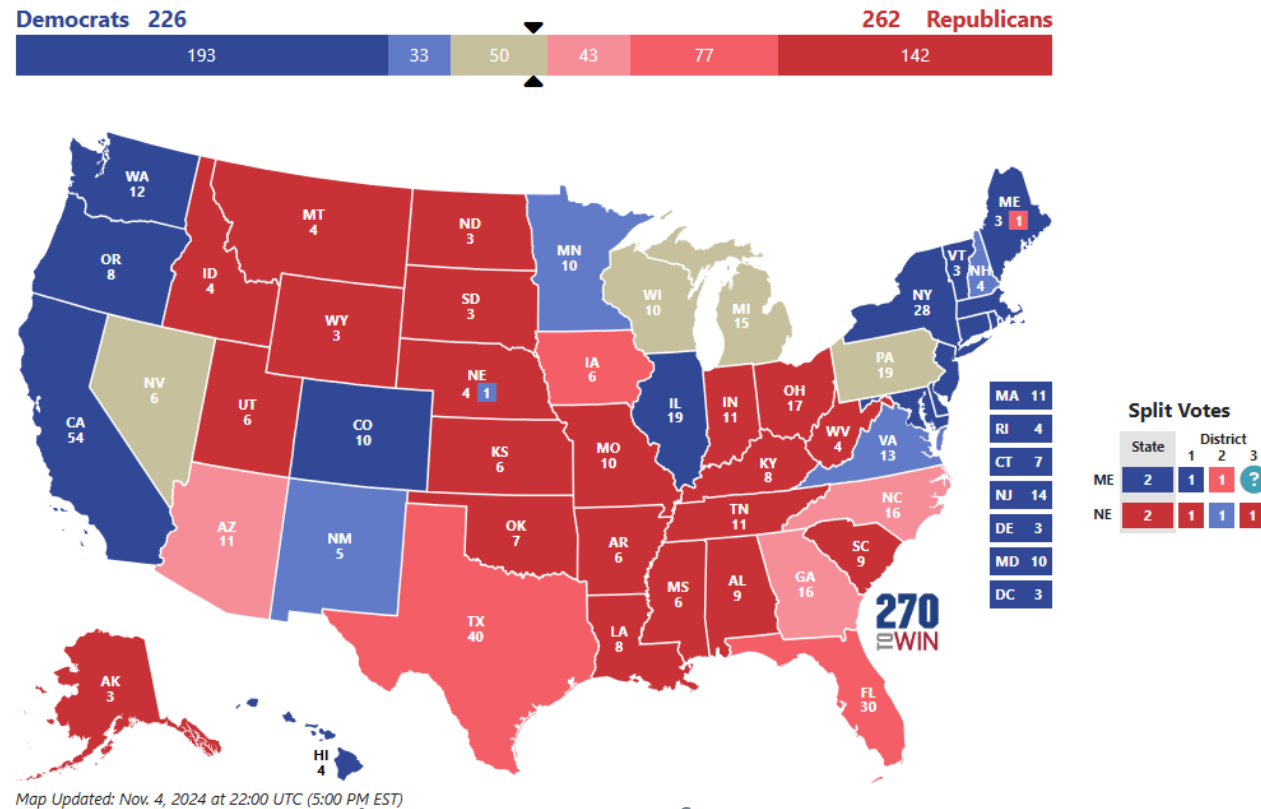
We're sharing the data and code behind some of our articles and graphics. We hope you'll use it to check our work and to create stories and visualizations of your own.

DATA SET	RELATED CONTENT	LAST UPDATED		
polls	Latest Polls	Sept. 26, 2023	info	↓
congress-generic-ballot	Do Voters Want Democrats Or Republicans In Congress?	Sept. 25, 2023	info	↓
mlb-elo	2023 MLB Predictions	June 21, 2023	info	↓
nba-forecasts	2022-23 NBA Predictions	June 14, 2023	info	↓
nba-raptor	The Best NBA Players, According To RAPTOR	June 14, 2023	info	↓
	Introducing RAPTOR, Our New Metric For The Modern NBA	Oct. 10, 2019		

<https://data.fivethirtyeight.com/>

Examples

FiveThirtyEight published a dynamic Presidential Forecast starting before Election Day, updating regularly until Election Day



Recommend reading: “How 538's 2024 presidential election forecast work”

(<https://abcnews.go.com/538/538s-2024-presidential-election-forecast-works/story?id=113068753>)

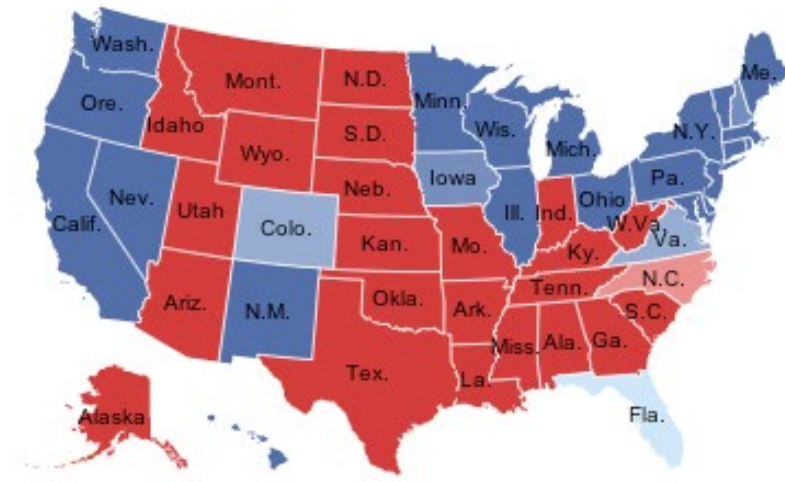


UNC
GREENSBORO

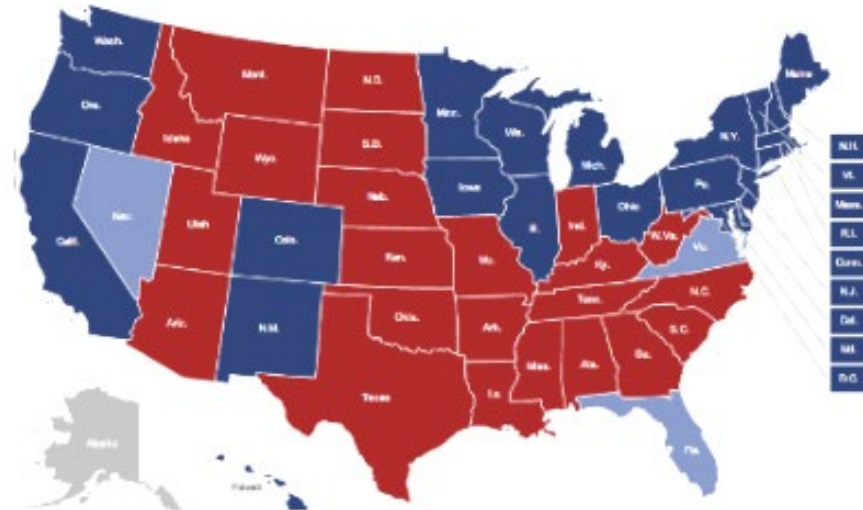
2024 Election Forecast from 538

<https://www.270towin.com/maps/538-forecast-2024-presidential-election>

Examples



Nate Silver's Map

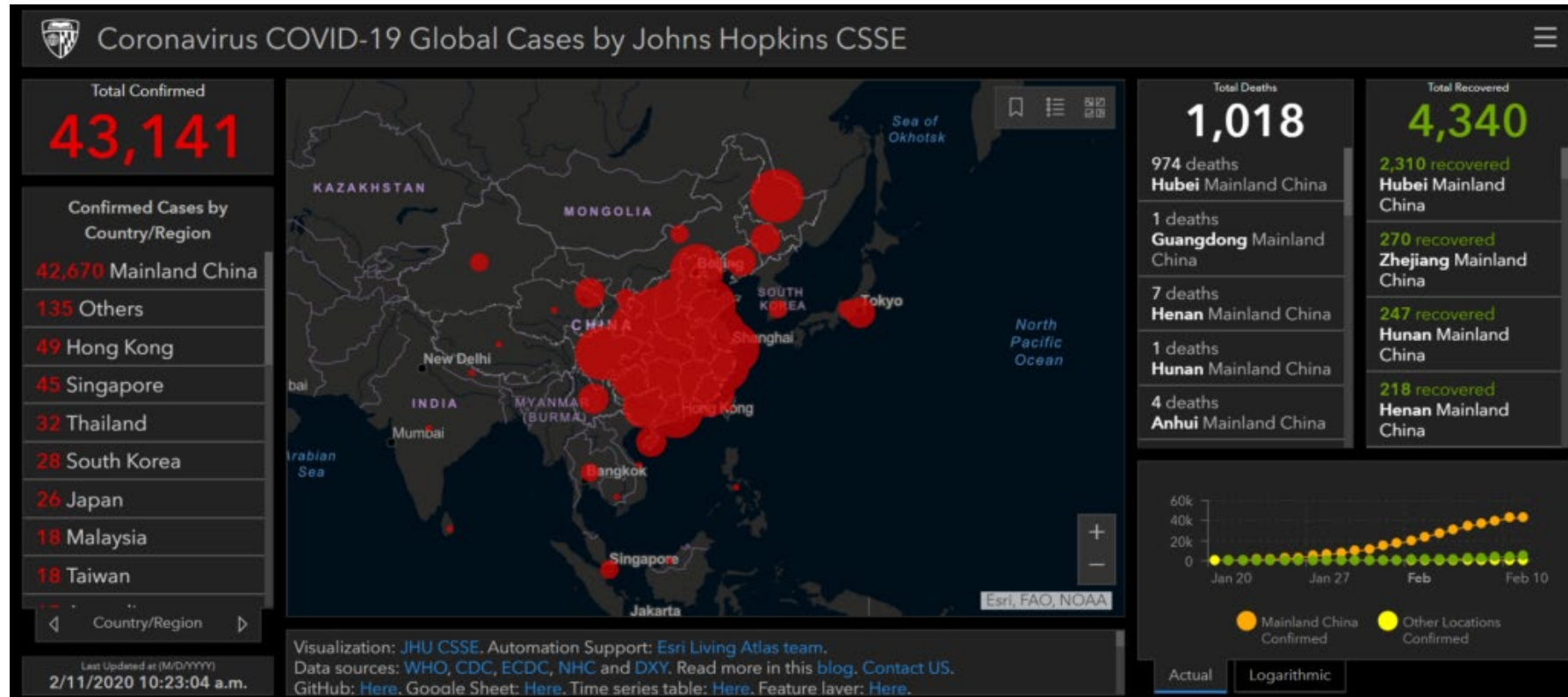


The Actual Map

2012 Nate Silver's prediction of election

Examples

John Hopkins University's Center for Systems Science and Engineering provides a real-time update on the new COVID-19 confirmed cases, deaths, and recoveries by country, province, and city.



Used by governments (CDC, WHO), airlines, and media, Helped scientists track spread and forecast hospital demand

Examples

Cyber-Security - 2013 Target Data

Data science models learn the “normal pattern” of network traffic or user actions, then they can flag anything unusual automatically in real time.

- login time & location
- number of failed login attempts
- file access patterns
- network packet characteristics

Target actually had an AI-powered anomaly detection system, but alerts were ignored.



NATIONAL

Hackers Stole 40 Million Credit, Debit Card Numbers From Target

DECEMBER 19, 2013 · 4:07 PM ET

HEARD ON [ALL THINGS CONSIDERED](#)

By [Elise Hu](#)

Hackers Stole 40 Million Credit, Debit Card Numbers From Target

▶ LISTEN · 3:48

+ PLAYLIST

...

Right in the middle of holiday shopping season, some 40 million credit and debit card numbers were stolen in a major breach of Target customer data. The thefts occurred in stores, not online. Target says it's working with a forensics company to investigate and prevent similar data thefts from occurring in the future. Security experts say one way to limit them is to switch from magnetic stripes on cards to embedded chips.

Examples

Cyber-Security – Data Privacy Leakage

To train data science models, companies often collect:

- Healthcare records
- Location data
- Financial data
- Browsing and communication logs
- ...

These datasets always contain **sensitive information.**

TIME

SECURITY & PRIVACY

How Target Knew a High School Girl Was Pregnant Before Her Parents Did

By Keith Wagstaff | Feb. 17, 2012

[f Share](#) [X Post](#) [in Share](#) [p Save](#) [Read Later](#)

In Charles Duhigg's new piece for the *New York Times*, a father finds himself in the uncomfortable position of having to apologize to a Target employee. Earlier he had stormed into a store near Minneapolis and complained to the manager that his daughter was receiving coupons for cribs and baby clothes in the mail.

Turns out Target knew his daughter better than he did. She really was pregnant.

(**MORE:** Google Wallet Hack Shows NFC Payments Still Aren't Secure)



Joe Raedle / Getty Images

