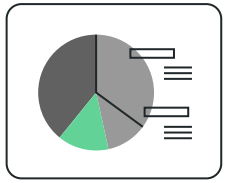


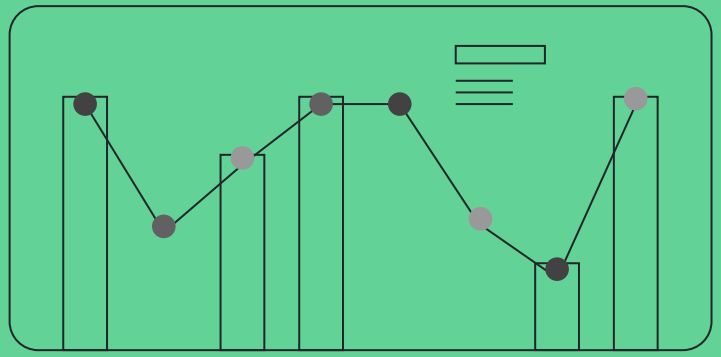
Introduction to Data Science

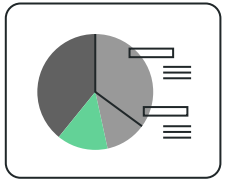
Henry Strecker

Learning Objectives

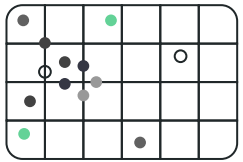
- Introduce the steps of the data science workflow
- Provide real examples of these steps in practice through my previous work
- Inspire you all to think about data and how it may relate to your work this quarter and beyond

What is Data Science?





Data science **combines** math and statistics, specialized programming, advanced analytics, artificial intelligence (AI) and machine learning with **specific subject matter expertise** to uncover actionable **insights hidden** in an organization's **data**. These insights can be used to **guide decision making and strategic planning** - IBM



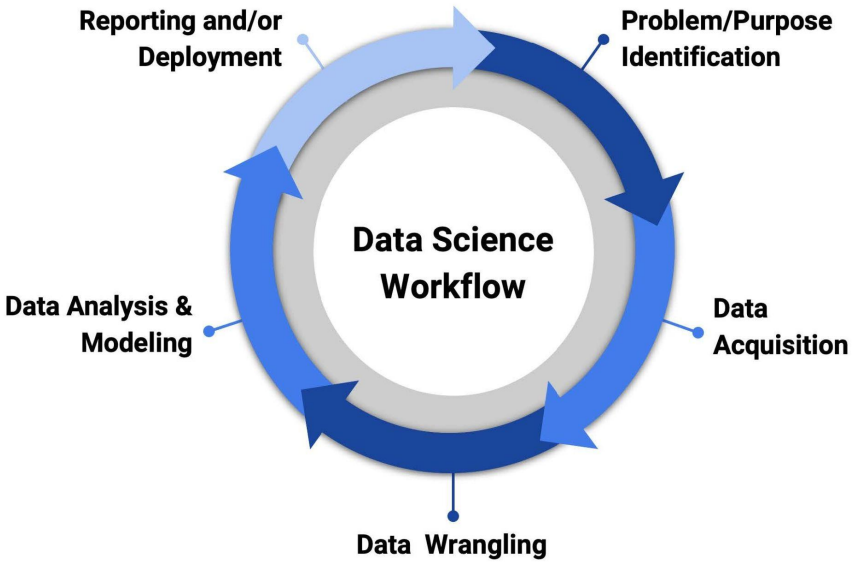
Importance and Usage

- Data storage has become extremely cheap, and most businesses have unprocessed data sitting around
- There is value in taking data and discovering or proving trends that aren't readily observable
- Most all businesses and organizations have dedicated data science teams working on products and projects

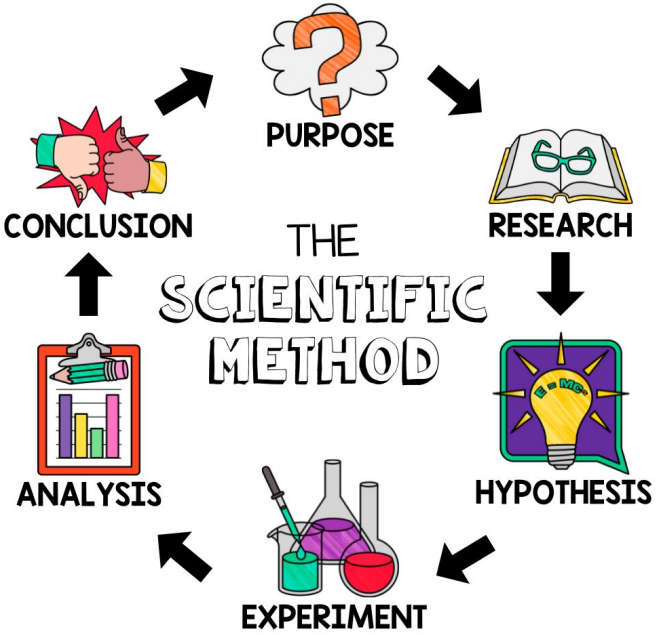
U.S. DEPARTMENT OF
ENERGY

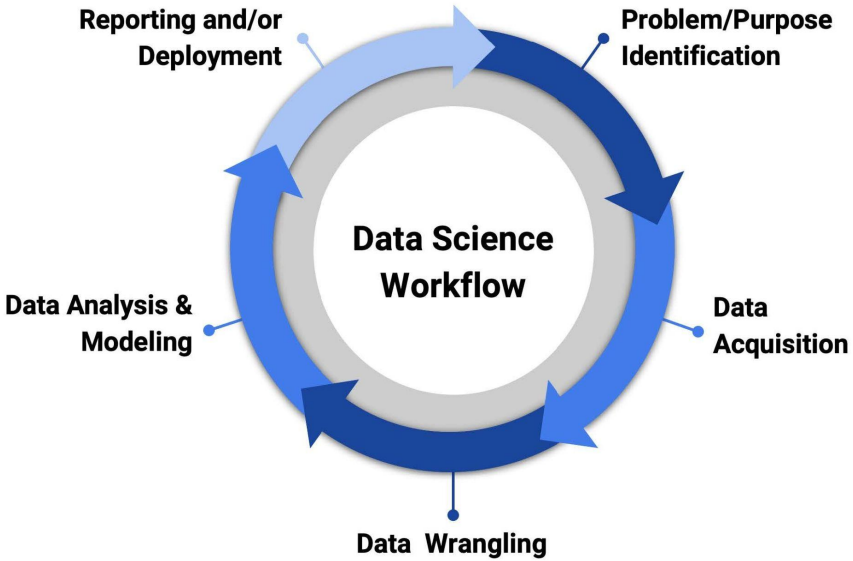
Data Science Workflow

- Sequential steps
- Iterative process, may need to move backwards if your first attempts don't work out
- This process should look somewhat familiar to you all...



Data Science Workflow





Problem Identification

What is the objective your work should achieve?

In data science, this purpose can take many forms:

- Discovering/quantifying trends
- Developing a model for explanation or forecasting
- Providing evidence for a report
- Creating a framework for automation
- Increasing efficiency

Data Science In Practice - My Bren Group Project

Client's mission: Help Federally Qualified Health Centers (FQHCs) in the US gain energy reliability via on-site solar panels and battery storage

Group project goal: Build a digital solution that will help our client to prioritize sites that are most vulnerable and would receive the greatest benefit

Data to achieve our goal: Power outages, social vulnerability, solar power potential, site energy consumption, and installation cost

Data Acquisition

	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02
Valiant	18.1	6	225.0	105	2.76	3.460	20.22
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00

- Seek quality sources
  - Government organizations, academic papers, public repositories
- Review metadata and documentation
  - Metadata is data about the data
  - Relays the units of measurement
  - Explains collection methods
- Assess the data's suitability for your purpose

Data Acquisition

- This article has credibility
  - Author works at a national lab (ORNL)
  - Released in *Scientific Data*
- Noted as the largest, most extensive dataset on this topic
  - For 2022 there's over 26 million rows
- Data was collected and combined from various other reliable sources
- Appears more useful and exhaustive than the alternatives

Data Descriptor | Open access | Published: 05 March 2024

A dataset of recorded electricity outages by United States county 2014–2022

Christa Brelsford, Sarah Tennille, Aaron Myers, Supriya Chinthavali, Varisara Tansakul, Matthew Denman, Mark Coletti, Joshua Grant, Sangkeun Lee, Karl Allen, Evelyn Johnson, Jonathan Huihui, Alec Hamaker, Scott Newby, Kyle Medlen, Dakotah Maguire, Chelsey Dunivan Stahl, Jessica Moehl, Daniel Redmon, Jibonananda Sanyal & Budhendra Bhaduri

Scientific Data **11**, Article number: 271 (2024) | Cite this article

Data Cleaning

This step normally accounts for **50-80%** of time spent on a project.

Most data is **not** collected and organized efficiently - it's inherently messy

Lots of effort is required to convert messy data into clean data that can be useful for your application

Data Cleaning - Exercise

Fips code	County	State	Customers Affected	Time of Outage
6083	Santa Barbara	California	292	2023-01-01 04:30:00
6083	Santa Barbara	California	292	2023-01-01 04:45:00
6083	Santa Barbara	California	32	2023-01-02 02:30:00
6083	Santa Barbara	California	32	2023-01-02 02:45:00
6083	Santa Barbara	California	32	2023-01-02 03:00:00
6083	Santa Barbara	California	169	2023-01-03 00:15:00
6083	Santa Barbara	California	107	2023-01-03 00:30:00
6083	Santa Barbara	California	107	2023-01-03 00:45:00
6083	Santa Barbara	California	107	2023-01-03 01:00:00
6083	Santa Barbara	California	116	2023-01-03 01:15:00
6083	Santa Barbara	California	0	2023-01-06 20:00:00
6083	Santa Barbara	California	0	2023-01-09 19:00:00

How is the data structured?

Is there anything weird to pay attention to?

Is this data ready to serve the purpose I need? What will it take to get there?

Data Cleaning - Exercise Results

Fips code	County	State	Customers Affected	Time of Outage
6083	Santa Barbara	California	292	2023-01-01 04:30:00
6083	Santa Barbara	California	292	2023-01-01 04:45:00
6083	Santa Barbara	California	32	2023-01-02 02:30:00
6083	Santa Barbara	California	32	2023-01-02 02:45:00
6083	Santa Barbara	California	32	2023-01-02 03:00:00
6083	Santa Barbara	California	169	2023-01-03 00:15:00
6083	Santa Barbara	California	107	2023-01-03 00:30:00
6083	Santa Barbara	California	107	2023-01-03 00:45:00
6083	Santa Barbara	California	107	2023-01-03 01:00:00
6083	Santa Barbara	California	116	2023-01-03 01:15:00
6083	Santa Barbara	California	0	2023-01-06 20:00:00
6083	Santa Barbara	California	0	2023-01-09 19:00:00

Observations:

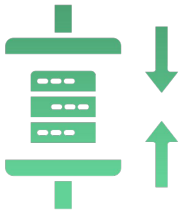
- One power outage can span multiple rows
- Need to split outage events when the number of customers changes
- Some outages don't have any customers without power?

Data Cleaning - Problem Solving & Coding

- Reached out to author of the paper about outages with zero affected customers

- Wrote about 200 lines of code using R that summarizes rows into outage events and then further into county summaries

- This code summarized our data from 26 million rows to about 3,000 - one for each county in the US



Data Cleaning - Results

The runtime for the code on all 7GB of data was about 5 hours

Results: Consolidated hundreds of thousands of rows for each county into a single summarized row

fips_code	county	state	year	total_outages	total_customers_affected	total_customer_minutes	total_customers
6079	San Luis Obispo	California	2023	8409	2727797	58694562	125727
6081	San Mateo	California	2023	15393	26550401	319184435	328408
6083	Santa Barbara	California	2023	7827	2195433	66091357	192080
6085	Santa Clara	California	2023	18792	32393447	378741095	844715
6087	Santa Cruz	California	2023	10101	15549741	258758550	118299
6089	Shasta	California	2023	5331	1491708	44364209	81282
6091	Sierra	California	2023	465	342512	26030675	4203
6093	Siskiyou	California	2023	4490	518625	18154472	26605
6095	Solano	California	2023	8649	1981809	50139585	187690
6097	Sonoma	California	2023	10267	5897490	107716890	214127
6099	Stanislaus	California	2023	2162	157623	6900760	219156
6101	Sutter	California	2023	3224	601593	16710795	42323

Data Analysis & Visualization

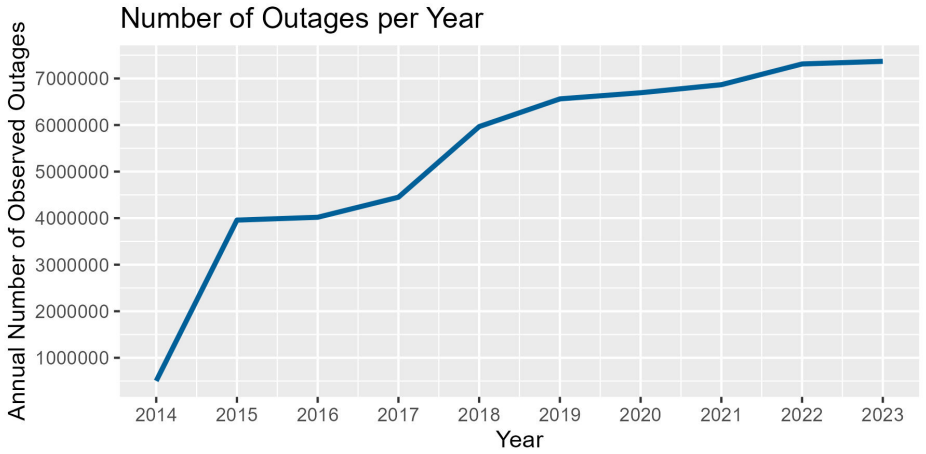
Now that the data is ready, it's time to search for insights!

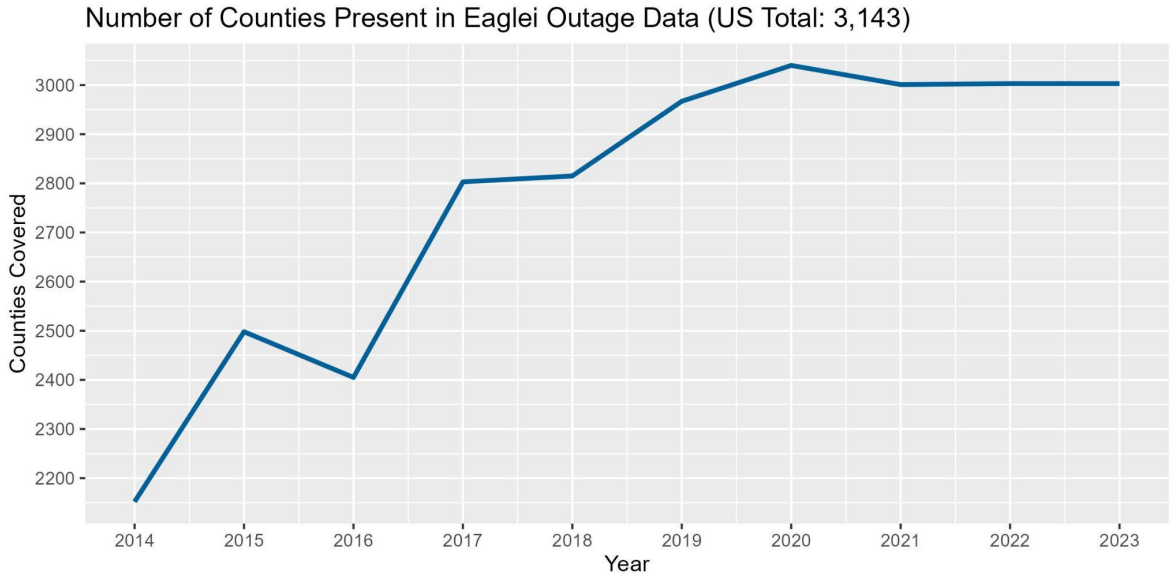
Analysis can come in many forms: plot generation, forecasting, discovering trends, etc.

Plots are normally the best way to convey results to a broader audience

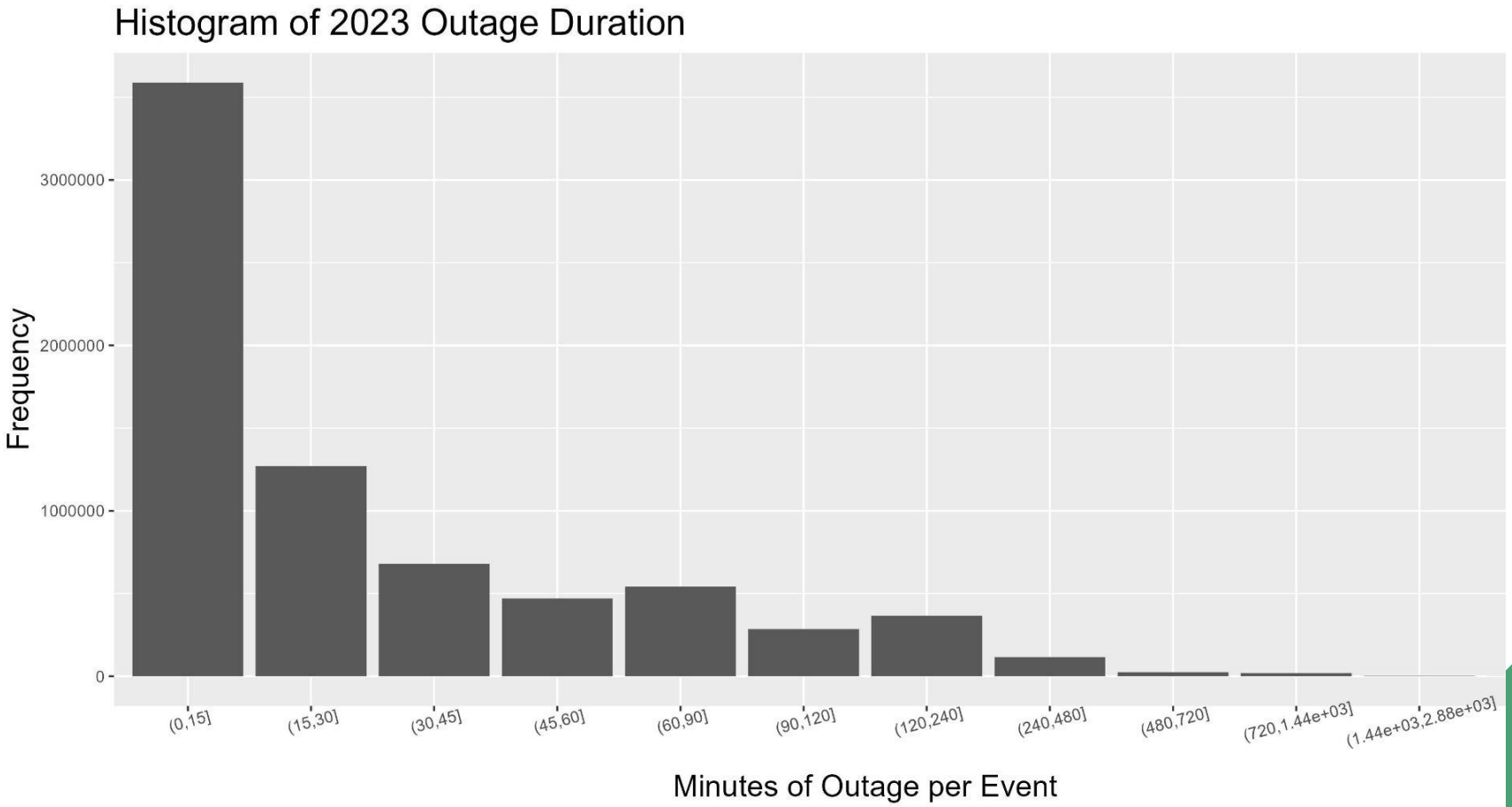
What are the components that make a good plot?

Assessing Power Outage Data Quality





Year	Total Sites Covered	Percent of Sites Covered
2016	9697	82.83%
2017	11013	94.07%
2018	10986	93.84%
2019	11303	96.55%
2020	11343	96.89%
2021	11376	97.17%
2022	11366	97.09%
2023	11375	97.16%


Histogram of 2023 Outage Duration
Frequency
3000000
2000000
1000000
0
(0,15]
(15,30]
(30,45]
(45,60]
(60,90]
(90,120]
(120,240]
(240,480]
(480,720]
(720,1.44e+03]
(1.44e+03,2.88e+03]
Minutes of Outage per Event

Documentation & Reporting

Goal is to consolidate and synthesize your work into a well organized deliverable

This should serve to explain the methods of your work, ensure reproducibility, and convey findings

Depending on the project's audience, there are various forms this may take:

- Slideshow presentation
- Internal documentation
- External reporting or publication

Lesson Summary

- The data science workflow is comparable to the scientific method
- Understanding your data before starting to work is crucial
- Cleaning data is arduous, but essential for generating results
- Visualizations are important to convey messages to an audience
- Documentation ensures transparency and reproducibility

Today's Lecture Free-write (Lecture 4.9)

- What are three data science concepts you learned from today's lecture? Provide explanations

Imagine that you are going to incorporate data science into your final project for this course:

- Where would you start to search for data relating to your topic?
- How could data contribute to the message or goal of your project?

Want more?

- UCSB Data Science Club
  - Normally have intro events at the beginning of each quarter, no experience needed to join!
- ENVS 193SW - Intro to Collecting, Wrangling, and Exploring Water Data (Offered Fall Quarter)
- Datacamp/Coursera - online courses in a broad variety of topics
- Bren School Master of Environmental Data Science (MEDS)
- Reach out to me to talk more about data science!