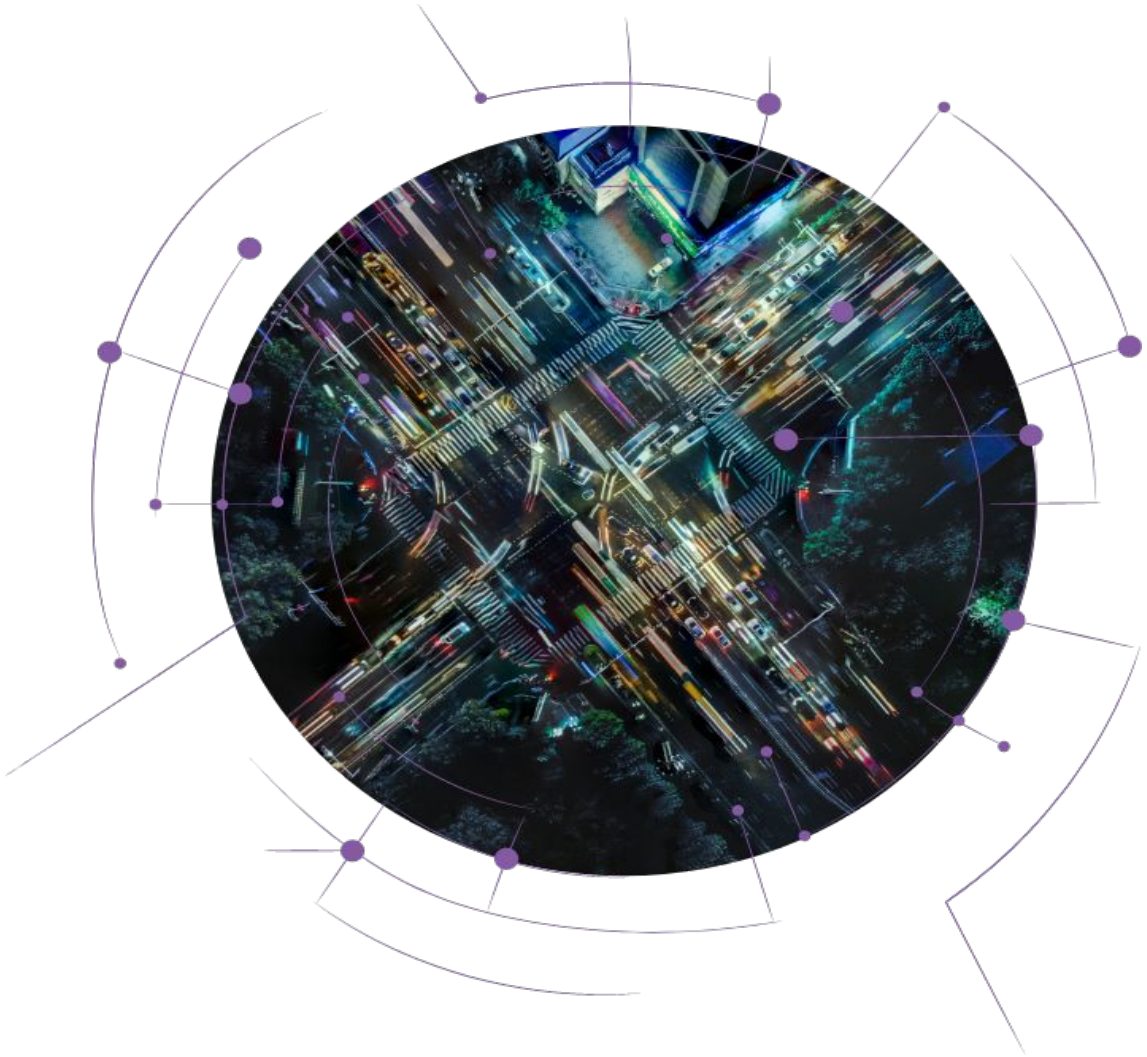
DATASOCIETY:

Data Literacy for All



About the course

- Instructor introduction

- Schedule:
  - 4 sessions
  - 11 am – 2 pm
  - 1 or 2 short breaks each session
  - Q and A during the last 30 minutes of class

Class materials

You should have received the following materials:

- Slides
- Participant guide
  - Needed during class
  - Contains discussion prompts, activities, a data science glossary, information about popular data science tools, and more!

Agenda

- **Day 1**
  - Fundamentals of data
  - Data analytics overview
  - Data governance
  - Data teams
  - Data tools
- **Day 2**
  - Data-driven cultures
  - Putting together a project
  - Foundational data science methods
- **Day 3**
  - Foundational ML methods
  - Advanced ML methods
- **Day 4**
  - AI methods
  - Refining a data project
  - Intro to data visualization
  - Best practices in data viz

Chat question

In the chat, share the following:

- Your name
- What comes to mind when you think of **“Data and AI.”** Whether it’s a word, phrase, tool, or concept, what do you associate with it?

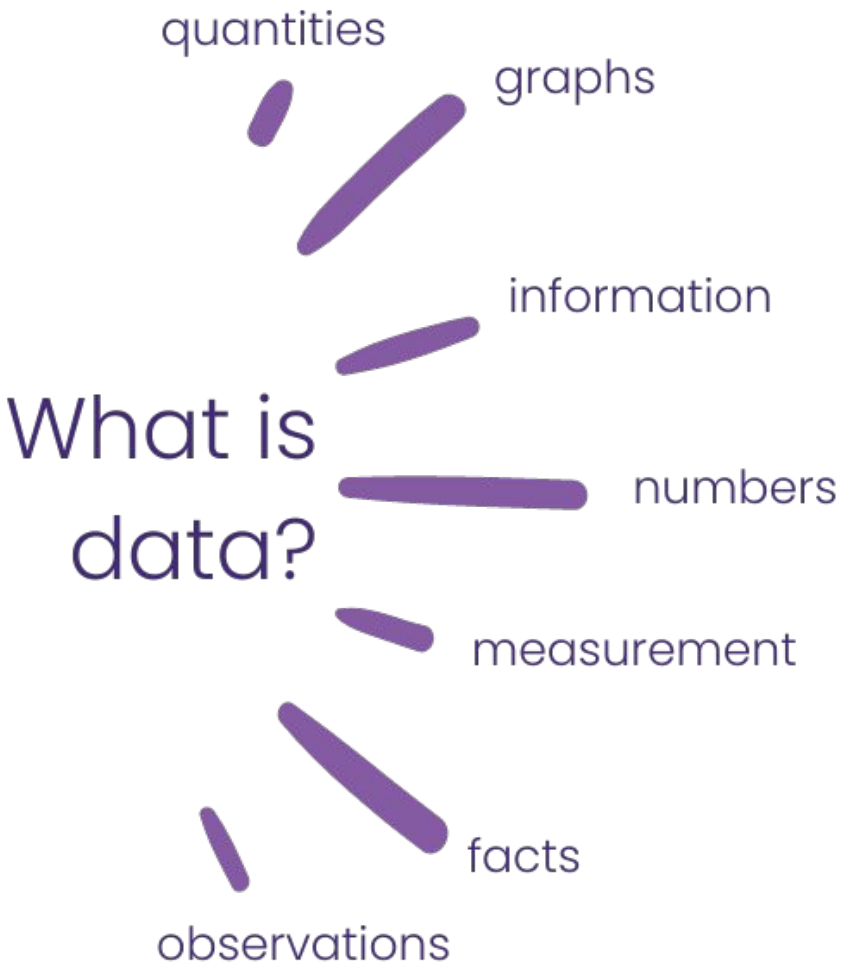
Agenda

Day 1

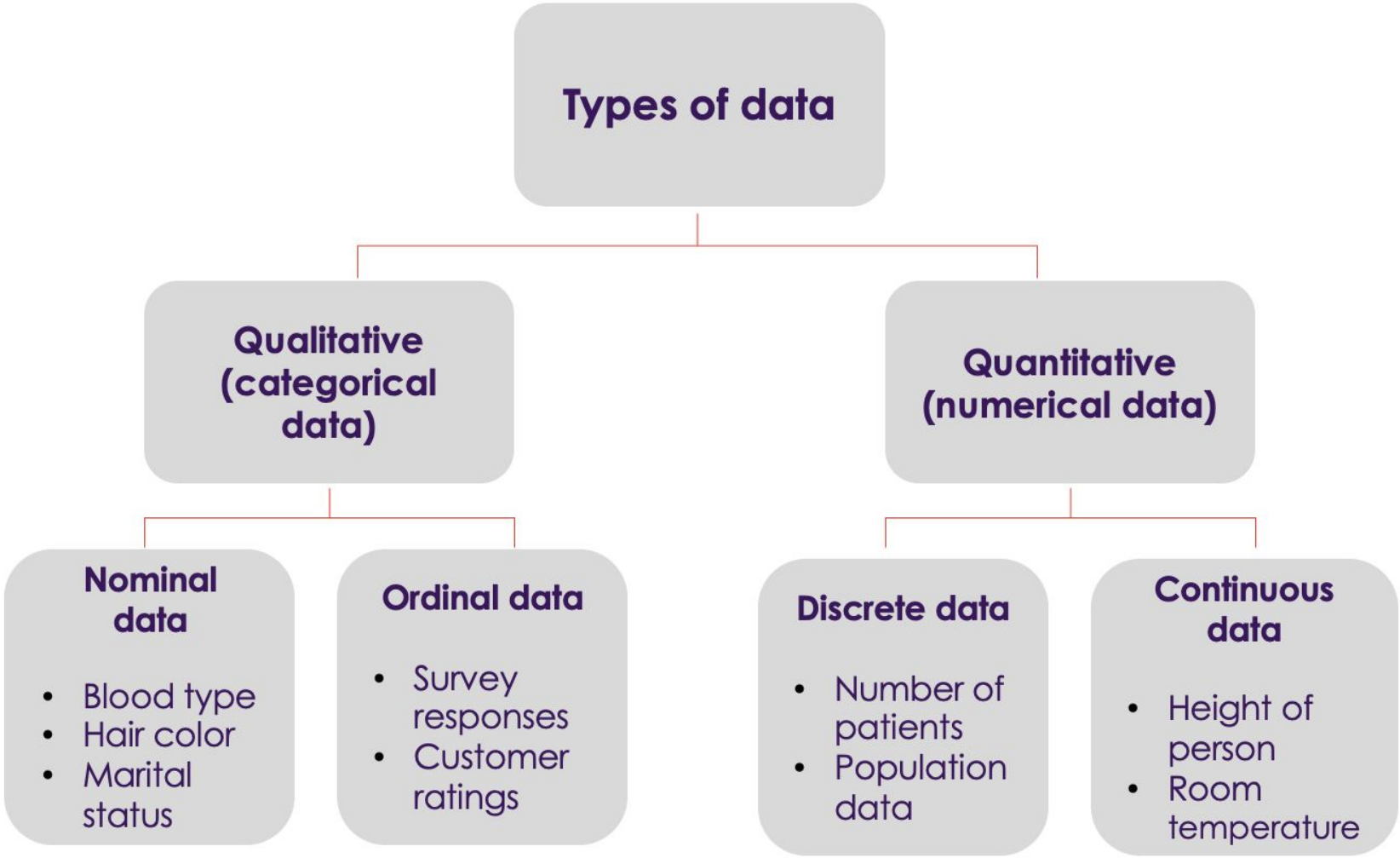
- **Fundamentals of data**
- Data analytics overview
- Data governance
- Data teams
- Data tools

What is data?

- **Data** is everywhere. So much so that there's possibly no field that doesn't deal with data.
- It is **information, facts, or numbers** collected to be examined, considered, and used to help decision-making.
- The data universe continues to grow relentlessly.

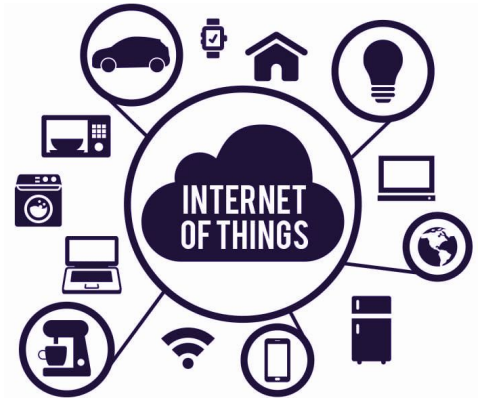


Types of data



Data formats

Unstructured



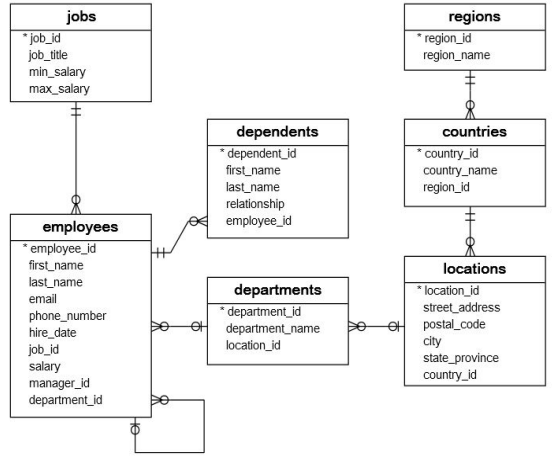
Semi/Quasi -structured

```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```

Sep 17 02:33:08.536 [debug] connection_edge_process_relay_cell(): Now seen 1802 relay cells here (command 2, stream 5845). Sep 17 02:33:08.536 [debug] connection_edge_process_relay_cell(): circ deliver_window now 933.

Structured

y1	x1	x2	x3



Quick quiz

Question #1:

- **What data format would emails fall under?**
    a. **Structured data**
    b. **Semi-structured data**
    c. **Unstructured data**

Quick quiz

Question #2:

- **What data format would research papers fall under?**
    a. **Structured data**
    b. **Semi-structured data**
    c. **Unstructured data**

Sources of data

Internal

- Information Resources Management (IT)
- Global Talent Management systems(HR)
- Comptroller and Global Financial Services systems (Finance)
- Supply chain records
- Service records

External

- Publicly-accessible APIs
  - e.g., api.data.gov
- Other open data sources
  - e.g., data.worldbank.org
- Large businesses are increasingly giving people access to their data
  - e.g., Expedia

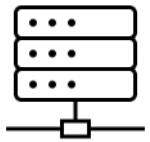
Chat question

What internal and external data do you interact with at work?

What is big data?

"Big data" refers to ***a large volume of data*** *that can be* ***mined for information*** *and* ***used in machine learning*** *projects and other analytics applications.*

Characteristics of big data



High volume

Data size is described in interabytes, petabytes, and exabytes!

High velocity

Big data flows from sources at a rapid and continuous pace

High variety

Big data comes in different formats from heterogeneous sources

Heterogeneous data

- Big data technology offers incredible opportunities for accelerating discoveries and innovations, but if not used responsibly, it poses significant risks.
- It is important to ensure there is a broad spectrum of data, i.e., different kinds of objects are represented.
- Data that reflects the **a variety of formats, sources, experiences and levels of details** can help fill in the gaps and provide a more rounded and accurate picture.

Image source: Voneff at The Noun Project

Agenda

Day 1

- Fundamentals of data
- **Data analytics overview**
- Data governance
- Data teams
- Data tools

Chat question

Where do you like to grocery shop? And do you like the way they have it set up?

Turning data into action

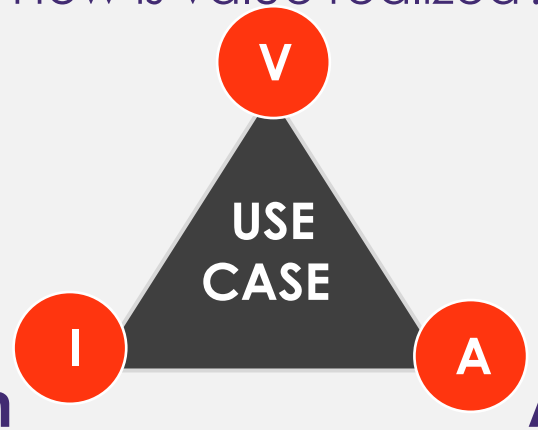
- Ever noticed how easy it is to find what you need in a massive supermarket?
- That's no accident!
- These stores use data and analytics to design their layouts to offer a smoother shopping experience to their customers.

Watch the video: *How Walmart is using data*

Framework for a data-driven approach

Value

What is the question, business problem or target outcome? How is value realized?



Information

What data or data sources are involved?

Analysis

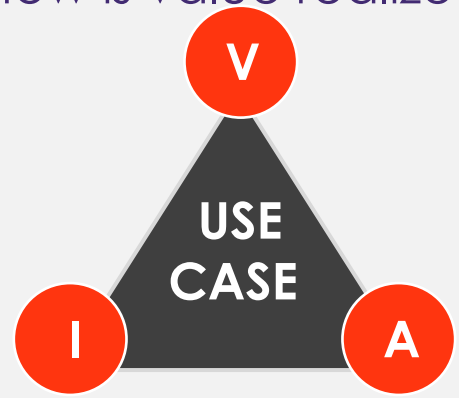
What analytical or data science methods are applied to the data?

1 Value: Defining the business problem

- Defining the problem is **articulating the business objective** and understanding the **key requirements and constraints.**
- A well-defined problem sets the stage for a successful data project that delivers value to the organization.

Value

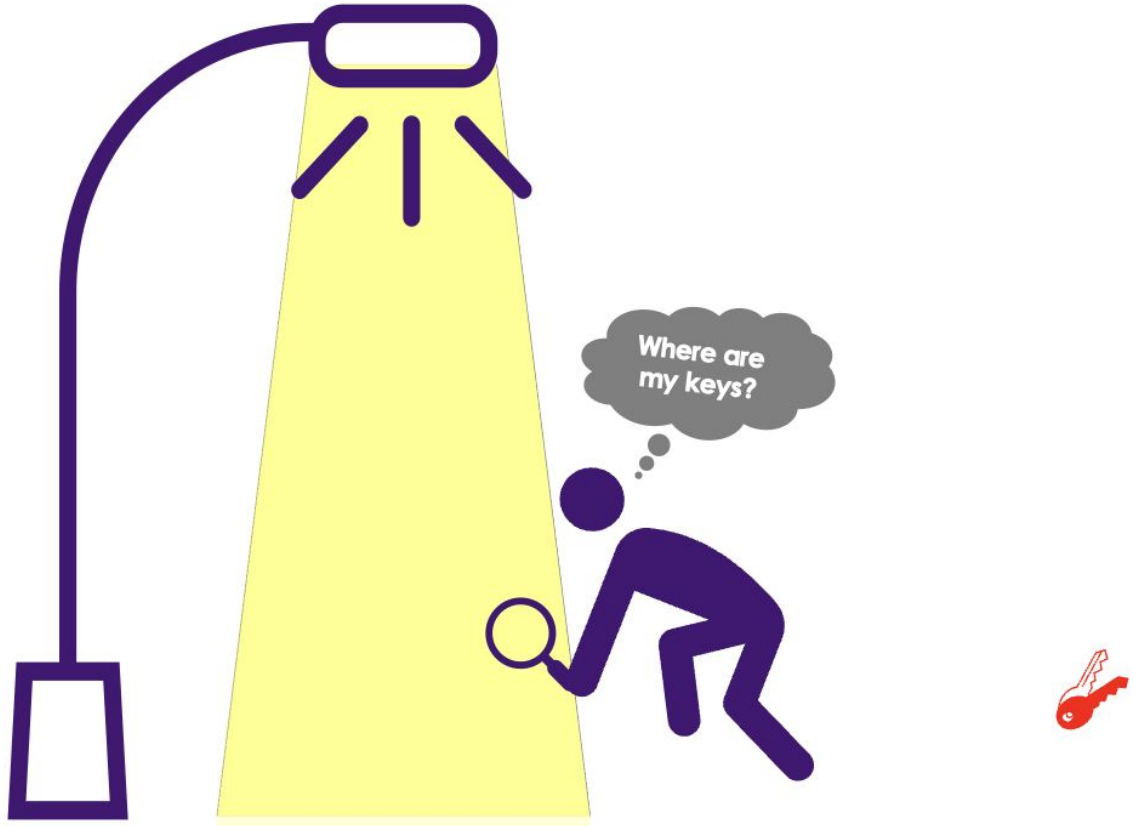
What is the question, business problem or target outcome? How is value realized?



The Data Lodge® | ISL Institute

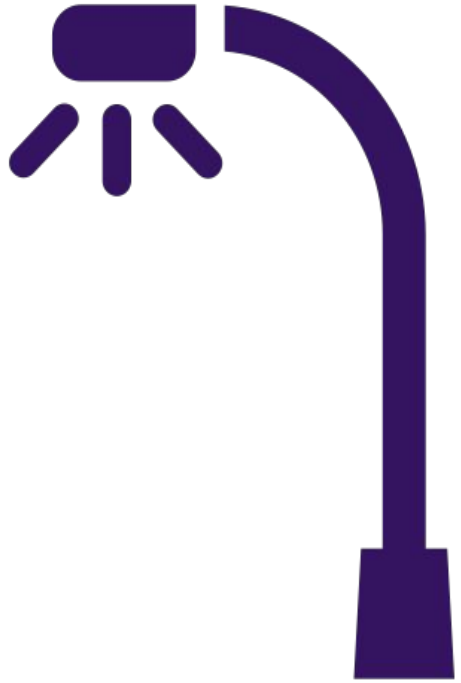
The streetlight effect

The streetlight effect is a term used to describe a situation where people look for information where is it most convenient.



The streetlight effect (cont'd)

- To gain a more comprehensive understanding, we should ask ourselves:
  1. What are you searching for?
  2. Where are you searching?
  3. What is your measure of success?
  4. In what ways could your own assumptions influence your data-gathering process?

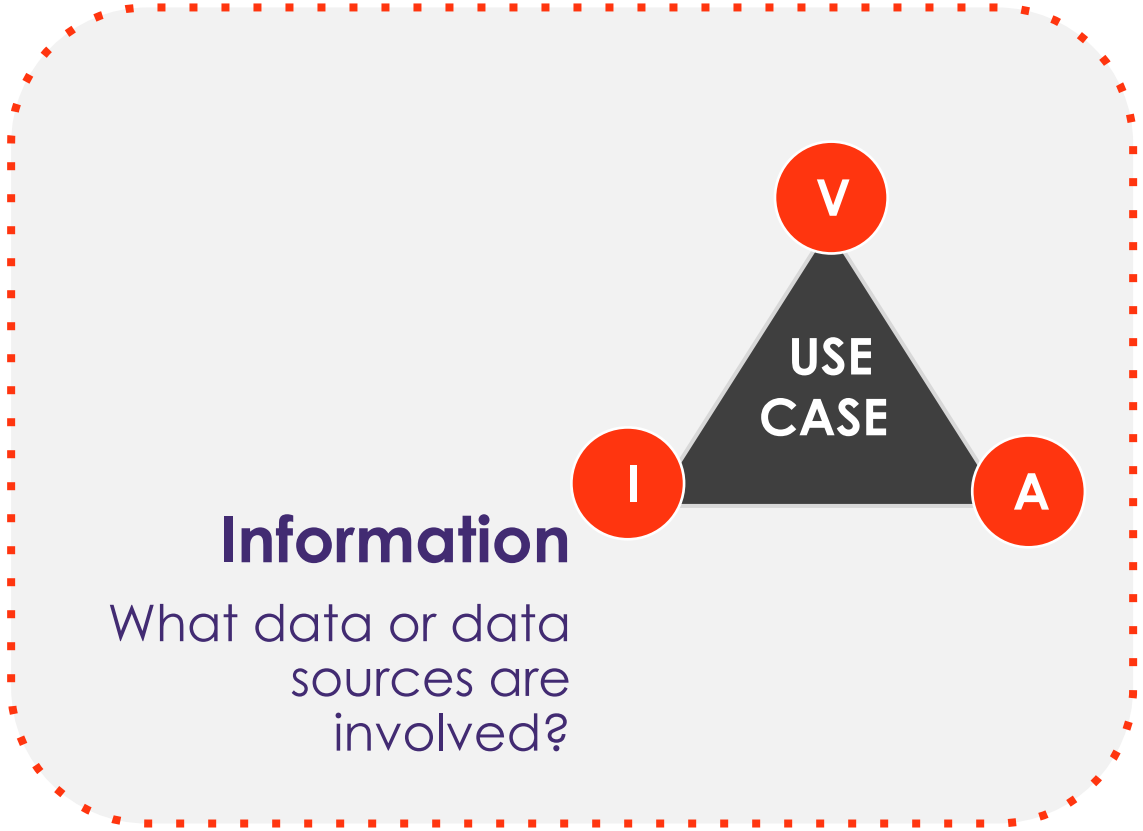


Chat question

How can you look beyond the obvious symptoms to identify the underlying causes of a problem?

2 Information: Finding the right data

- The next step is to identify the data that can help you answer your questions or solve your business problem.
- The data required can be primary data *(collected by yourself or your team or your organization)* or secondary data *(obtained from existing external sources.)*
- It is essential to consider the reliability, validity, and relevance of each data source before using it.

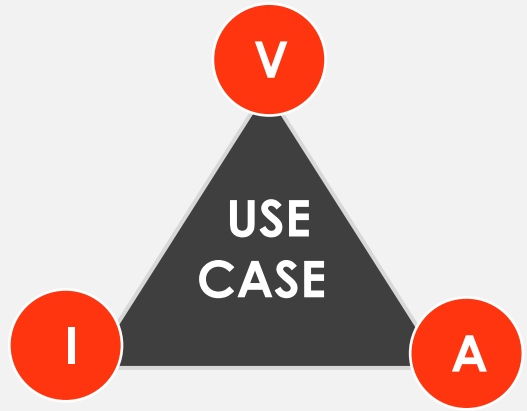


The Data Lodge® | ISL Institute

3 Analysis

- **Data analysis** focuses on processing and performing statistical analysis on existing datasets.

- Analysts **capture, process, and organize data to uncover actionable insights** for current problems and establish the best way to present this data.

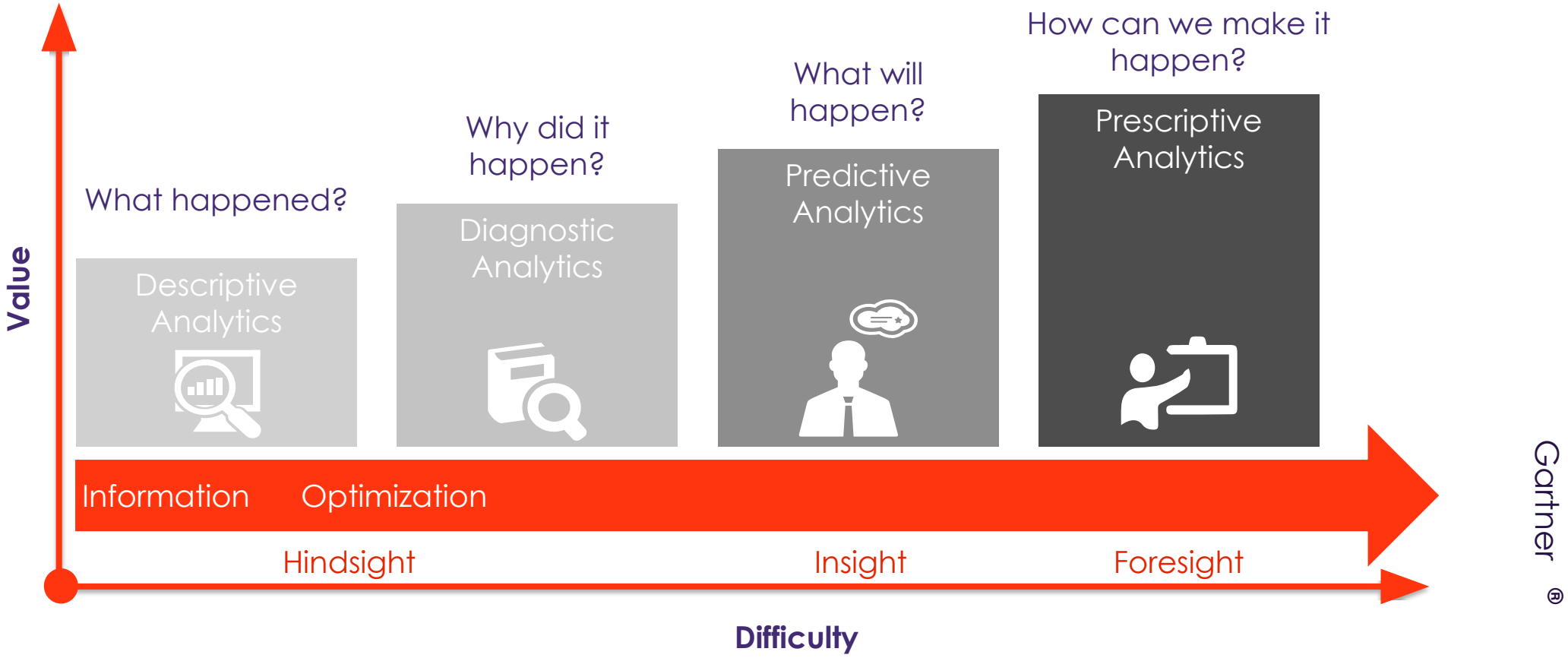


Analysis

What analytical or data science methods are applied to the data?

The Data Lodge® | ISL Institute

Data analytics maturity model



Model revisited

Chat question

Have you come across any of these data analytics terms?

Stage 1: Descriptive analytics

What questions does it answer?	What has happened in the past?
How valuable is it?	Provides some value, but doesn't provide causation or prediction
How labor intensive is it?	Easy to deploy provided you have the right data

Stage 2: Diagnostic analytics

What questions does it answer?	Why did something happen in the past?
How valuable is it?	Provides insights into a particular problem, and can help you identify some root causes for past trends and behaviors
How labor intensive is it?	Requires detailed data, but doesn't have to be overly intensive

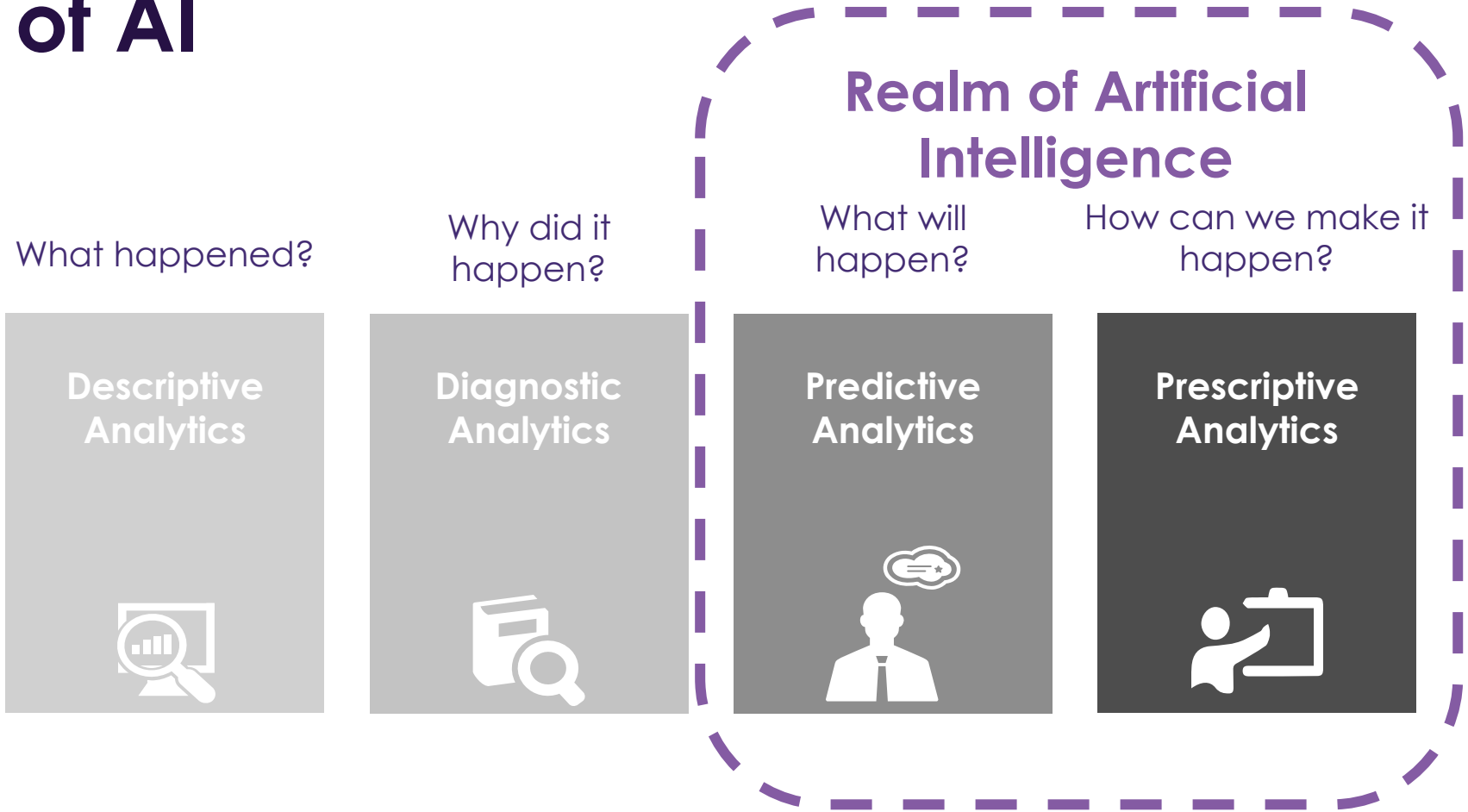
Stage 3: Predictive analytics

What questions does it answer?	What is likely to happen?
How valuable is it?	Provides trends / behaviors that are likely to happen
How labor intensive is it?	Requires detailed data, and may require a moderate to high level of computer power, depending on the method and the amount of data

Stage 4: Prescriptive analytics

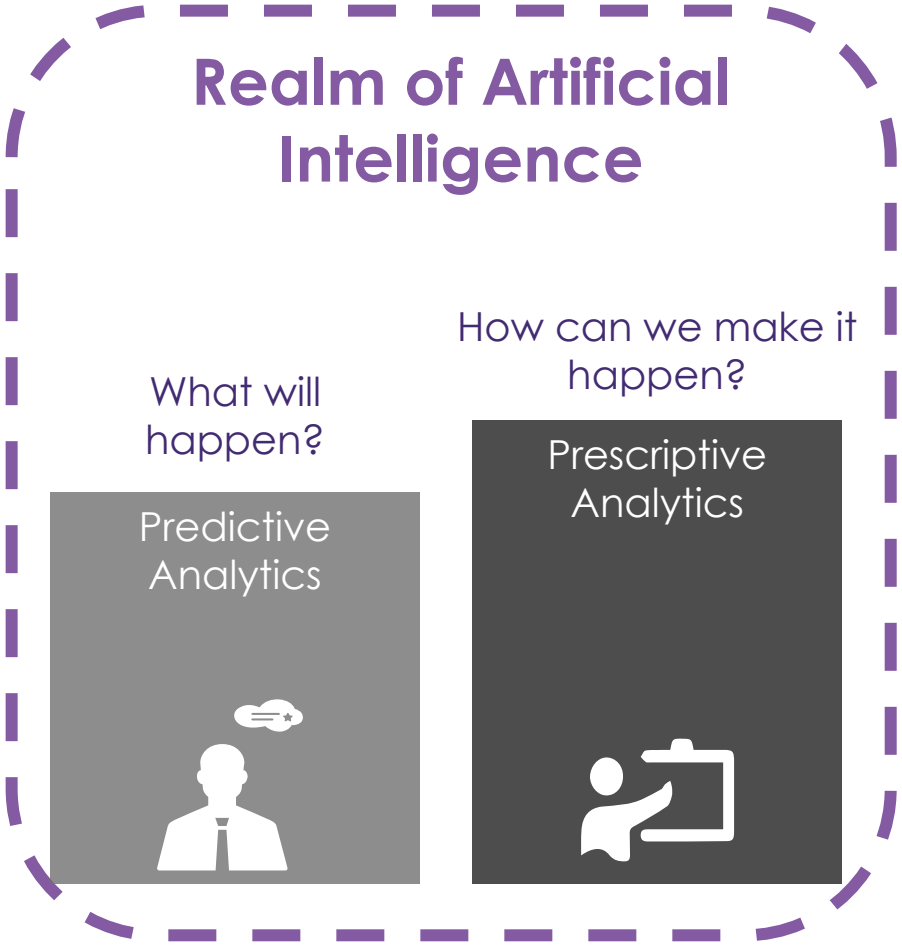
What questions does it answer?	What action should I take next?
How valuable is it?	Provides recommendations for future actions
How labor intensive is it?	Requires a lot of detailed data, as well as data from other external sources that will impact the model; very labor intensive

Realm of AI



AI for data analysis

- **Artificial intelligence** technology processes, comprehends, and extracts insights from huge, complex datasets.
- Some examples of implementing AI systems in data analysis:
  - Automatically classifying and categorizing unstructured data
  - Creating models to anticipate future trends
  - Identifying correlations and patterns



Example: FHA

- The Federal Housing Authority (FHA) forecasts default, repayment, and claim rates using big data analytics.
- They also leverage analytics to anticipate the premiums needed to maintain positive cash flow.
- During the 2008 recession, they were the only subprime mortgage insurance fund that did not require bail out.

Example source: Five Examples of How Federal Agencies Use Big Data (linked)

Chat question

[Select ALL that apply]

What type of data analytics is demonstrated by the Federal Housing Authority example?

- ❑ Descriptive
- ❑ Diagnostic
- ❑ Predictive
- ❑ Prescriptive

Break

Agenda

Day 1

- Fundamentals of data
- Data analytics overview
- **Data governance**
- Data teams
- Data tools

Chat question

What does “quality data” mean to you?

What is quality data?

- Data quality is crucial as it assesses whether information can serve its purpose in a particular context.

Clean or quality data is:

- ✔ Valid
- ✔ Accurate
- ✔ Consistent
- ✔ Complete
- ✔ Uniform

Clean or quality data is **not**:

- ✘ Corrupt
- ✘ Incorrect
- ✘ Duplicate
- ✘ Incomplete
- ✘ Wrongly formatted

Why use quality data?

- Poor-quality data is often pegged as the source of operational snafus, inaccurate analytics, and ill-conceived business strategies.

- Clean data has a range of benefits, and it helps with:
  - ✓ Staying organized
  - ✓ Avoiding mistakes
  - ✓ Improving productivity
  - ✓ Avoiding unnecessary costs
  - ✓ Presenting the true story of the data

Take action: prevent data issues

- Security! Track who has access to the data and who has the permission to modify it.
- Employ version control, backups, and redundancy.
- Use redundancy and other methods to regularly check data quality.
- Identify where human error occurs; keep track of data's travel path.
- Establish organization-wide standards for:
  1. Data entry
  2. Data checking
  3. Records structure
  4. Data ownership
- Train analysts and data owners on data quality.

What is data governance?

- **Data governance** is a collection of practices and processes that help to ensure the formal data management within the organization.
- Within each process, data governance is concerned with:
  - Awareness and communication
  - Policies, standards and procedures
  - Tools and automation
  - Skills and expertise
  - Responsibility and accountability
  - Goal setting and measurement

Key drivers for data governance

1. **Regulatory compliance**
   With increased regulation comes compliance that needs to be implemented and followed

2. **Reduce risk**
   Effective data governance enhances data security and privacy

3. **Improve processes**
   When everyone follows the same standards, projects and management become more efficient

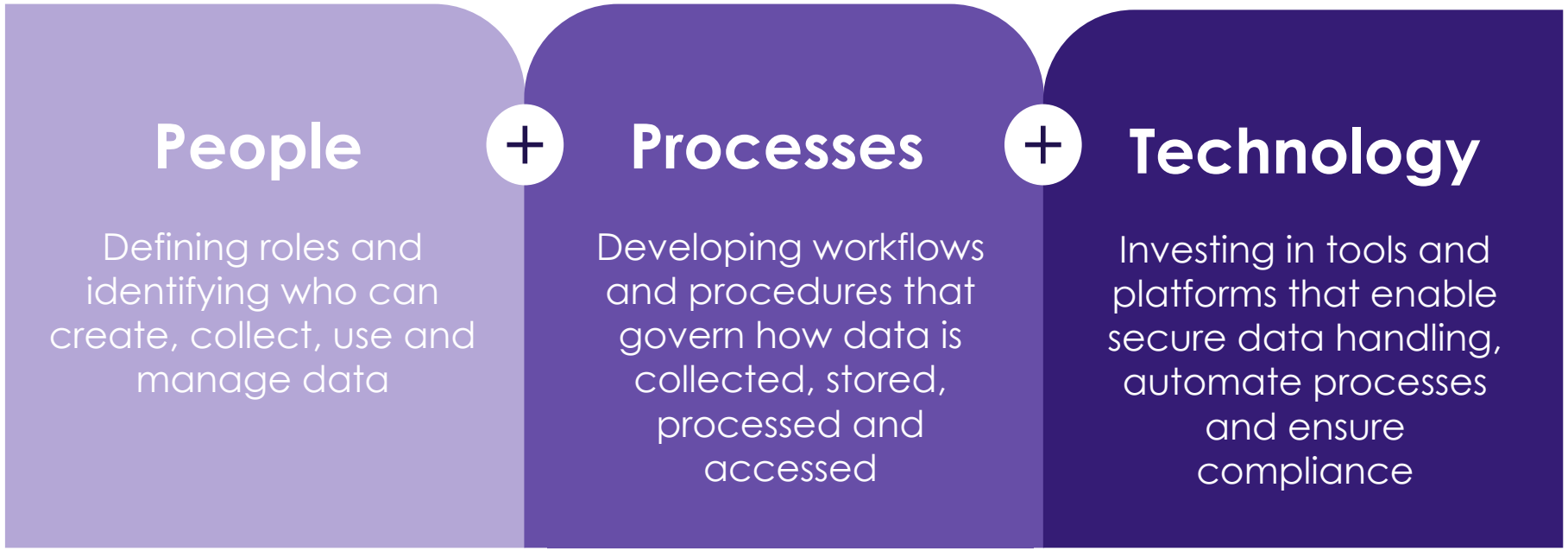
4. **Rise of AI**
   There are potential risks related to data privacy, accuracy, and ethical usage

Examples of existing regulations

- National and international governments draft, publish, and enforce data ethics rules.
- Some of the most widely recognized data governance guidelines in the world include:
  - California Consumer Privacy Act (CCPA)
  - Health Insurance Portability and Accountability Act (HIPAA)
  - Family Educational Rights and Privacy Act (FERPA)
  - EU Data Governance Act (DGA)
  - General Data Protection Regulation (GDPR)

Data governance framework

- Data governance of an organization relies on a balanced approach, encompassing not just policies and standards on processes, but also the people and technology.

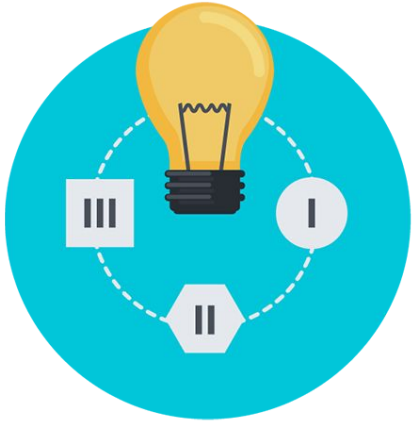


Data governance strategy

A data governance program might be documented using:

Charters

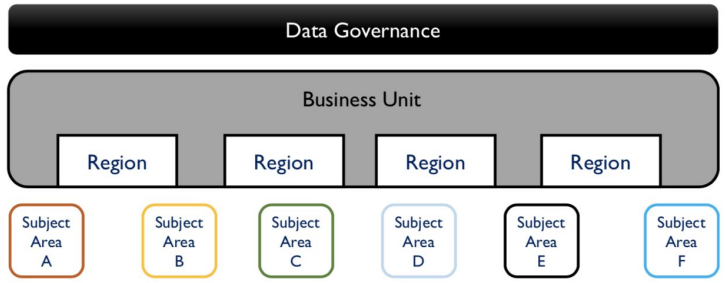
Implementation roadmaps



Operating frameworks / accountabilities

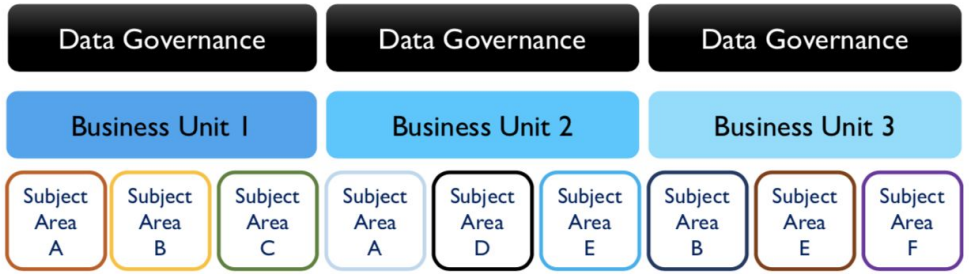
Plans for operational success

Data governance models



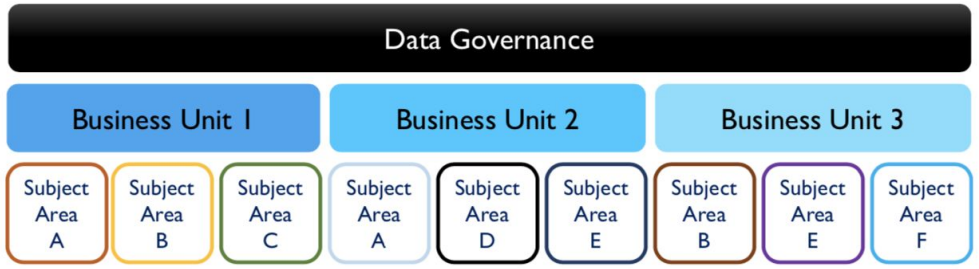
Centralized

One overarching data governance organization applies to all sectors.



Replicated

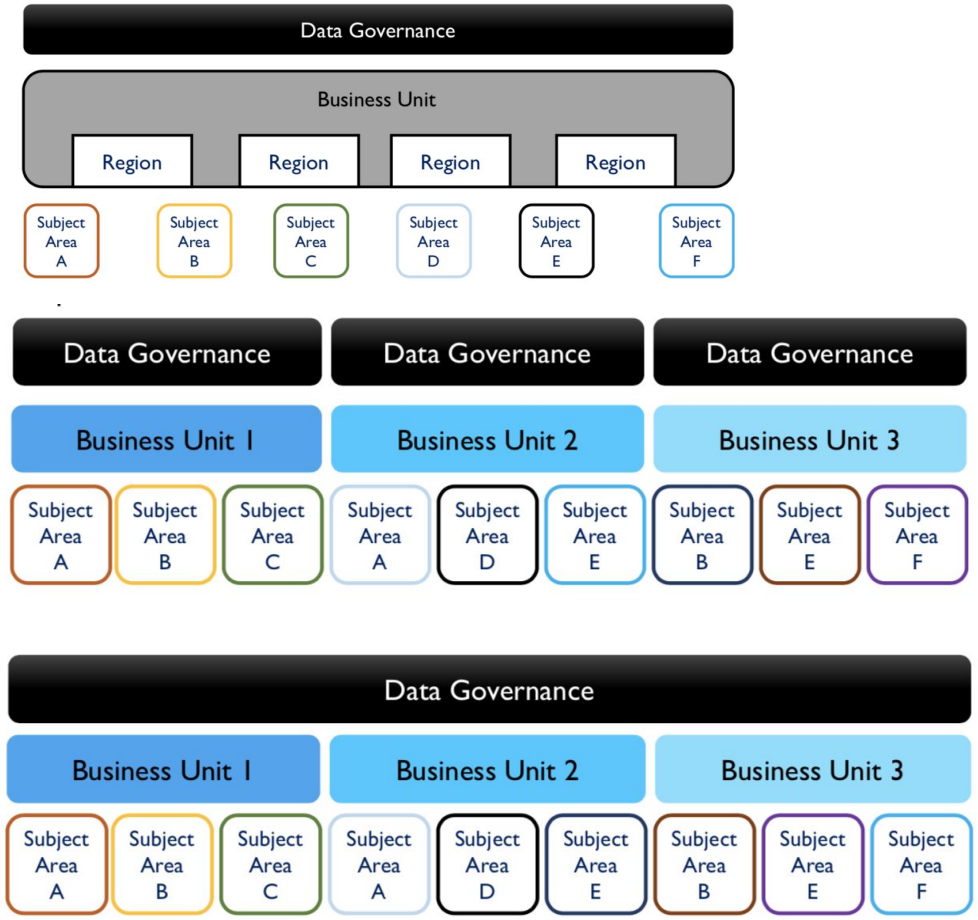
Each data governance section is repeated across departments but may have multiple governing bodies.



Federated

An overarching data governance organization works with multiple departments to maintain consistency.

Chat question



Which of the three models feels most applicable, in your experience, to how your organization operates?

Top: Centralized
Middle: Replicated
Bottom: Federated

Promoting good governance

- Catalog the data "owned" by your team or office. Which elements are **critical**?

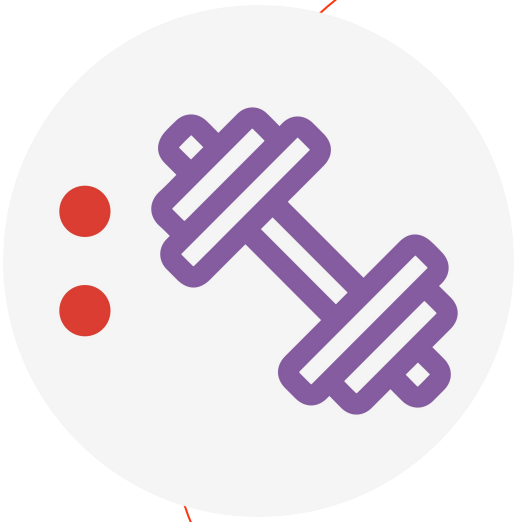
- Define roles and responsibilities:
  - **Data owners** are accountable for the state of the data.
  - **Data stewards** make sure that data policies and standards are adhered to and stay abreast of changes.

- Develop standardized data definitions and educate stakeholders on them.

- Implement preventative and detective controls to improve data quality.

Activity: evaluate yourself!

- Turn to your participant guide to the **data governance assessment**, which begins on page 4, to see how far along you and your team are in the data governance cycle.
- You'll measure the foundational components, such as **awareness**, **formalization**, and **metadata**, as well as the project components of **stewardship**, **data quality**, and **master data policies.**
- Then, assess your progress and set goals for where you want your team.
- In the chat, share your main takeaways



Agenda

Day 1

- Fundamentals of data
- Data analytics overview
- Data governance
- **Data teams**
- Data tools

Data team: Key functions

Data Owner

Typically a senior business leader who is responsible for data domains within the organization.

Data Steward

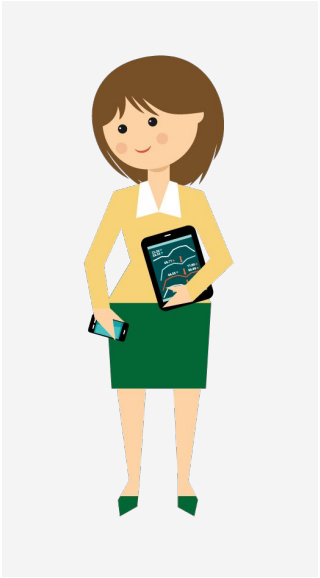
A “go-to” person for data-related questions within a specific data domain, acting as an interface between the technical team and the data owner.

Data Architect

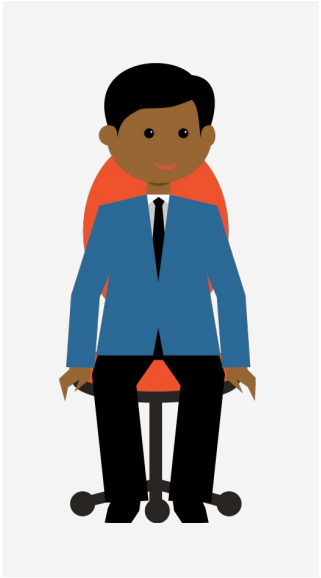
Responsible for designing and maintaining the data infrastructure in line with the policies and standards outlined by the data owner.

Typical data and AI team personnel

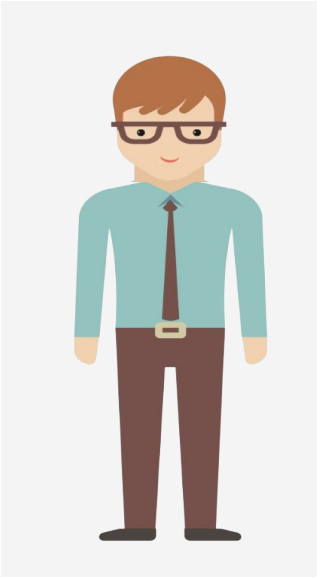
DATA ANALYST



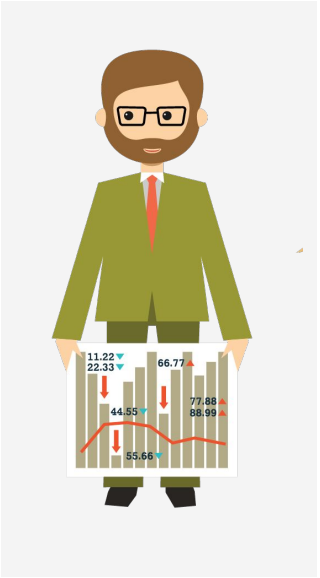
DATA SCIENTIST



DATA ENGINEER



SOFTWARE/AI ENGINEER



MLOPS EXPERT

DATA SCIENCE MANAGER

- Note: **Collaboration is key in data teams**, where responsibilities can overlap. Similarly, job titles are not uniform and may differ based on an organization's specific structure.

Data analyst

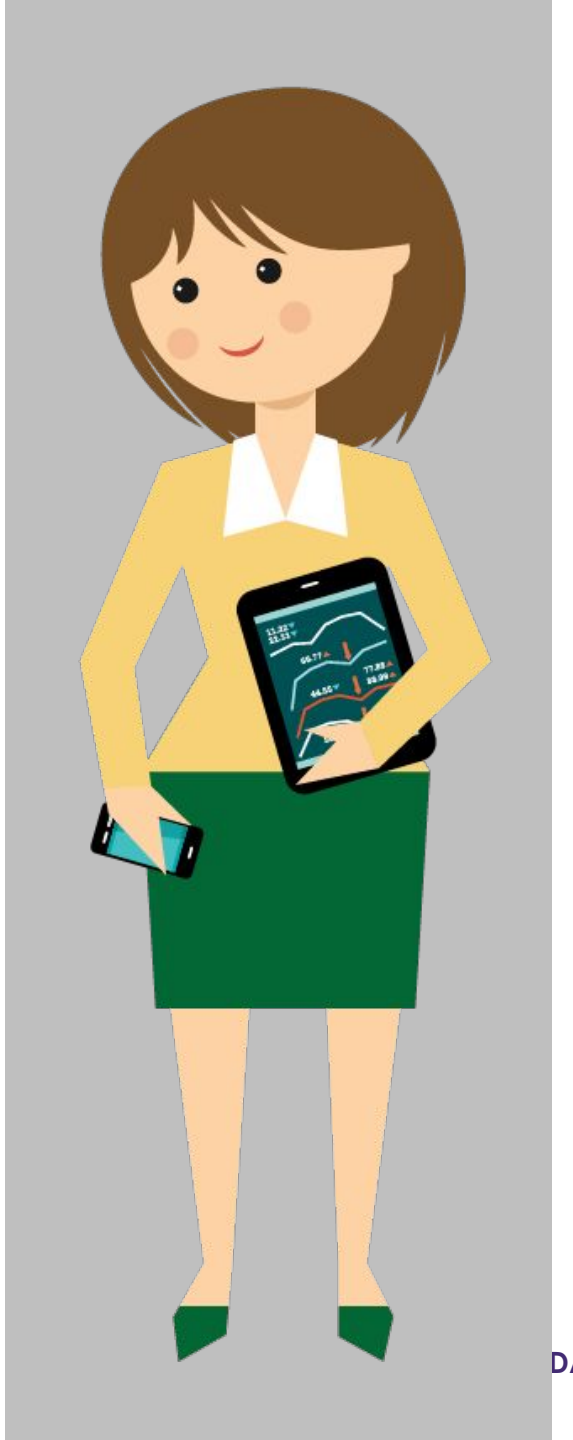
- Ensures that collected data is relevant and exhaustive while also interpreting the analytics results

- Main responsibilities include:
  - Wrangling the data
  - Managing the data
  - Creating basic analyses and visualizations

Data scientist

- Builds upon the analysts' data work to develop predictive models and complex algorithms

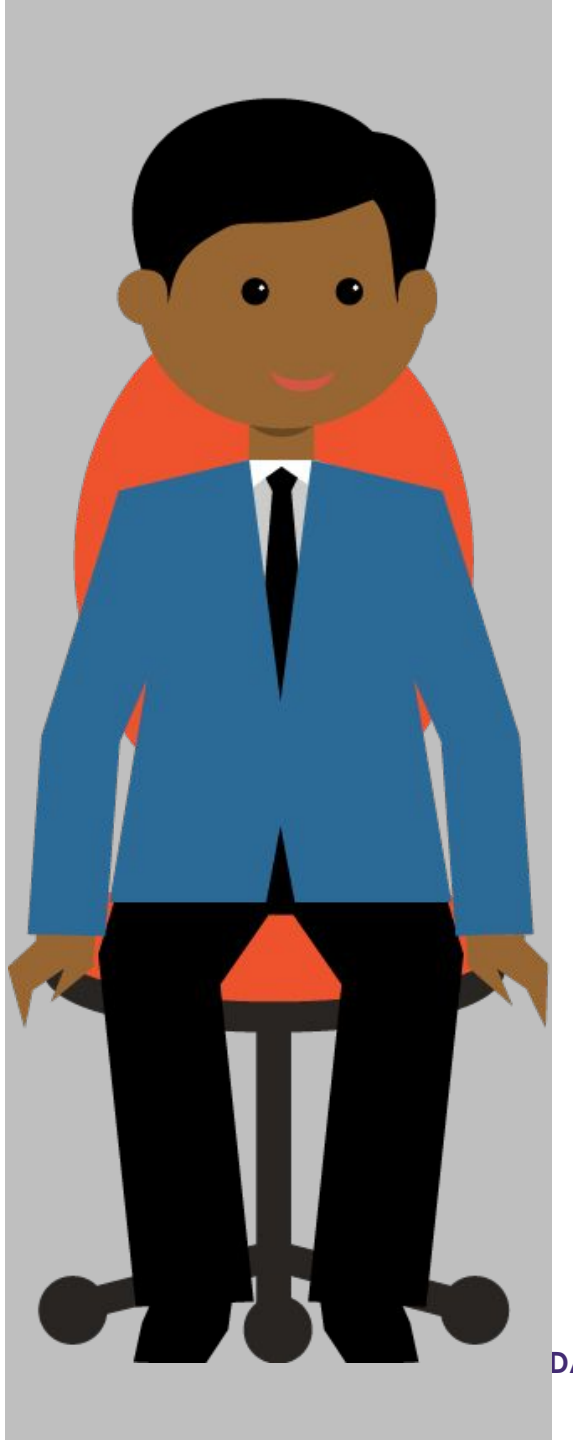
- Main responsibilities include:
  - Asking the right questions from the data
  - Building more complex predictive models
  - Interpreting the results critically and communicating them well



Data engineer

- Develops the infrastructure to house the data and maintains the structural components

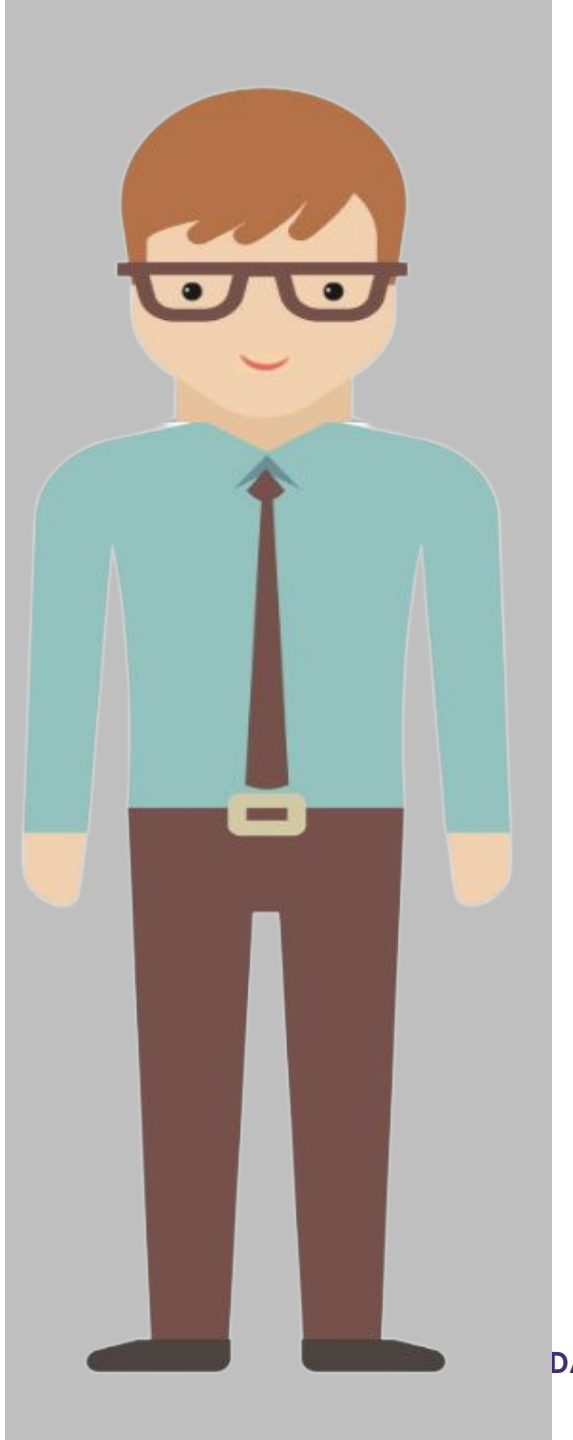
- Main responsibilities:
  - Ensuring data integrity across different data sources
  - Building out additional data warehouses as needed
  - Maintaining data pipelines and access



Software/AI engineer

- Responsible for building and maintaining AI software

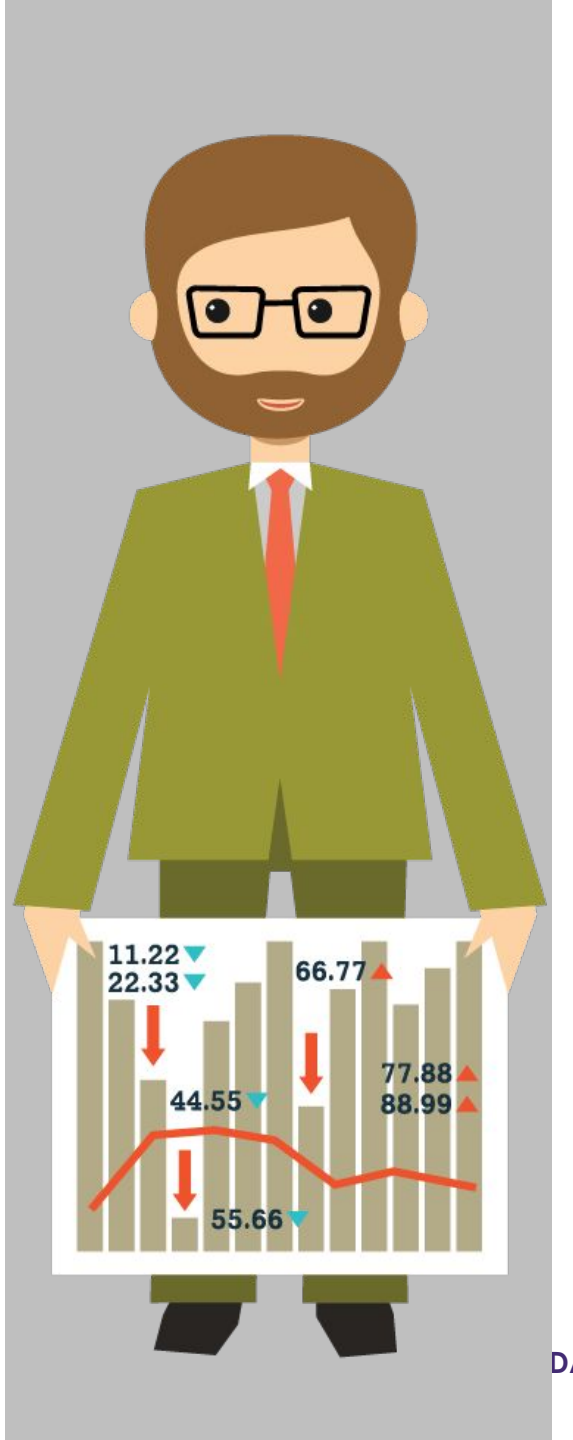
- Main responsibilities:
  - Construct frameworks that enable AI functions
  - Ensures seamless installation and integration of AI models into software applications
  - Designs front-end systems and integrates data sources



Machine Learning Operations Engineer

- Aims to deploy and maintain machine learning systems in production reliably and efficiently

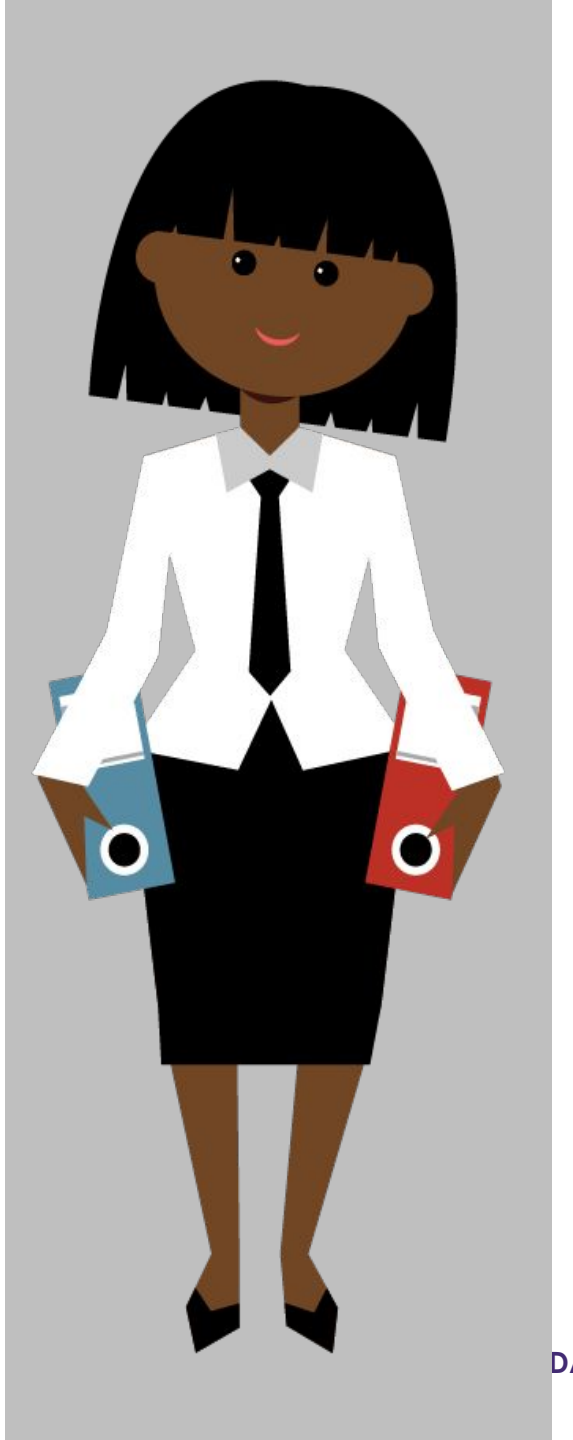
- Main responsibilities:
  - Requirements engineering
  - System design
  - Implementation and testing
  - Maintenance, support, troubleshooting, etc.



Data science manager

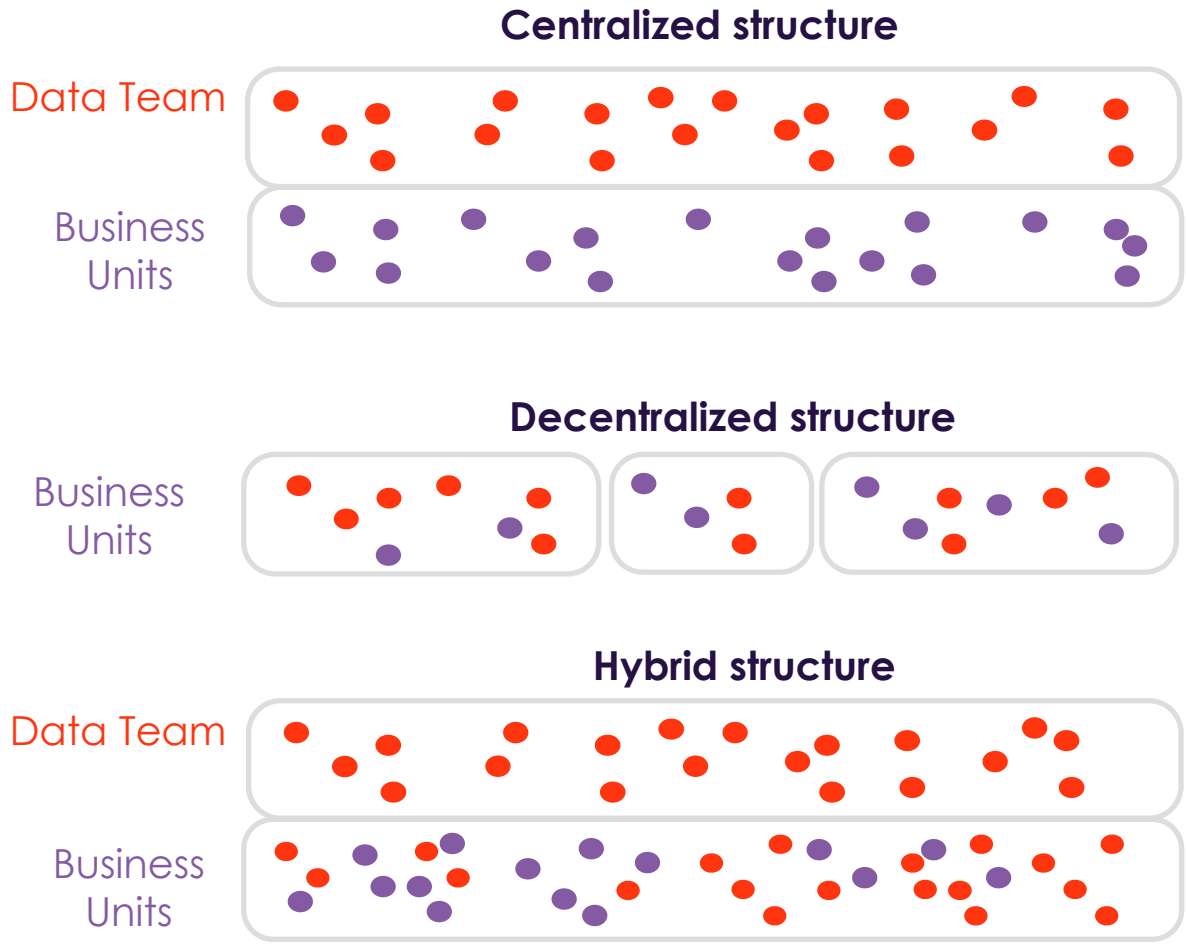
- Oversees and directs data teams and projects and bridges data and non-data people.

- Main responsibilities include:
  - Planning out people and resources for projects
  - Communicating results to executives and stakeholders
  - Running the data science teams



Team structures

- Data analysts
- Business analysts



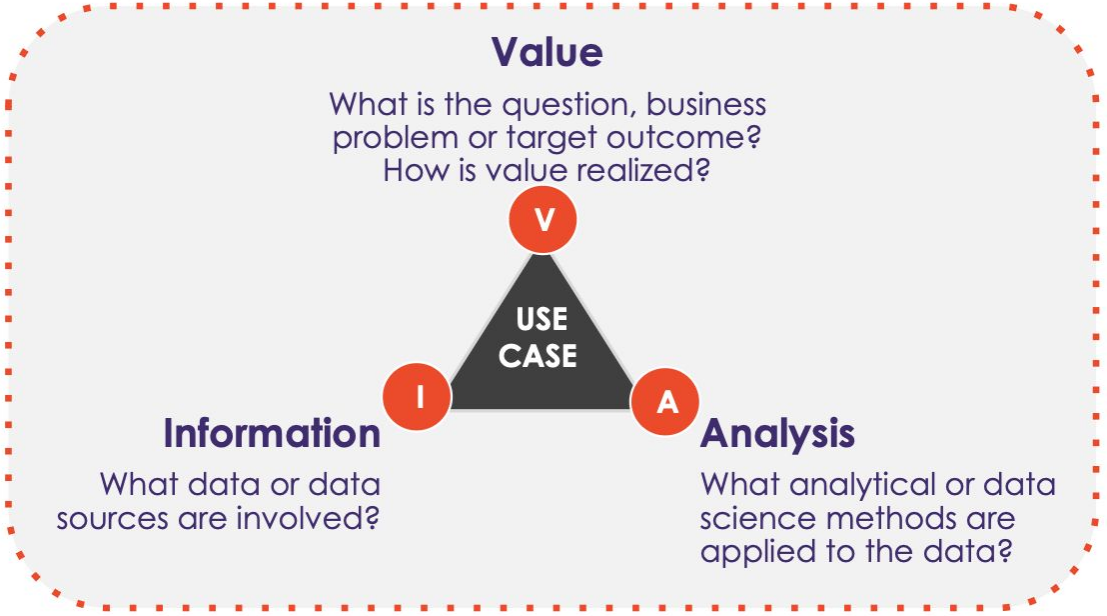
Chat question

Have you ever wondered:

- What if I am just starting out with using data?
- How can I effectively contribute to the process?

How can you contribute?

- To fit into the process, non-data scientists can contribute by:
  - Ask questions and help identify what projects to tackle
  - Think about the data requirements and be clear about what is needed
  - Ask questions about the process and the results



The Data Lodge® | ISL Institute

Agenda

Day 1

- Fundamentals of data
- Data analytics overview
- Data governance
- Data teams
- **Data tools**

Chat question

What software and applications are integral to your daily tasks and how do they support your responsibilities?

Tools

- There are many types of data tools that make working with data more efficient and faster.

Tools (cont'd)

A wide range of data tools are available, designed to accelerate data processing and often include built-in AI functionalities. Some common tool include:

Storage

- Databases
- Data warehouses
- Data lakes
- Cloud computing

Cleaning

- Drake
- OpenRefine
- Data Wrangler
- Data Cleaner
- Winpure

Analysis

- Excel
- R
- Python

Visualization

- Excel
- Power BI
- Tableau
- R and RStudio
- Python
- Power BI

Collaboration

- Git
- GitHub

Questions to guide tool selection

- Which steps are required in the data pipeline from ingestion to analysis?
- Which technologies are available for working with data at various stages of the data pipeline?
- How do different tools and technologies for working with data compare in their functionality, strengths and weaknesses?
- Do you have staff who can be trained or know how to use particular tools?
- Do you have budget constraints you need to be mindful of?
- Is it on the approved software list?

Summary

In this module, we covered:

- The benefits of data
- Data analytics overview
- Data governance
- Data teams
- Data tools

In the next module, we will cover:

- Data-driven cultures
- Putting together a project
- Foundational data science method

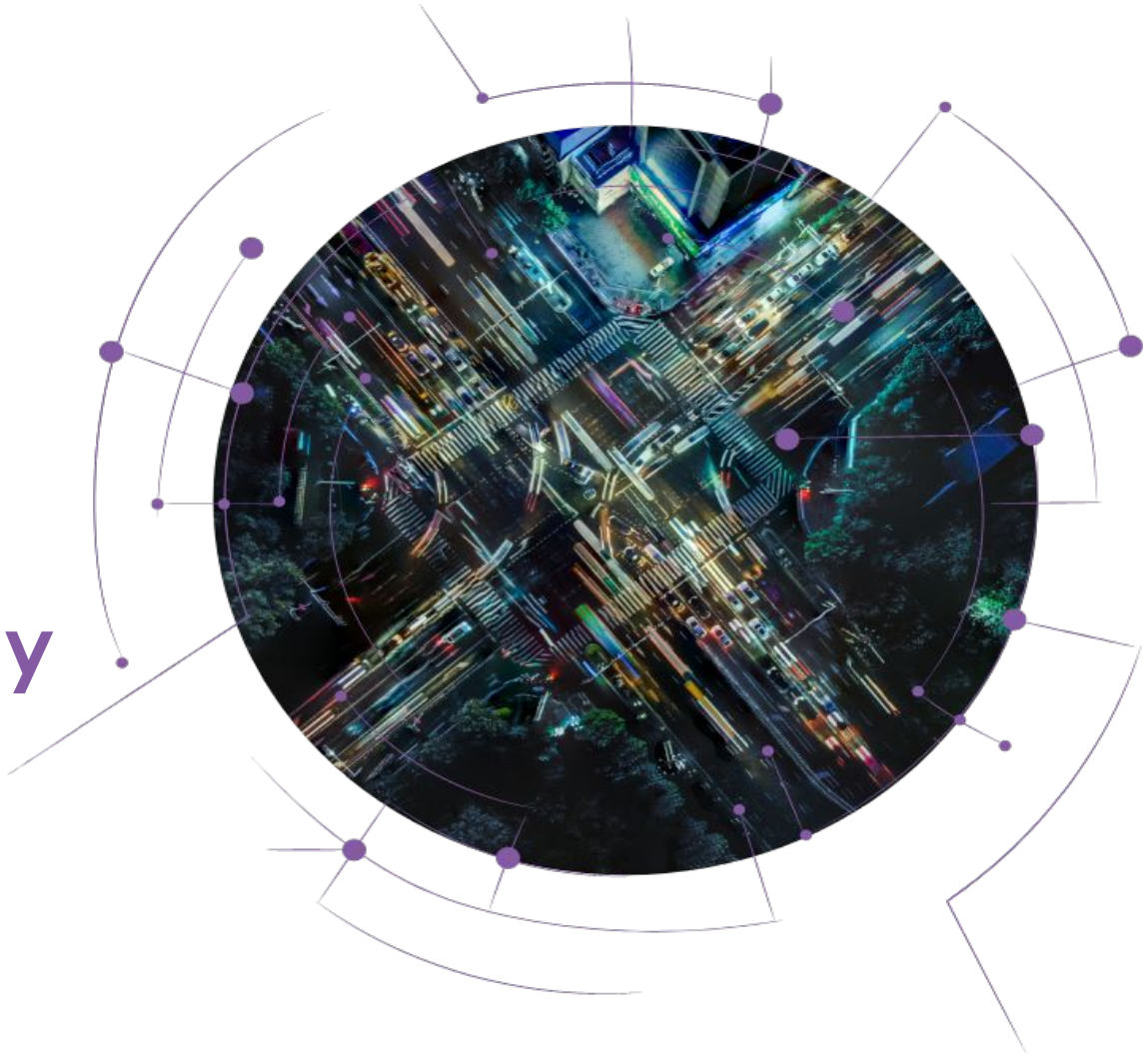
End of Day 1

Questions? Comments?

DATASOCIETY:

Fundamentals of Data Literacy

Day 2



Warm-up questions

- What is the most commonly used app on your phone?
- Do you think this app uses data? How?

Agenda

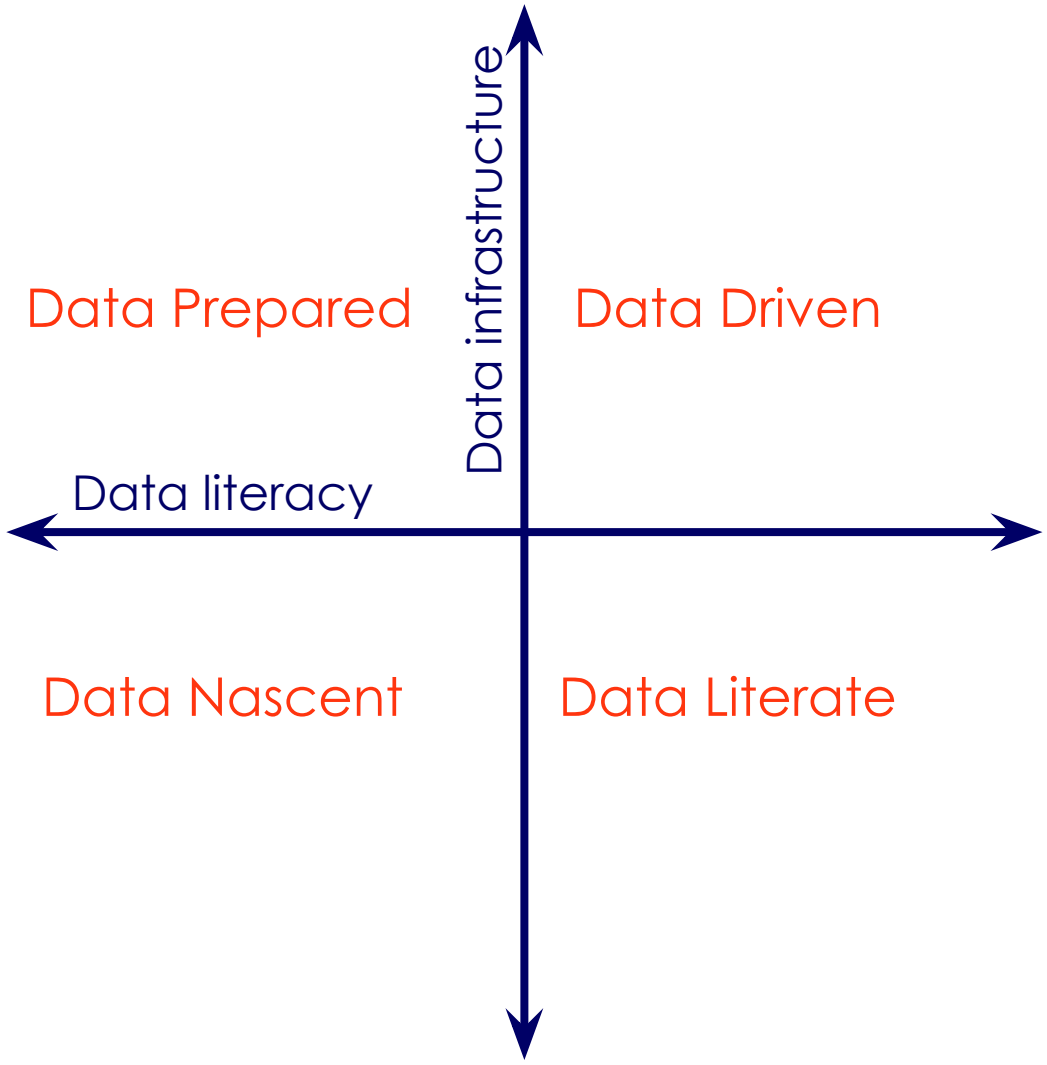
Day 2

- **Data-driven cultures**
- Putting together a project
- Foundational data science method

What is a data-driven culture?

- A data-driven culture incorporates data and analysis into its business decisions, systems, and processes.

- It can be separated into two main categories:
  - Data infrastructure
  - Data literacy



Data infrastructure

- Components of data infrastructure include:

DATA ACCESS

Can staff access data easily and in a timely manner?

DATA STORAGE

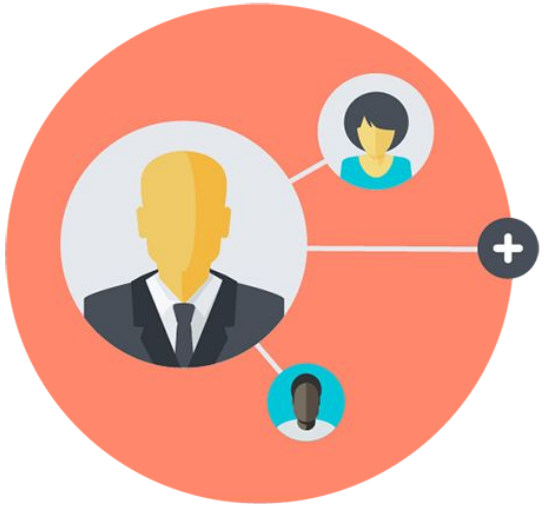
Is the data stored securely with a backup?

DATA COLLECTION

Is data collected in a timely and clean way?

Data literacy

- Components of data literacy include:



DATA LEADERSHIP

Do executives champion data usage?

DATA GOVERNANCE

Are staff aware of data standards and practices?

DATA KNOWLEDGE

Does staff understand how to ask questions of data?

Why is it important to be data driven?

- Identify trends. Trends can inform effective practices, help you become aware of issues, and illuminate possible innovations or solutions.

- Reduce bias. Making decisions based on data is far more reliable than ones based on instinct, assumptions, or perceptions.

- Benchmark performance. Benchmarking allows staff to connect their actions to business results, which will reveal new opportunities for improvement.

A study from the MIT Center for Digital Business found that organizations driven most by data-based decision making had 4% higher productivity rates and 6% higher profits.

Example: Walmart

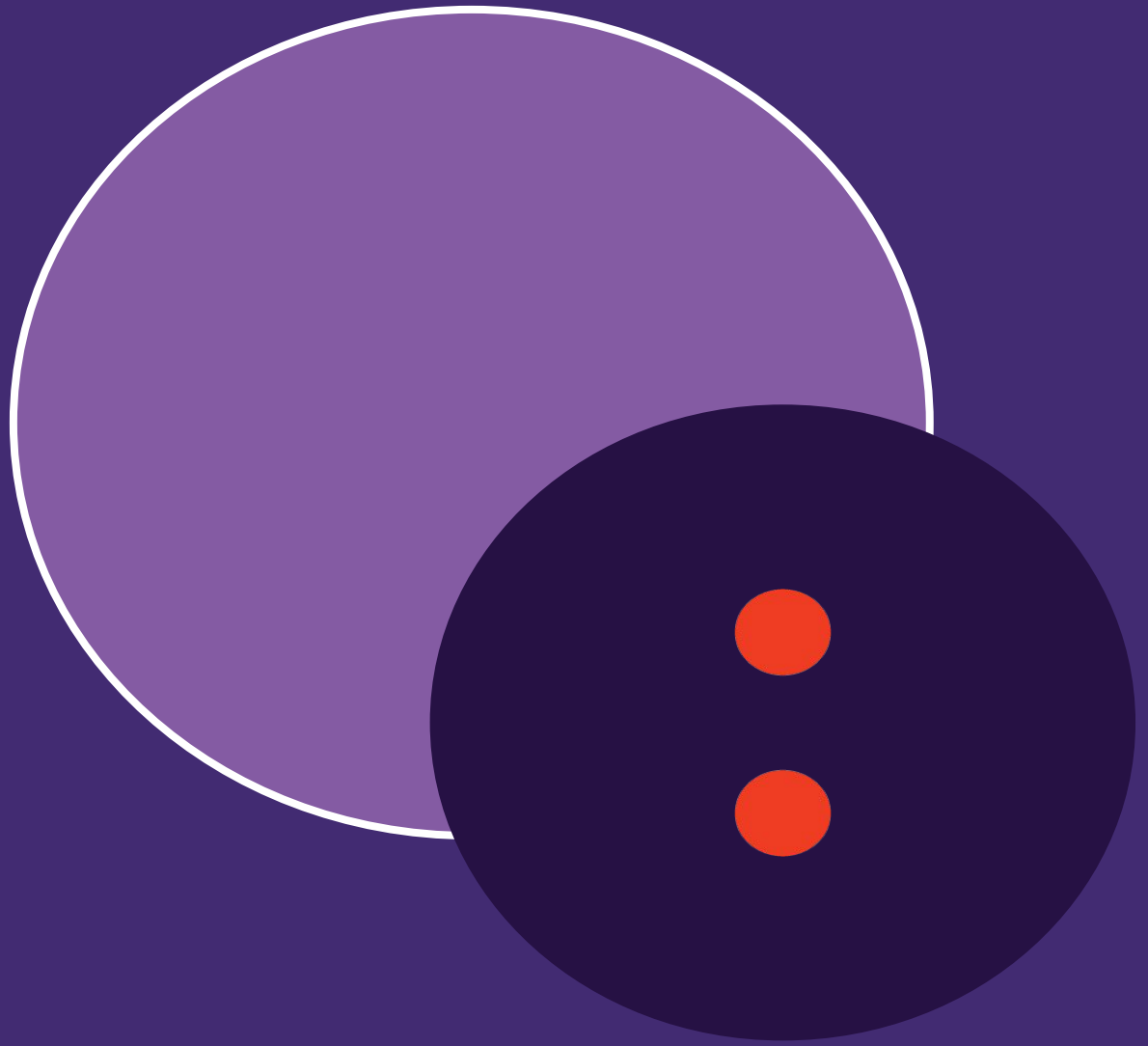
- Walmart executives wanted to know what items to stock before Hurricane Frances in 2004.
- Analysts mined a terabyte of purchase history from other Walmart stores under similar conditions.
- Turns out, in times of natural disasters, Americans want strawberry Pop-Tarts and beer! Stores were stocked accordingly.

Walmart Corporate, via Flickr

Discussion question

What is one thing **you** can do to personally create a data-driven culture within your team or organization?

How to encourage data-driven thinking & innovation



Step 1: Create data-driven guideline

- All the data may be irrelevant if it is not used correctly.
- Hence, organizations need to know how to extract information and knowledge from their data.
- They must incorporate data by developing objectives and laying out a broad roadmap for the data.

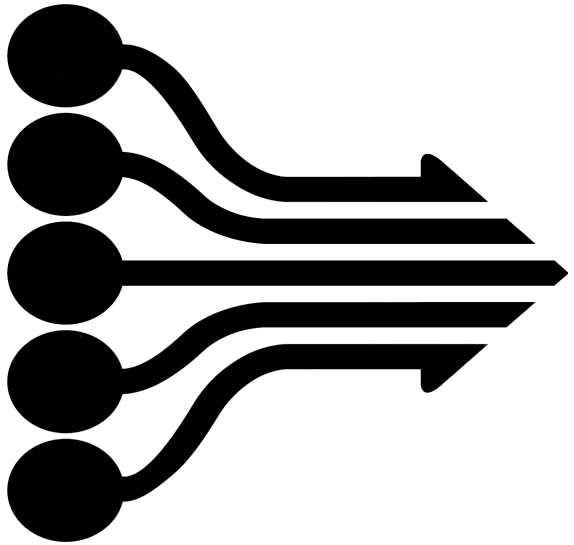
Step 2: Invest in data infrastructure and strategy

- Determine the space required to manage data for your organization and develop systems to support data collection, storage, and analysis.
- Collaborate with the IT department to establish databases and install software for data reporting, modeling, and analysis.
- Using the right tool to perform data analysis.

Step 3: Encourage careful and comprehensive methods of data collection

- Create policies for gathering data
- Establish practices to measure the success
- Discuss the importance of collecting data records in the future
- Meet with other managers or department leaders to communicate individual data collection methods

Step 4: Streamline data collection process:



- Every department gathers relevant and valuable data and hence have a central repository for all the collected data is recommended.
- Data analysts evaluate the data and provide understandable analytical reports and insights back to the head of each department.
- Each team can turn the outcomes from these insights into actions, execute them in their domain, and share results with other departments/teams.

Step 5: Improve and maintain the data quality

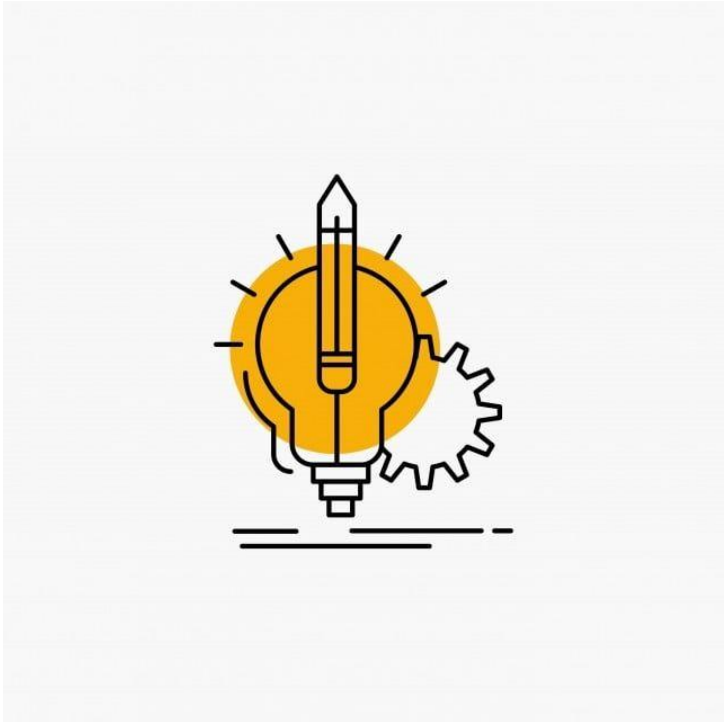
- Data quality is just as crucial as data quantity.
- If you do not have new data in your repository, you might be looking at outdated data and fake reality.
- Keep collecting more relevant, new data.
- Use data mining and software tools to clean and maintain the quality of the data automatically.

Step 6: Train your team

- Educate employees with the right data-related skills and knowledge.
- Plan or suggest official training sessions (like this one!) on data literacy and information analysis.
- Have your team complete a tutorial/training when the organization incorporates a new software or database system.

Step 7: Share insights and knowledge

- Encourage your team by sharing data directly relevant to them
- Show the value and impact of data by preparing information about the team's performance and sharing it with them during quarterly and yearly reviews.
- It is crucial to ask the team how they came to a conflict, analyzed it, and decided on the resolution. It gives your data team a deeper understanding of the data.



Step 8: Applaud your team

- Identify a successful analytics project / team and highlight their success through a newsletter, event, or lunch and learn.

- Recognize the right things—including when mistakes move you to another level.

Recap: 8 Actionable steps to establish data-driven culture

Here is the checklist of the steps:

- ✔ Create Data-Driven guideline
- ✔ Invest in data infrastructure and strategy
- ✔ Encourage careful and comprehensive methods of data collection
- ✔ Streamline data collection process
- ✔ Improve and maintain the data quality
- ✔ Train your team
- ✔ Share insights and knowledge
- ✔ Applaud your team

Discussion question

- Which idea(s) that we've discussed could you implement in the near term? What specifically would you do?

- What challenges do you expect to face when implementing those ideas? How will you overcome them?

What is a data-driven culture?

- An organization with a data-driven culture incorporates **data and analysis** into its business decisions, systems, and processes.

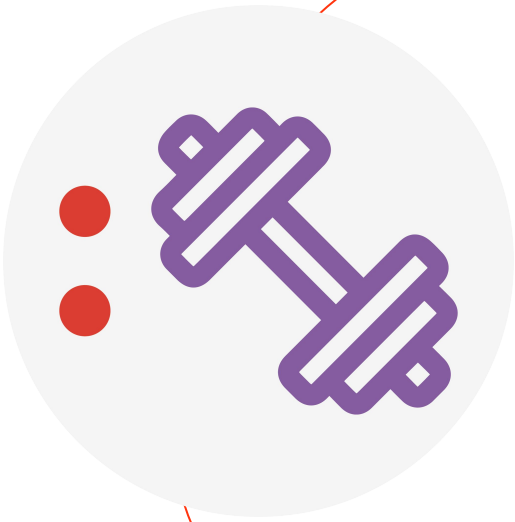
What companies or organizations come to mind when you think of a data-driven culture?

Activity: Are you data driven?

Turn to page 8 of your participant guide to the **data-driven culture assessment** to evaluate your team.

In the chat, answer the following questions:

1. **Which quadrant are you in?**
2. **What are key areas you'd like to improve?**

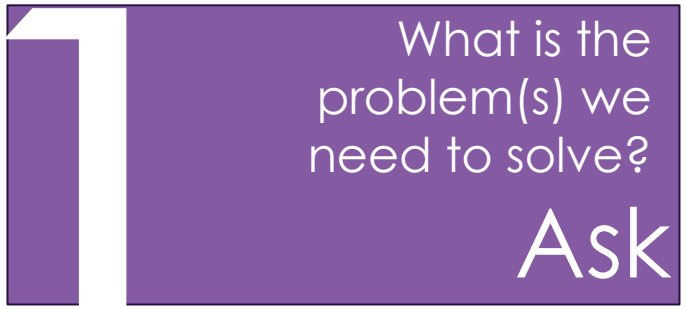


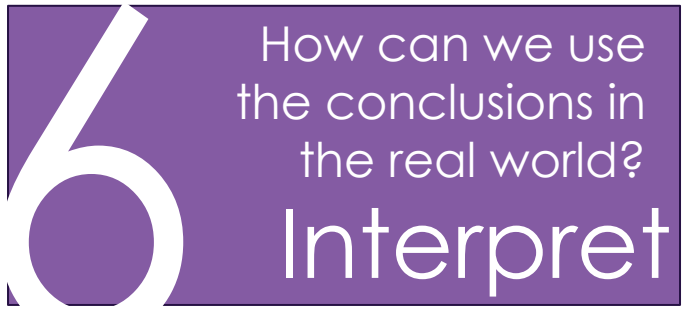
Agenda

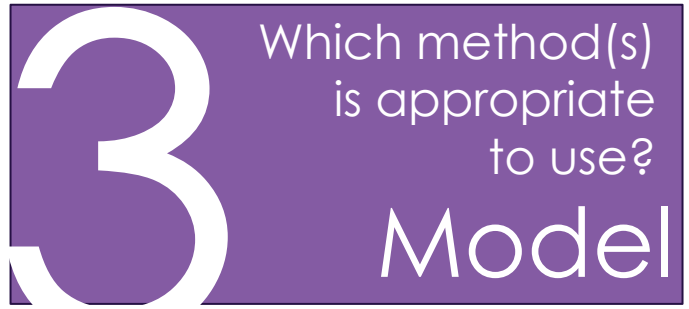
Day 2

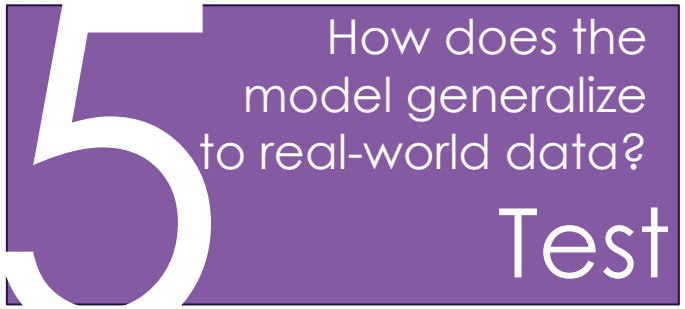
- Data-driven cultures
- **Putting together a project**
- Foundational data science method

Typical data science process









1 What is the problem(s) we need to solve? Ask

- The business and data teams should work together to develop a question that is **specific**, **measurable**, and **objective**.
- Domain knowledge comes into play.

Examples

How can I make my policies more effective? → Which 3 policies have demonstrated the best results, and did they have anything in common?

We'll use an indicator that shows the most improvement. → We'll use the calculated ROI and the percent difference in desired behaviors from before and after.

2 What data do we need and how do we get it? Research

- The data team, with input from the business, **gathers information** about the data needed to get a relevant answer.
- *Is it already collected, or is time needed to get it? What format is it in?*

Examples

I'm sure we have the data somewhere. → We'll use the datasets from the policy report that can be found in X repository.

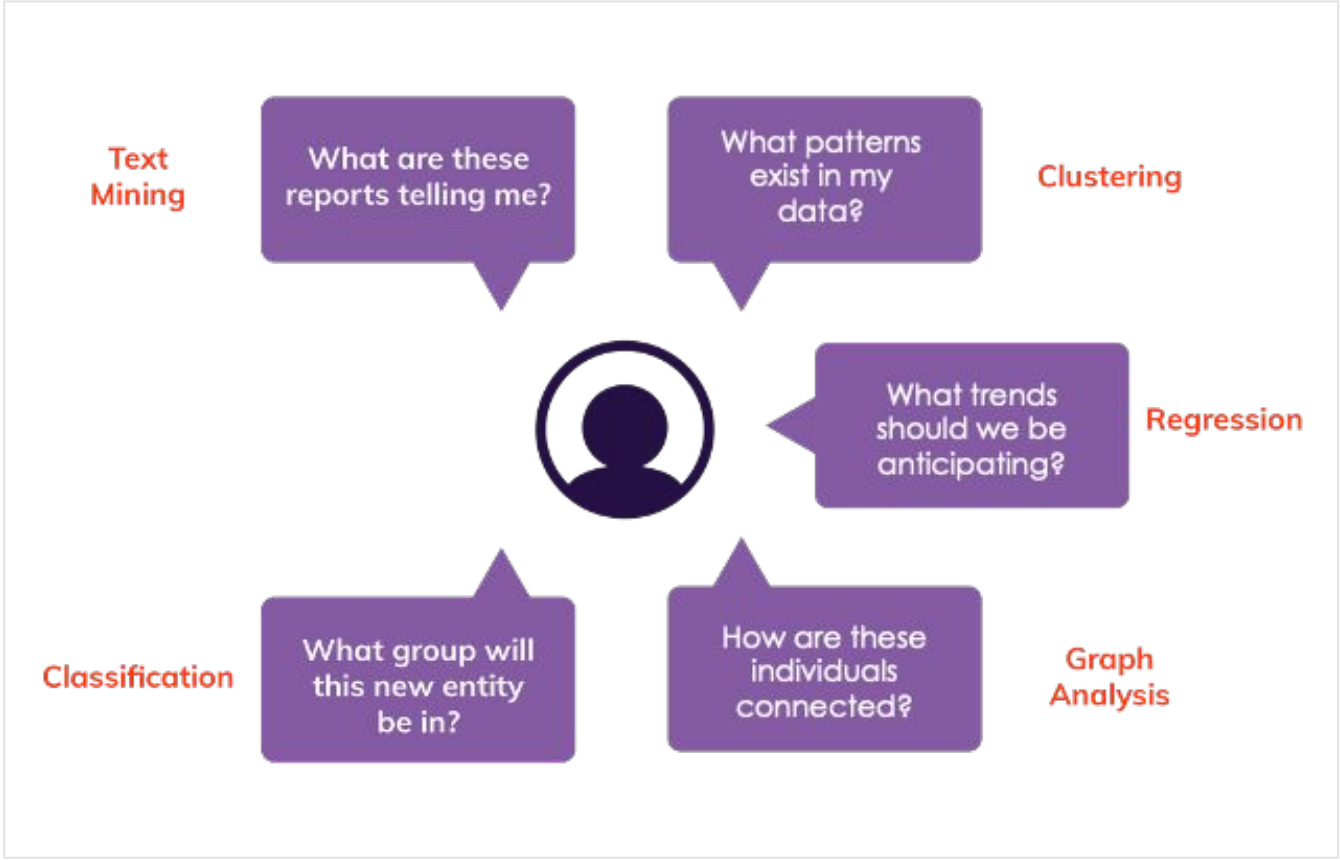
I'm sure the data is good enough as is. → Where can I read about how the data was collected and how the metrics are defined?

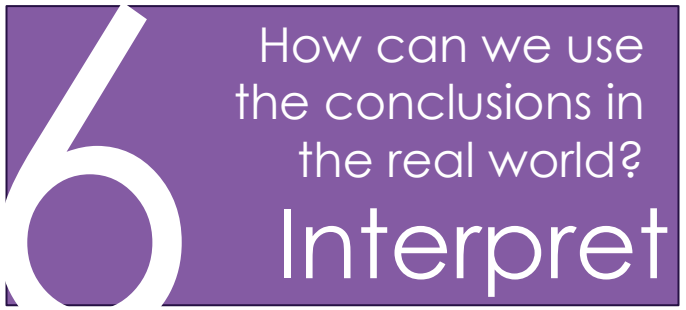
3 Which method(s) is appropriate to use? **Model**

4 Do the model and assumptions work as expected? **Validate**

5 How does the model generalize to real-world data? **Test**

- Models take questions and **provide answers and outputs**.
- The methods chosen by the data team are based on the questions asked and the type(s) of data that you have.
- **Multiple iterations** are required to ensure the model works well.





- The data team looks at **what the results are telling them**—not what they were expecting the results to be.

- They **present** the data and **make recommendations** based on the data, their domain knowledge, and stakeholder needs.

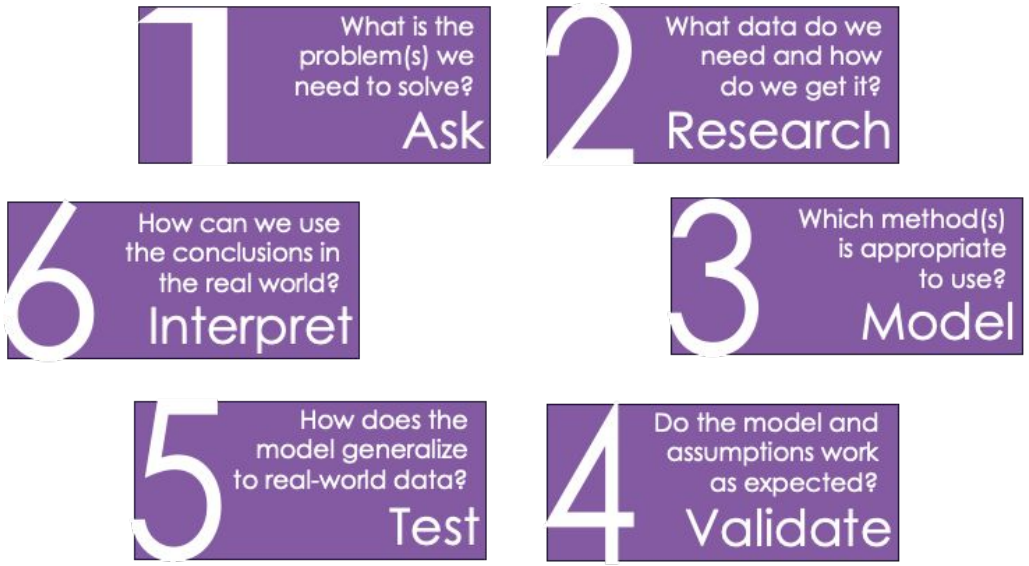
Example

I'll put the results in the same format as I usually do.

How can I best convey the results that matter most to my end users?

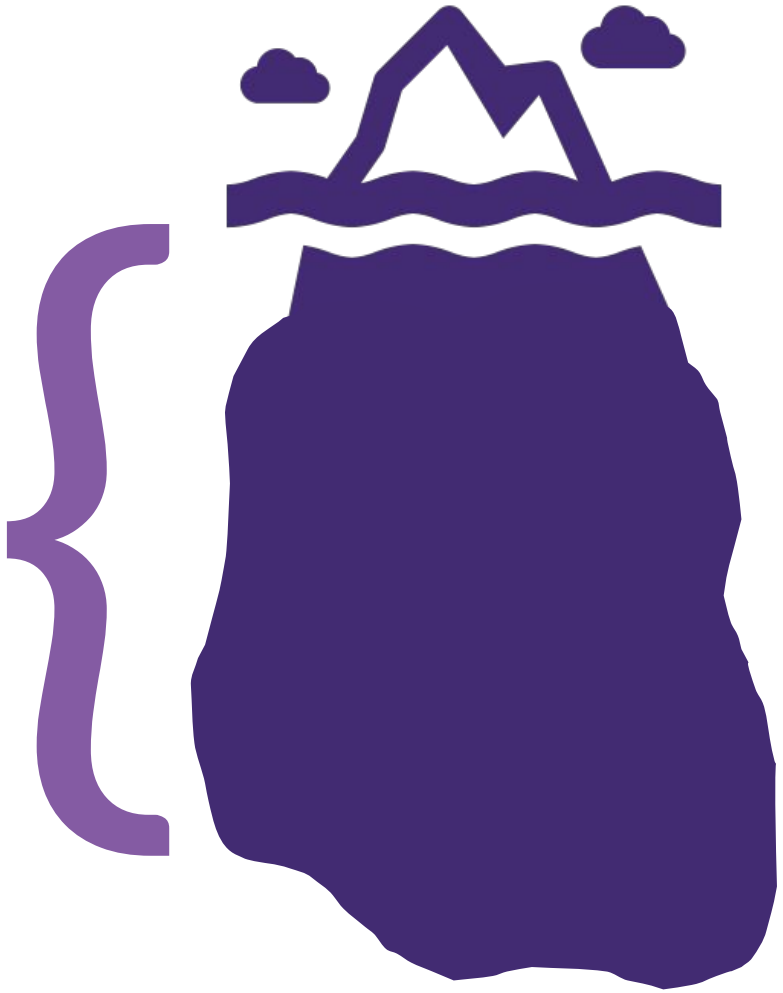
Discussion question

Which phase do you think takes the longest? Why?



The iceberg of data analysis

Cleaning data often takes 70-80% of project time.

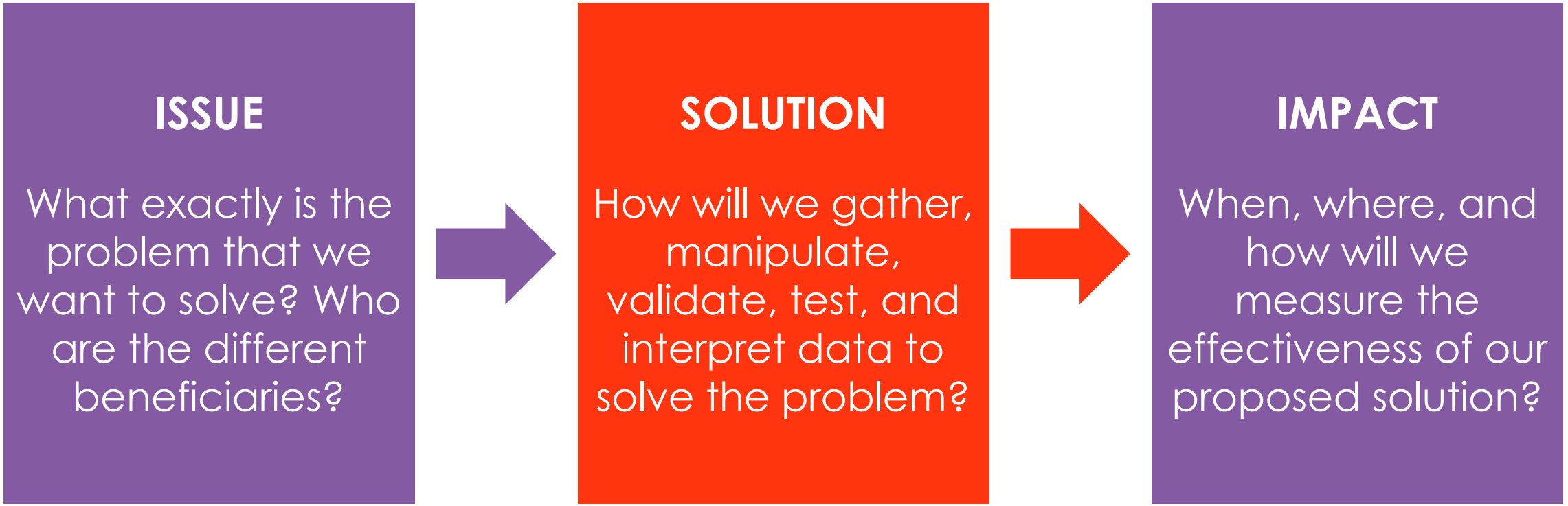


How do I fit in?

- How do non-data scientists fit into this process?
  - Help decide what projects to tackle
  - Be clear about what is needed
  - Ask questions about the process
  - Ask questions about the results

From issue to impact

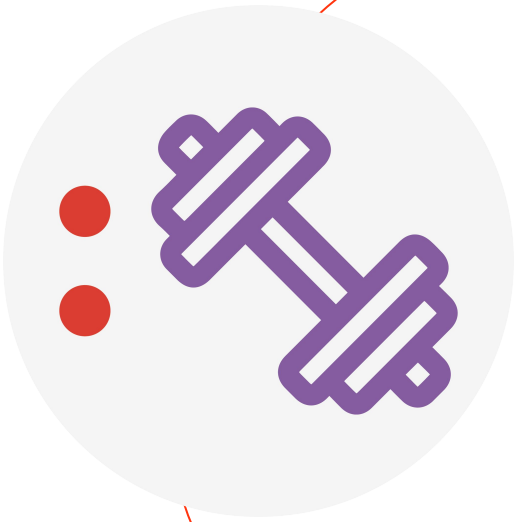
- When thinking about data science projects, it can be useful to consider the arc from **issue** to **impact.**



Activity: brainstorm ideas

- Turn to page 11 of your participant guide to the **Project brainstorm** activity.

- Identify 3-5 ideas for leveraging data in your workplace. Then, assess their feasibility and impact.



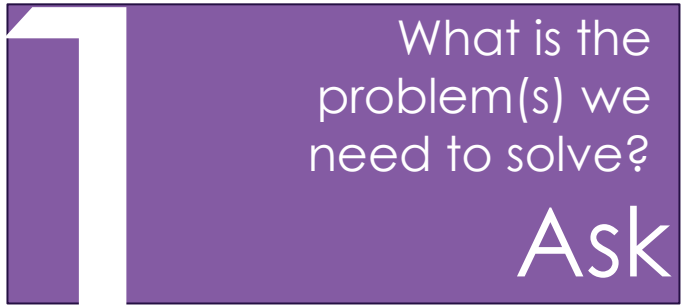
Break

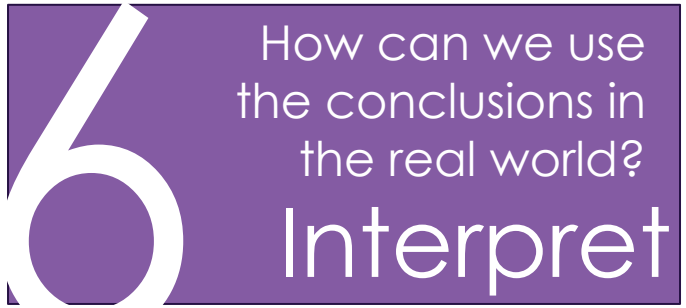
Agenda

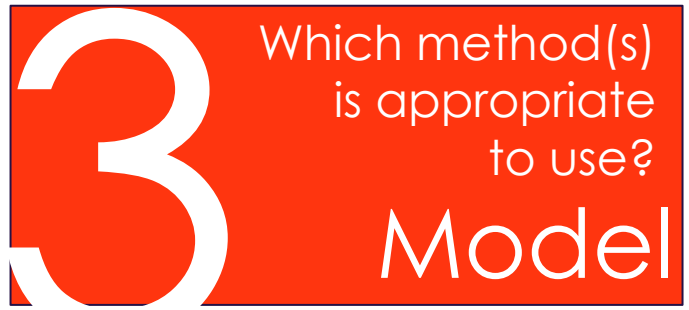
Day 2

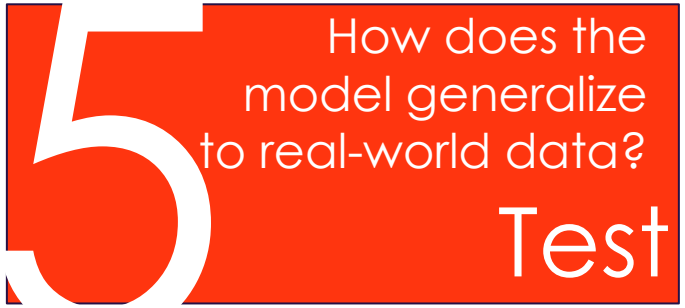
- Data-driven cultures
- Putting together a project
- **Foundational data science methods**

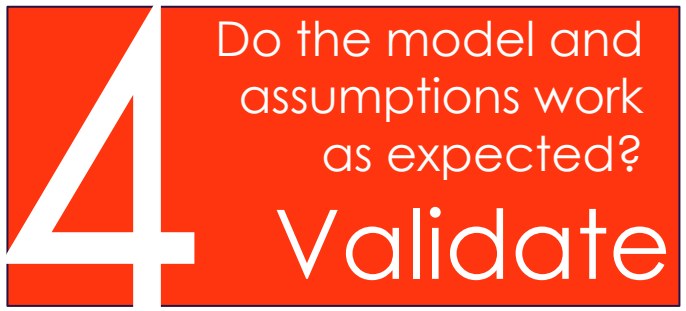
Data science process: model, validate, test











Why learn these terms and concepts?

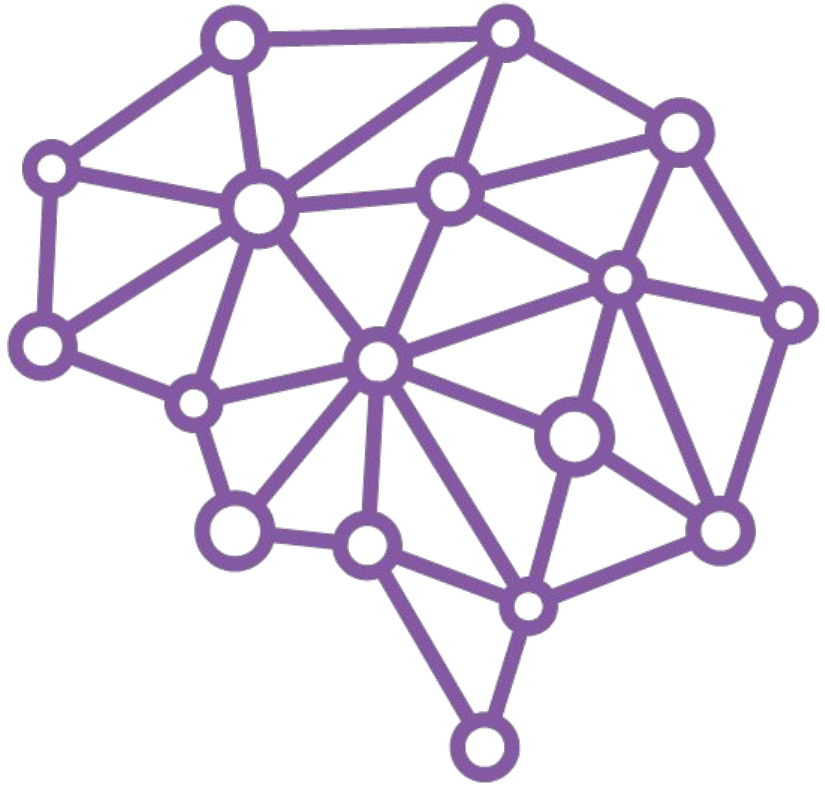
1. To develop a **common vocabulary** with the data science team
2. To direct data science projects and **make recommendations**
3. To understand what **options** are available for finding new insights and becoming more efficient

Discussion question

What types of AI/ML do you interact with on a day-to-day basis, either professionally or personally?

What is AI?

- **Artificial intelligence (AI)** refers to systems or machines that **mimic human intelligence to perform tasks and can iteratively improve themselves** based on the information they collect using machine learning (ML).
- It is a form of intelligence that is used to solve problems, come up with solutions, find answers, make predictions or recommendations.



Spotlight: Increased productivity

Yes, AI Increases Productivity, Study Suggests

Joe McKendrick Contributor ⓘ

I track how technology innovations move markets and careers

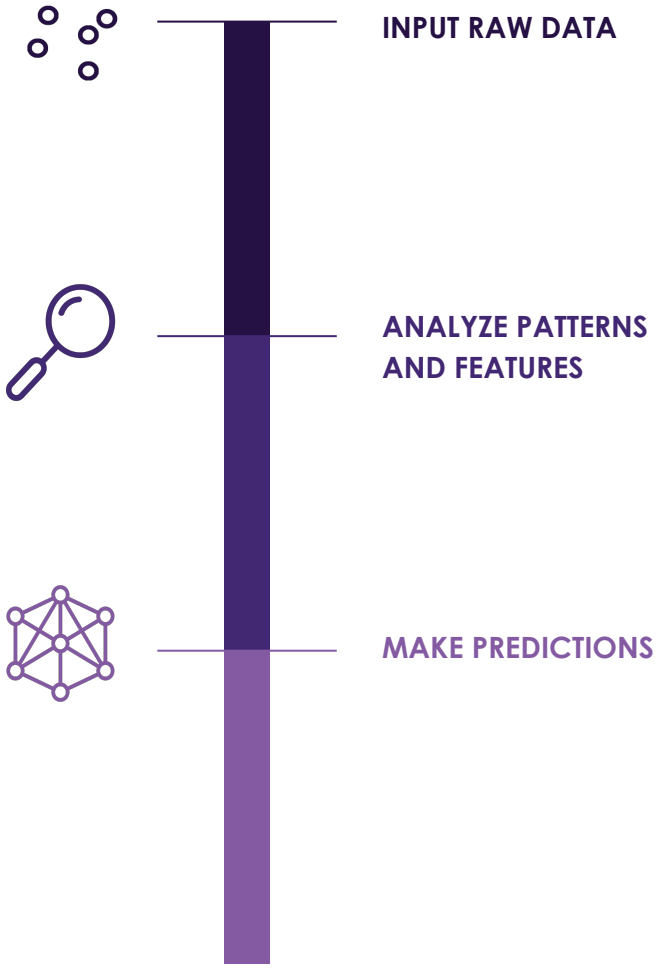
Follow

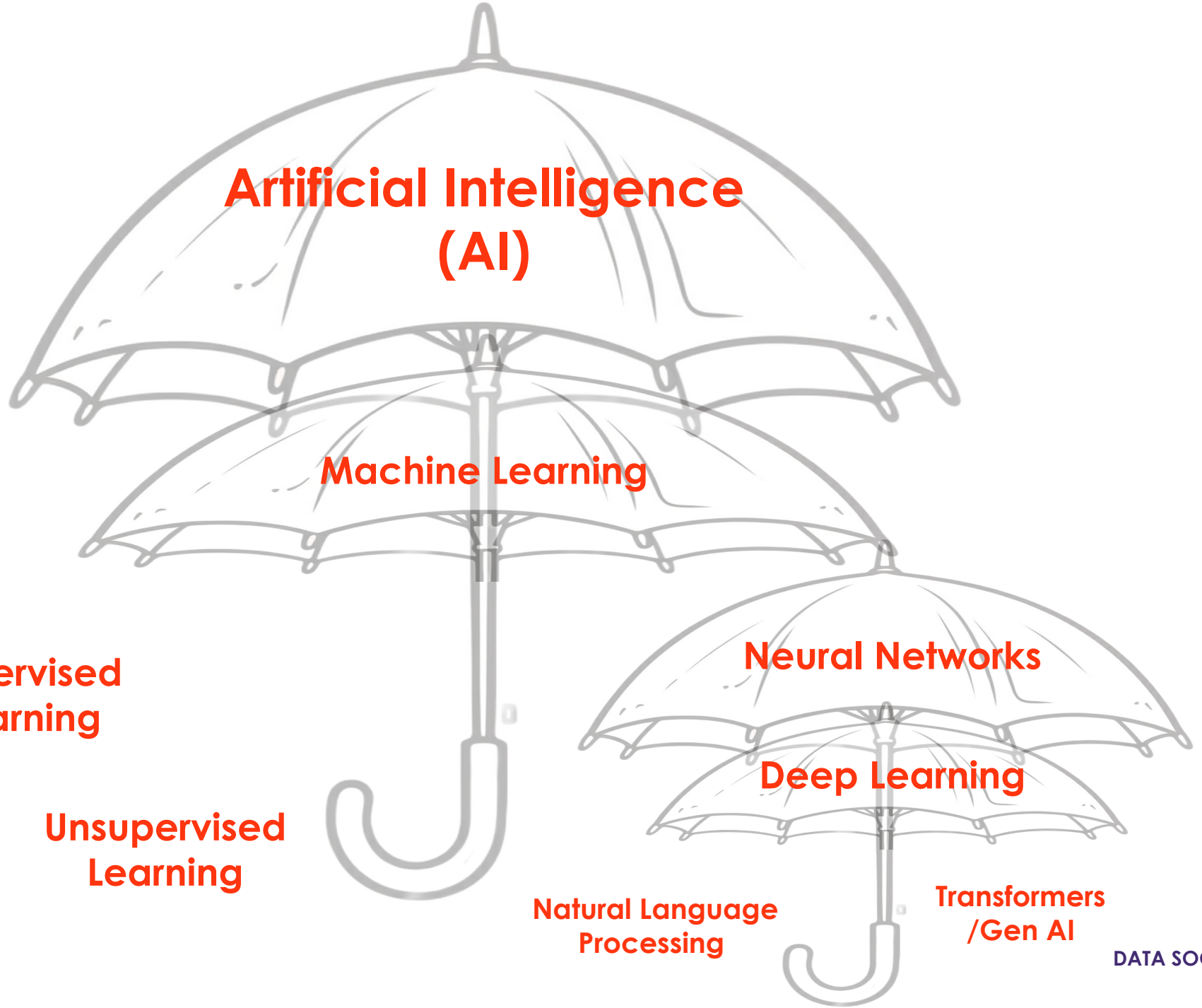
Apr 25, 2023, 12:39am EDT

Source: Forbes (link)

How AI works?

- **To apply AI, you need data**. A lot of data!
- AI "algorithms" are trained using large datasets so that they can identify patterns, make predictions and recommend actions, much like a human would, just faster and better.
- There are several data science methods, including machine learning, applied to build AI technologies.





What is machine learning?

- **Machine learning (ML)** uses **algorithms** to find patterns in massive amounts of data and **predict** future results with minimal human intervention.
- It powers many of the services we use today:
  - Recommendation systems like those on Netflix
  - Search engines like Google
  - Voice assistants like Siri and Alexa
- Most is categorized as either supervised, unsupervised or semi-supervised.

Supervised learning

- An algorithm is given **tagged and labeled** input data to create such a solid map that we can predict the output for any novel input data.

- Goal: mapping known input to a known output

Unsupervised learning

- An algorithm is allowed to detect and learn patterns based on **untagged** input data, usually to **generate categories.**

- Goal: mapping known input to a unknown output

Chat question

Can you think of any uses of AI/ML in the field you work for?

Summary

In this module, we covered:

- Data-driven cultures
- Putting together a project
- Foundational data science methods

In the next module, we will cover:

- Foundational ML methods
- Advanced ML methods

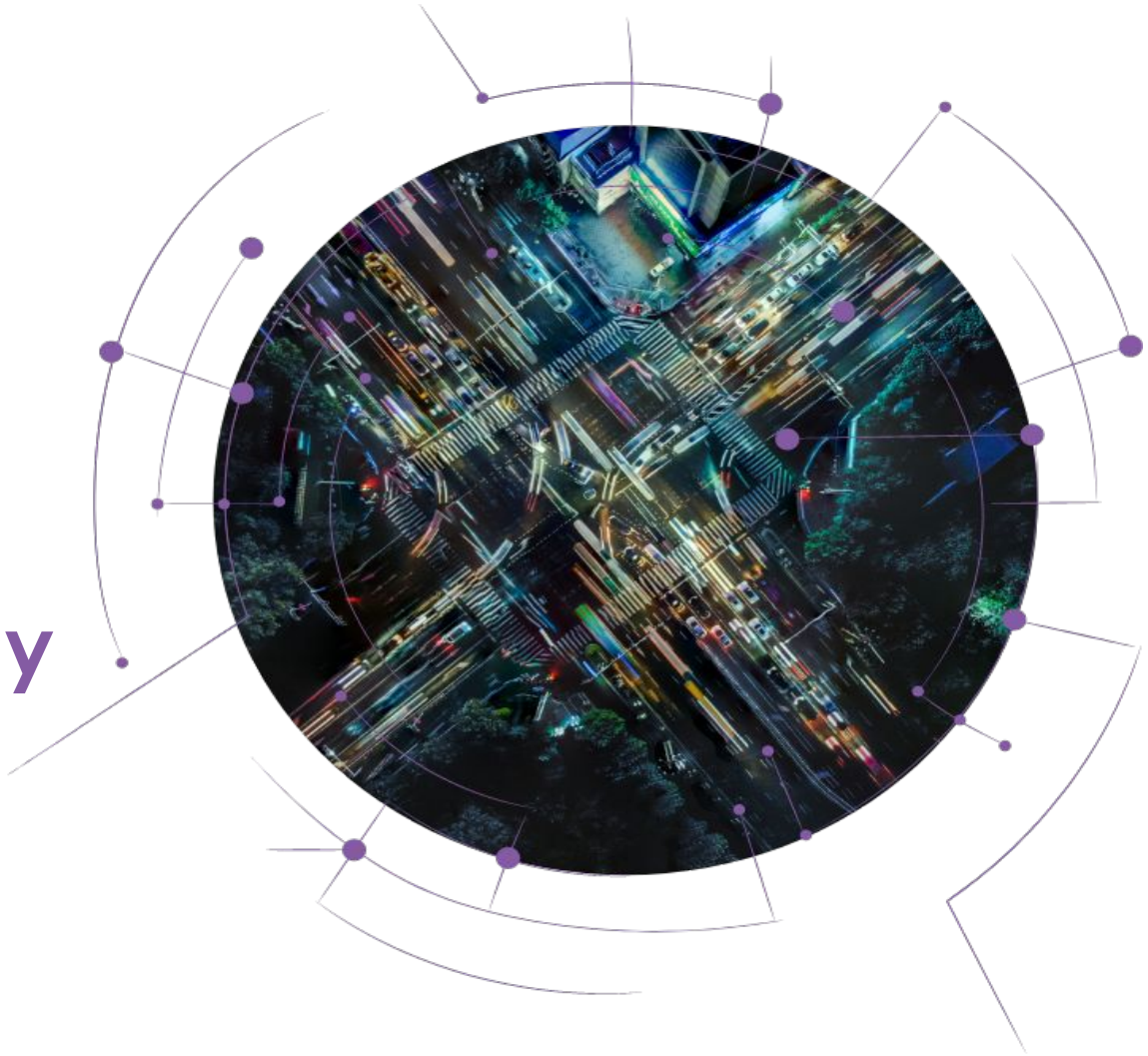
End of Day 2

Questions? Comments?

DATASOCIETY:

Fundamentals of Data Literacy

Day 3



Recap: data projects

- Yesterday we saw a few different data projects based on your needs and experiences

- In the time since the end of yesterday's session, have you had any new thoughts about:

    - your project's scope?
    - availability of data?
    - the method or process?
    - your ideal outcome or findings?

Take 5 minutes to reflect. When you're ready, leave a comment in the chat!

Agenda

Day 3

- **Foundational ML methods**
- Advanced ML methods

What is an algorithm?

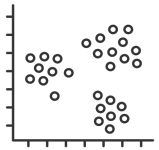
Before we go further…

- Remember that most data science projects combine a few methods to extract the full picture.

- The two big components that drive the decision for which method to use are: the question you're asking, and the data you have.

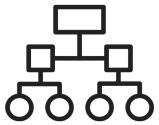
ML foundational methods

- Machine learning provides **a set of methods for data analytics**, and projects often involve strategically applying multiple methods to address the challenges.



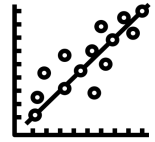
CLUSTERING

generating labels from unlabeled data



CLASSIFICATION

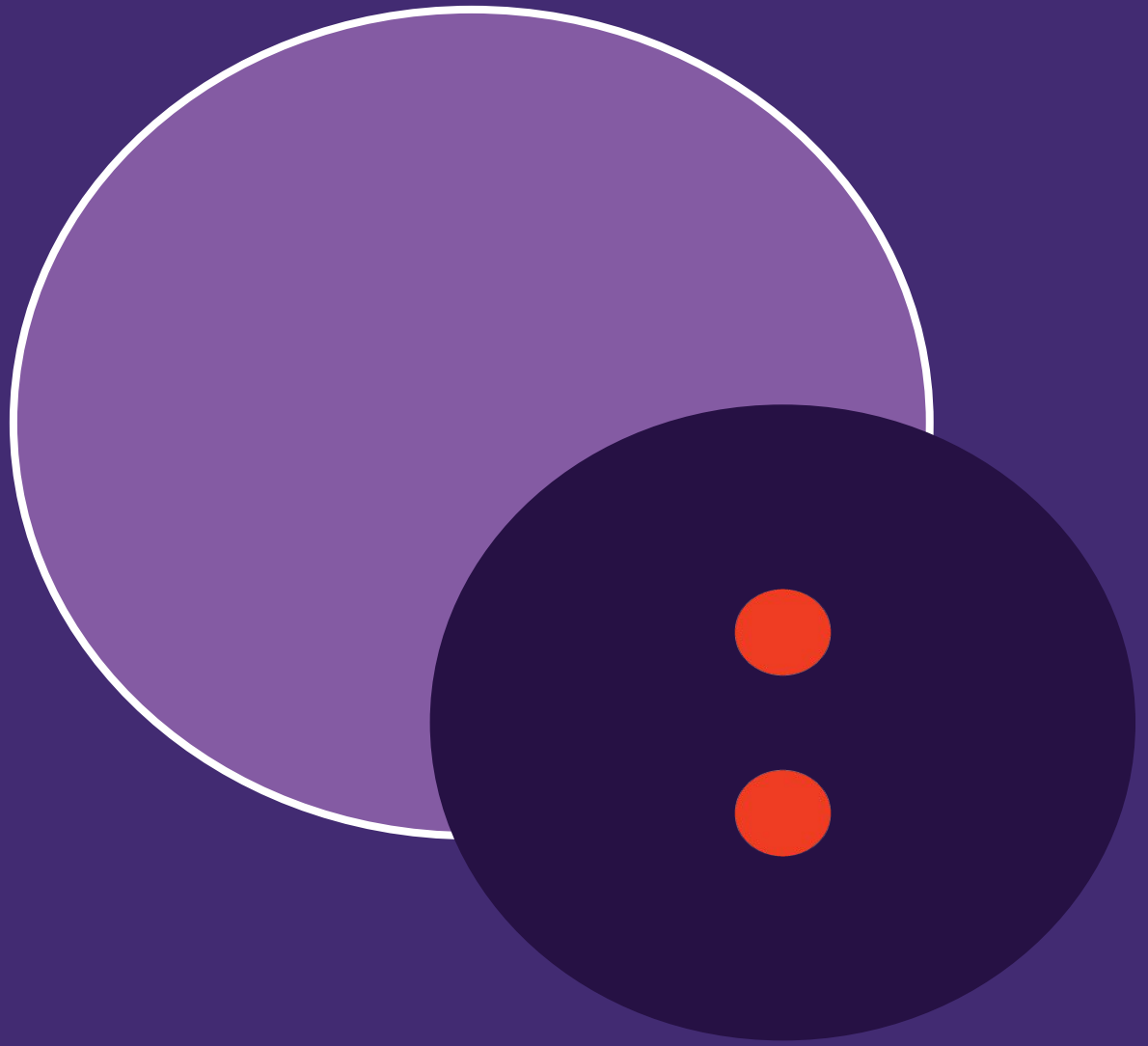
applying labels to novel data points



REGRESSION

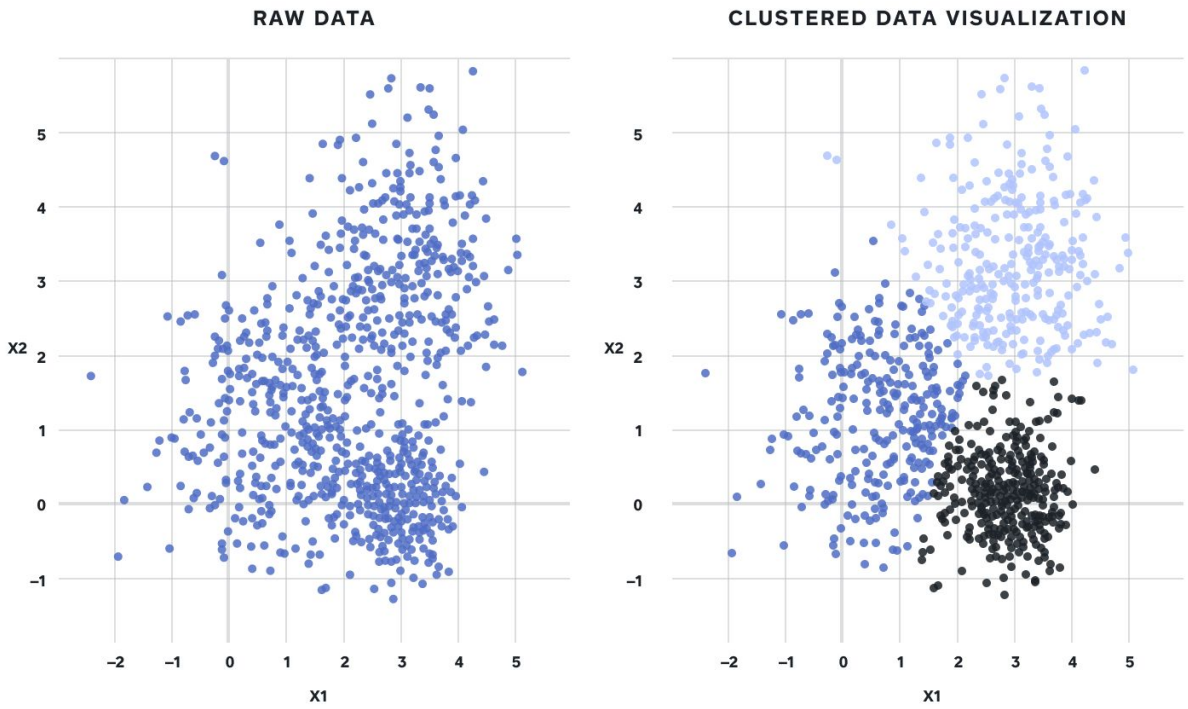
predicting the influence of various factors

Clustering



A clustering problem

- A grouping or prioritization problem like this is best solved by **clustering.**
- A clustering algorithm will arrange the data points to **generate groups,** also known as clusters, based on their similarities.
- You can use the resulting clusters to determine the suitable group for a specific health campaign and consider what resources might be required.



How can you use clustering?

Clustering answers the questions:

1. Who/what is this person/object similar to?
2. Is there a hidden pattern in the data that we can't see?
3. Are there groups of data with similar attributes?

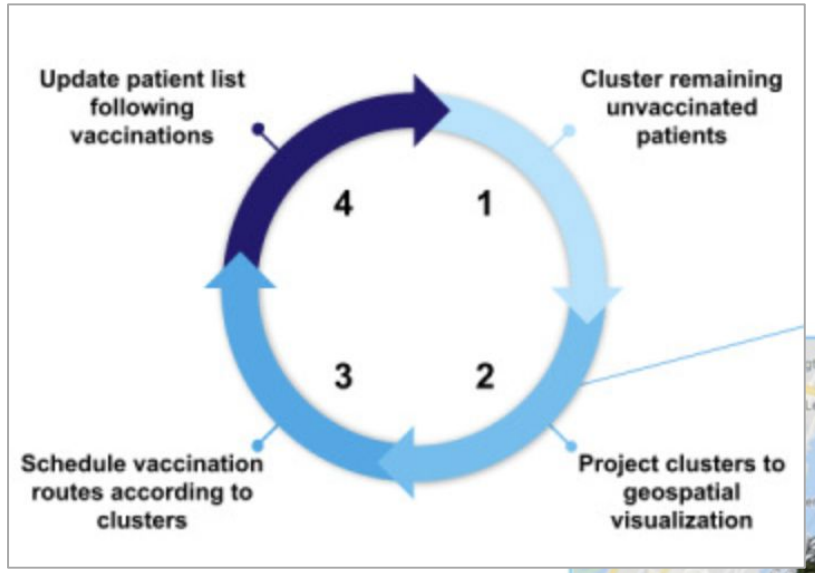
Domain knowledge is key!

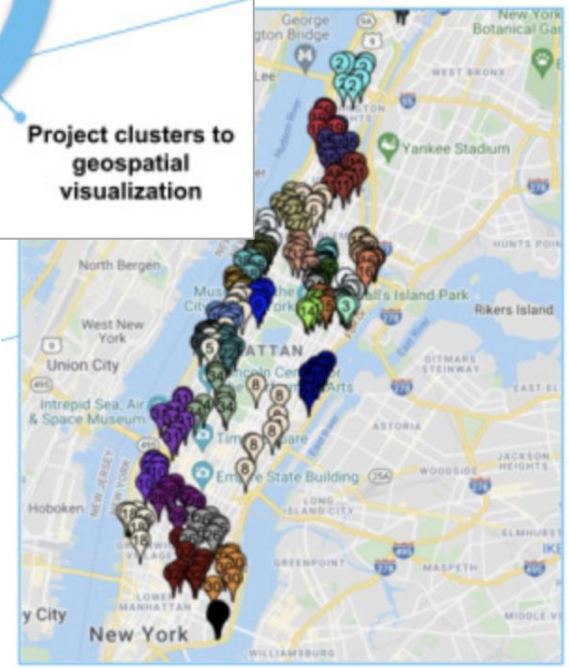
- If we know that certain policies are more effective, we can model more policies off of the similar metrics.
- If we had projects with similar objectives and outcomes, we can consolidate ones that overlap to streamline progress.

Example: COVID-19 immunization campaign

- A home-based primary care program used Python to cluster patient addresses using coordinate points.
- These data points were clustered into groups of 12-15, and each cluster was assigned a provider to administer the COVID vaccine within their assigned geographic space.
- This process was repeated weekly with the remaining unvaccinated patients.

Source: Using Machine Learning to Efficiently Vaccinate Homebound Patients Against COVID-19: A Real-time Immunization Campaign (linked)





Common pitfalls with clustering

- Clustering algorithms don't scale well to large datasets
- Different data types need to be formatted correctly (i.e., mixing categorical data with numerical data may not be the best way to find similar points).
- Make sure you use the right clustering model for the data!

Questions to ask

- How was the distance measure identified?
- Did you scale the data appropriately?
- How many clusters do you expect or want? Why?
- What can we learn from the groups that the algorithm identified?

Recap: when should you use clustering?

- **Use clustering when:**

  ✓ You have an unlabeled dataset

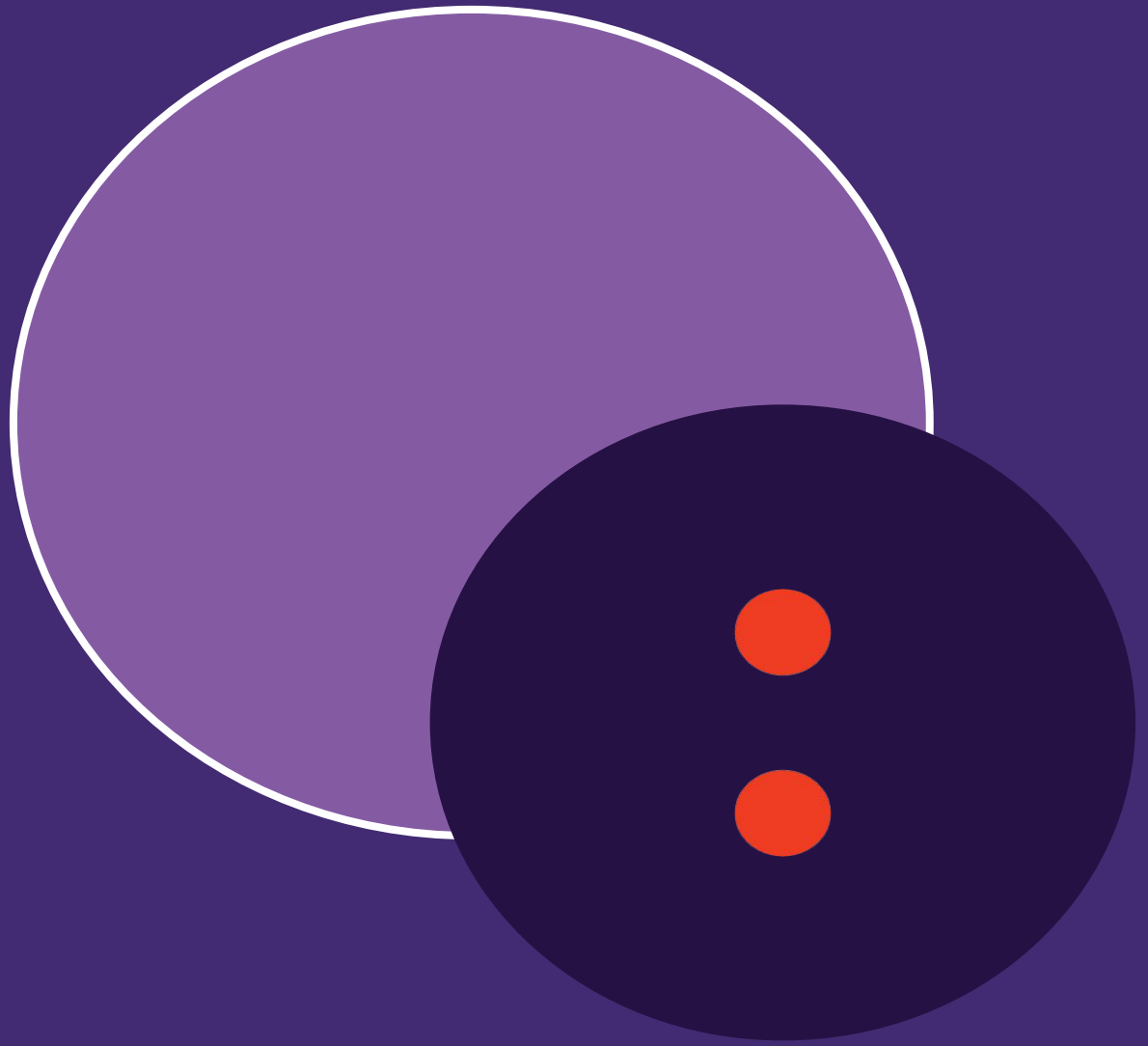
  ✓ The dataset has multiple attributes

  ✓ You need to identify patterns in your data

  ✓ You need to find groups in your data

Break

Classification



Chat question

- What data would you need to answer this question?
- What kinds of relationships might you expect to see?
- What domain knowledge is most important for interpreting the results?

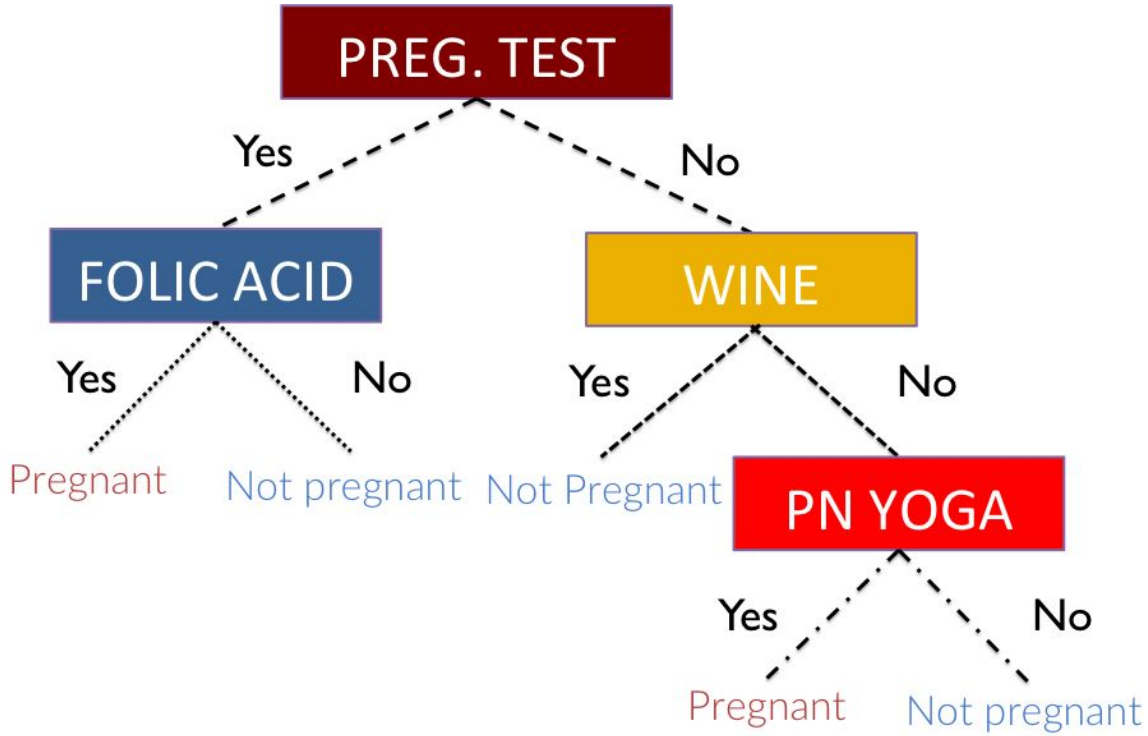
"Out of all the web traffic to our website, how can we tell which visitors are bots and which are humans?"

A classification problem

- A detection problem like this can be solved using **classification.**
- A classification algorithm will **sort visitors into pre-existing categories,** specifically “bot” or “human”
- The algorithm will need to be trained on data that clearly demarcates these two categories:
    - You could use labelled data based on typical hallmarks of bot visitors (duration, repeated visits, fake conversions, refilling / refreshing)
    - You could also use **clustering** on your web traffic data to see what patterns or fine-grain categories emerge!

Classification

- Classification is a type of **supervised** machine learning.
- Assigning new points to classes is based on their similarity to existing data points with known class assignments (i.e., a category or behavior pattern).
- Models should be retrained and labels updated as data and needs change.

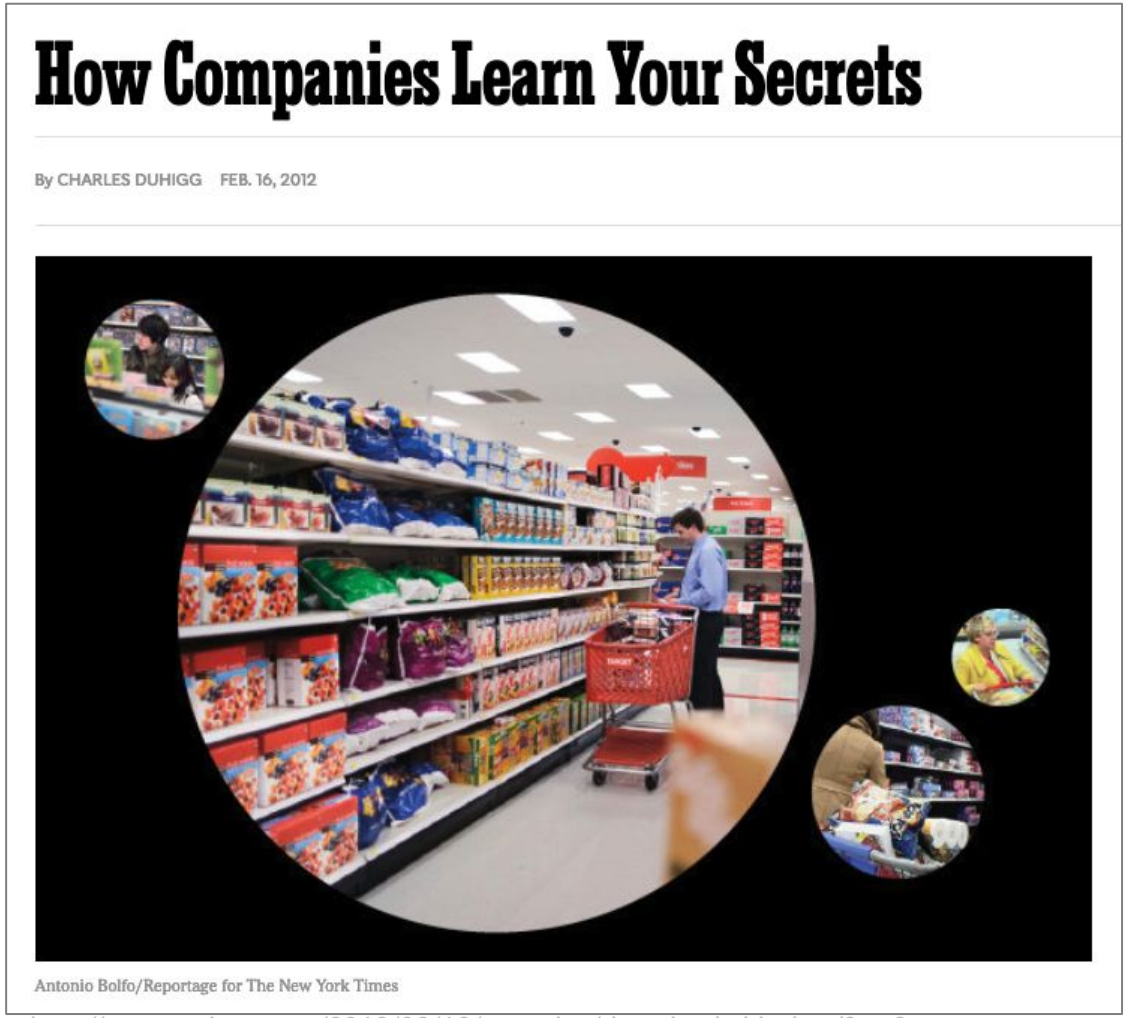


Example: predicting pregnancy

- In 2002, Target implemented data analytics to analyze buying patterns in customers.

- New parents often get bombarded with advertising offers, so Target wanted a way to anticipate who is expecting in order to get ahead of the competition.

- They were able to predict pregnancy of their customers based upon their purchases and sent out targeted coupons.



How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

Antonio Bolfo/Reportage for The New York Times

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0

Chat question

What ethical implications might Target's pregnancy predictions have raised?

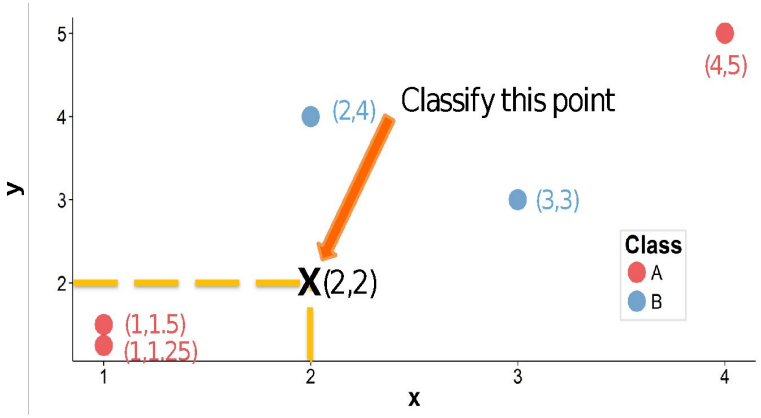
How can you use classification?

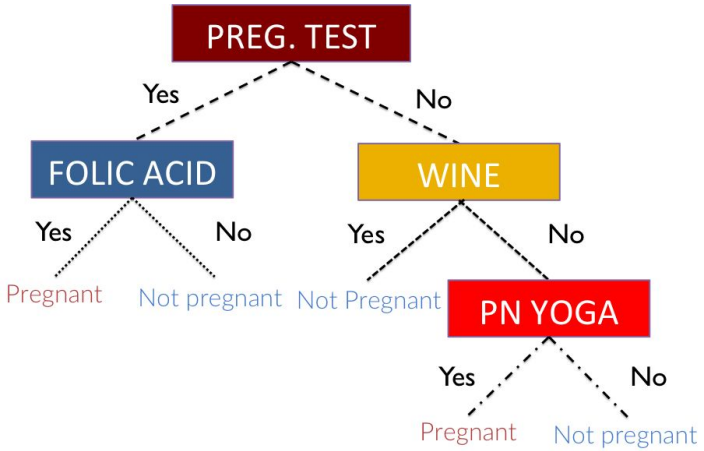
- Classification answers the questions:
  1. Which is the probability of an object / person being in a particular group?
  2. What category is this person / object in?
  3. What is this person / object most similar to?

- Domain knowledge is key!
  - If we know that certain policies are most likely to be successful, we can predict if new policies will also be successful
  - If we see behavioral outcomes based on certain decisions, we can predict similar behaviors

Common classifiers

- k-Nearest Neighbors (KNN) – assumes that similar things exist in close proximity; classifies a data point based on how its neighbors are classified
- Decision tree – uses a tree-like graph or model of decisions and their possible consequences to classify data
- Logistic regression – determines the probability of a data point to be part of a certain class or not





Recap: when should you use classification?

- **Use classification when:**

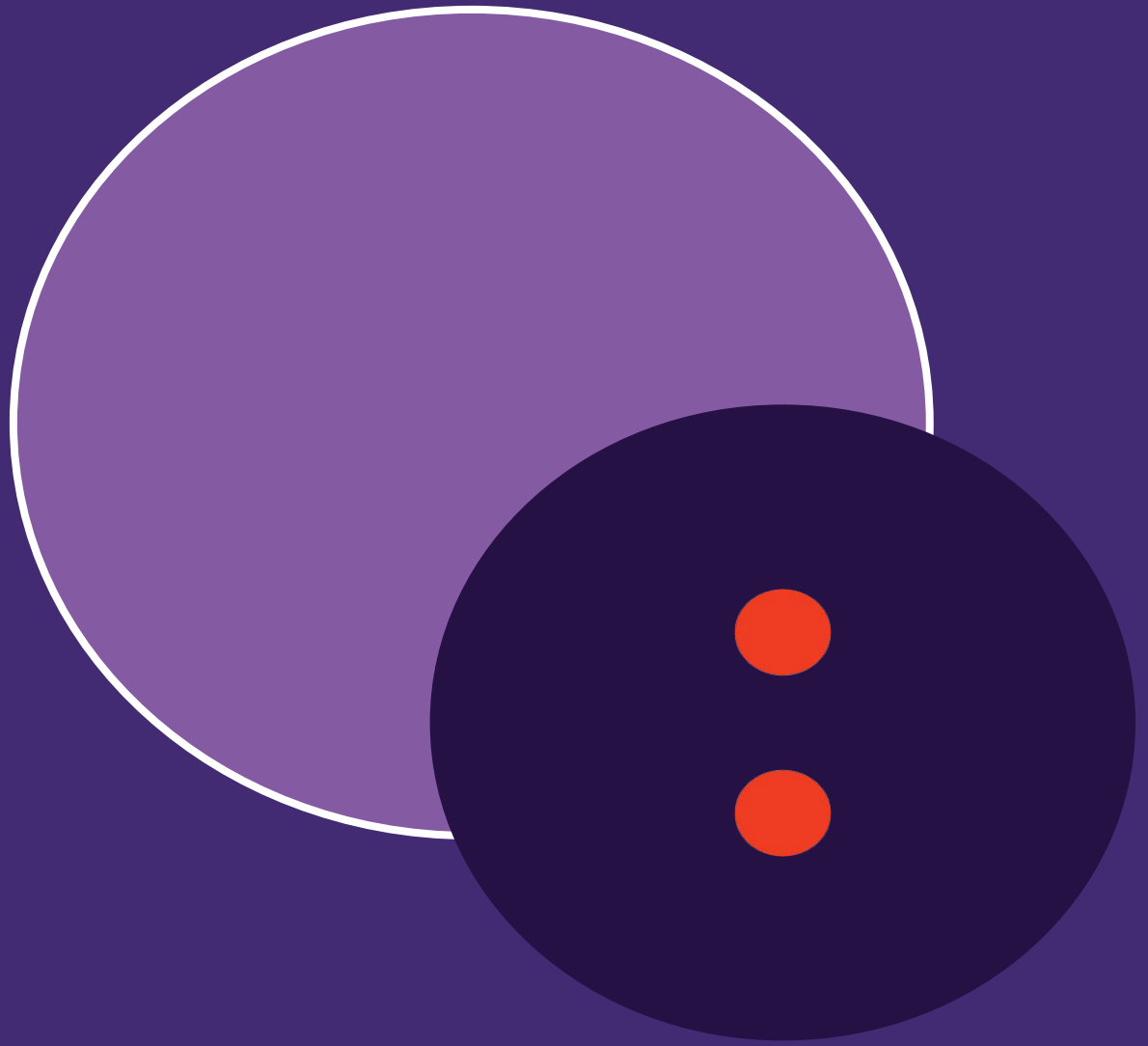
  ✓ You have a labeled dataset

  ✓ You want to predict group assignments

  ✓ You want to predict behaviors / events

  ✓ You want to identify important attributes

Regression

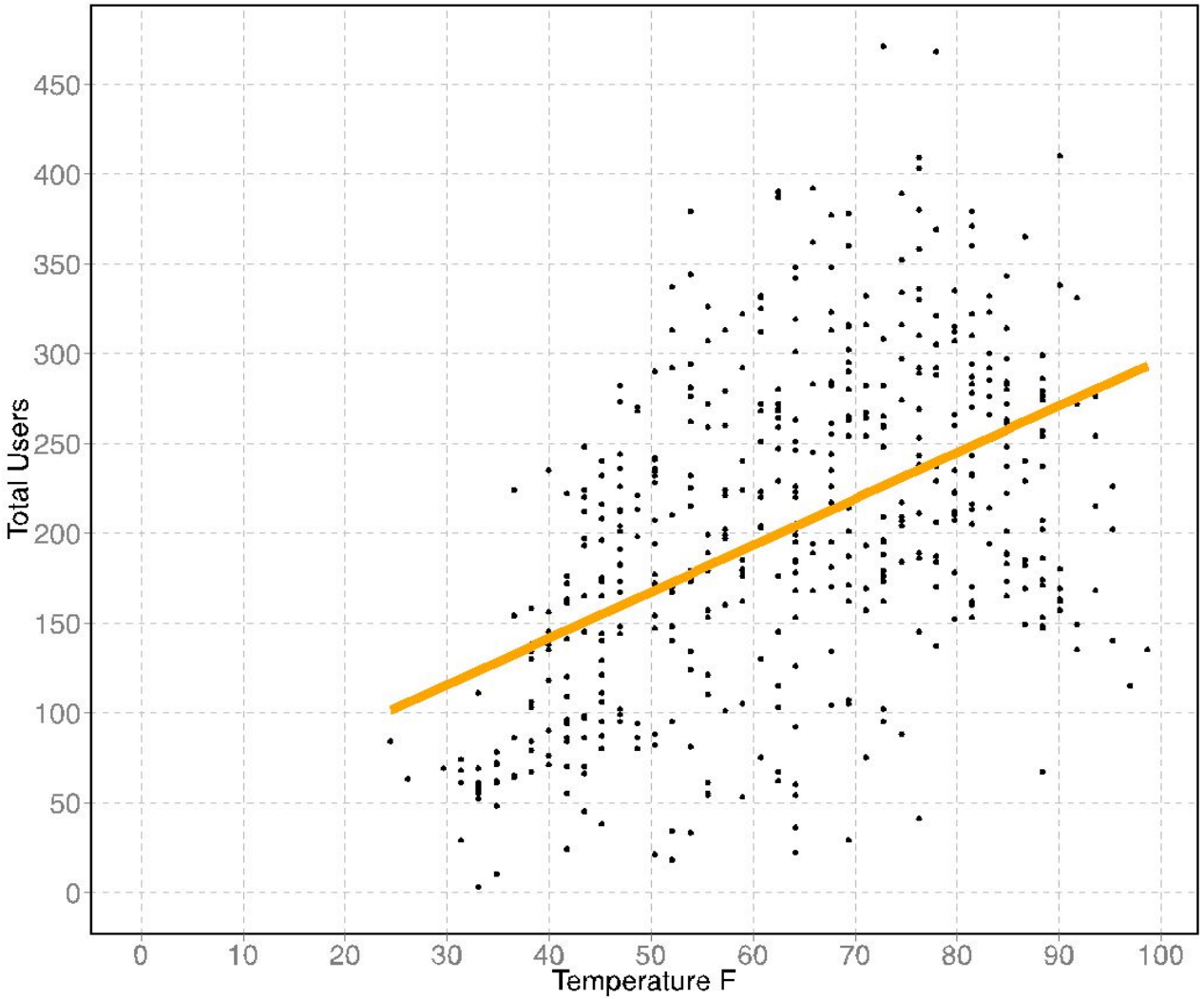


A regression problem

- Regression models are statistical tools to examine and **quantify the relationship** between a variable of interest and one (or more) explanatory variables:
    - The variable of interest, like *post bid success*, is considered a **dependent variable**
    - It depends, or is subject to influence by, the various factors that might explain it, or the **independent variables**
    - Once you know how a dependent variable depends on certain factors, you can determine which independent variables are key **predictors**

Regression

- Regression is a type of supervised machine learning.
- The example to the right attempts to determine the degree to which a single independent variable (temperature) affects the independent variable (total users).
- Regression can quickly become complex given multiple factors, and if factors must be numerically encoded.



Use case: predicting city movements

- There are over 500 bike-sharing programs around the world with over 500,000 bikes.
- Automated systems track numerous data points providing a treasure trove of data about the mobility of residents.
- Data can be used to forecast the number of bikes required and adjust pricing based on demand.

Chat question

What factors do you think might drive demand for bike-share use?

How can you use regression?

Regression both explains predictors (inference) and enables forecasting (prediction) by answering the following questions:

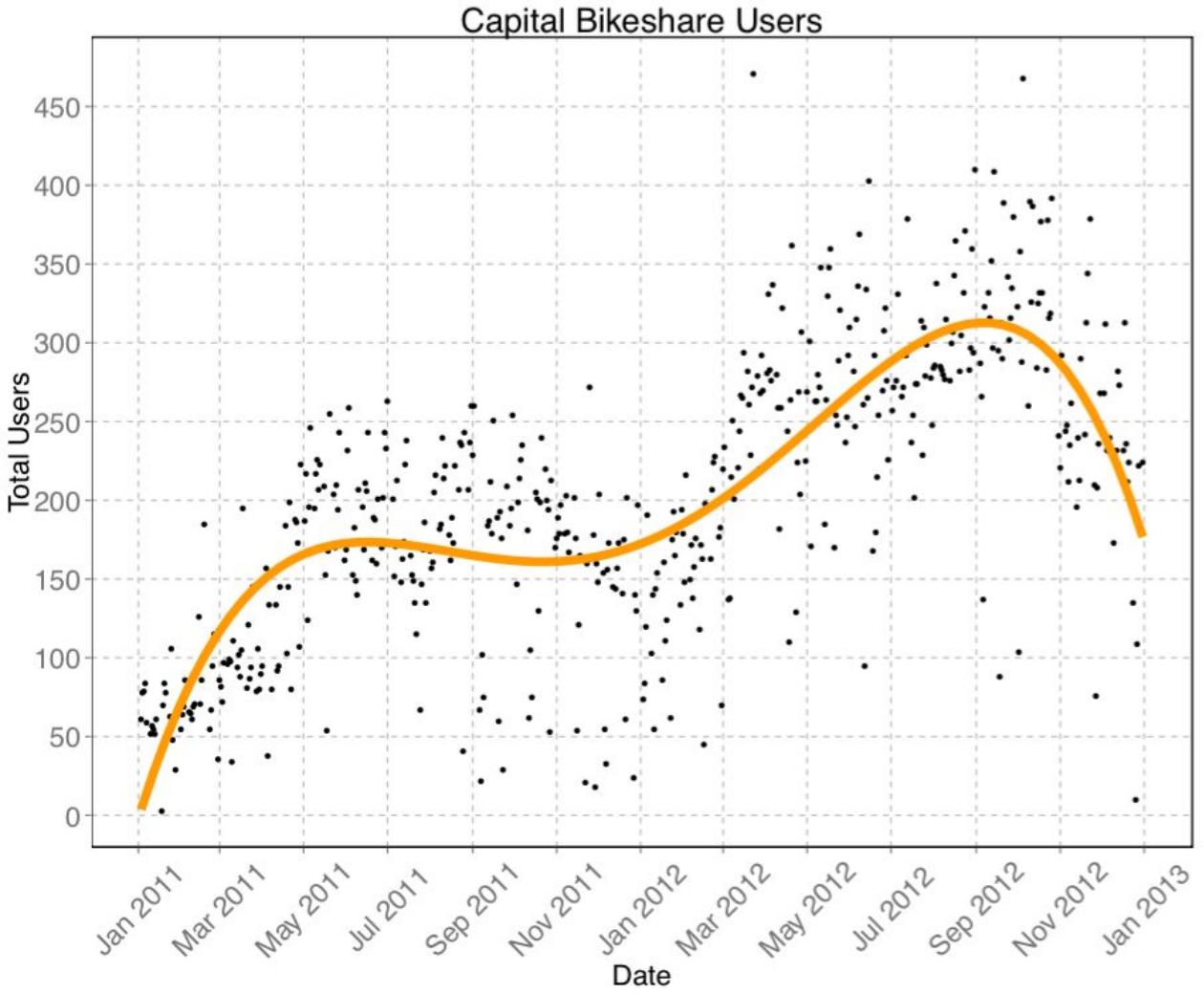
1. Which factors matter most?
2. Which can we ignore?
3. How do those factors interact with each other?
4. How certain are we about all of these factors?
5. What happens to the outcome if we change a factor?

Domain knowledge is key!

- We can predict political instability in countries
- We can predict how tourism season affects a country's economy

Common regression techniques

- Different regression techniques attempt to best fit a line to the data in different ways.
  - Linear regression
  - Polynomial regression
  - Lasso regression
  - Ridge regression
  - Nonlinear regression
  - Binary logistic regression



Questions to ask about regression

- How well do we understand the underlying data distribution?
- Did you identify any outliers? Were they significant? Did you remove them?
- Are you sure each of the independent variables really is independent? Are we double counting anything?

Recap: when should you use regression?

- Use regression when:

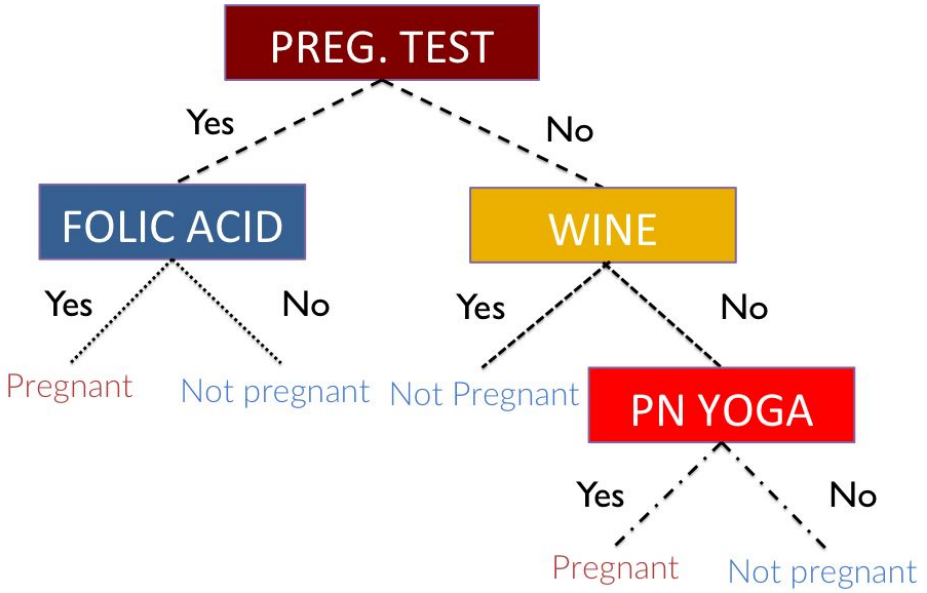
  ✓ You have a labeled dataset

  ✓ You want to predict trends

  ✓ You want to anticipate needs or shortages

Chat question

Do you think the decision tree depicts a classification method?



Chat question

Would you use clustering, classification, or regression to anticipate what candidate a person would vote for?

Break

Agenda

Day 3

- Foundational ML methods
- **Advanced ML methods**

ML advanced methods families

- Advanced machine learning methods build upon fundamental concepts to tackle complex problems and achieve higher accuracy and efficiency.

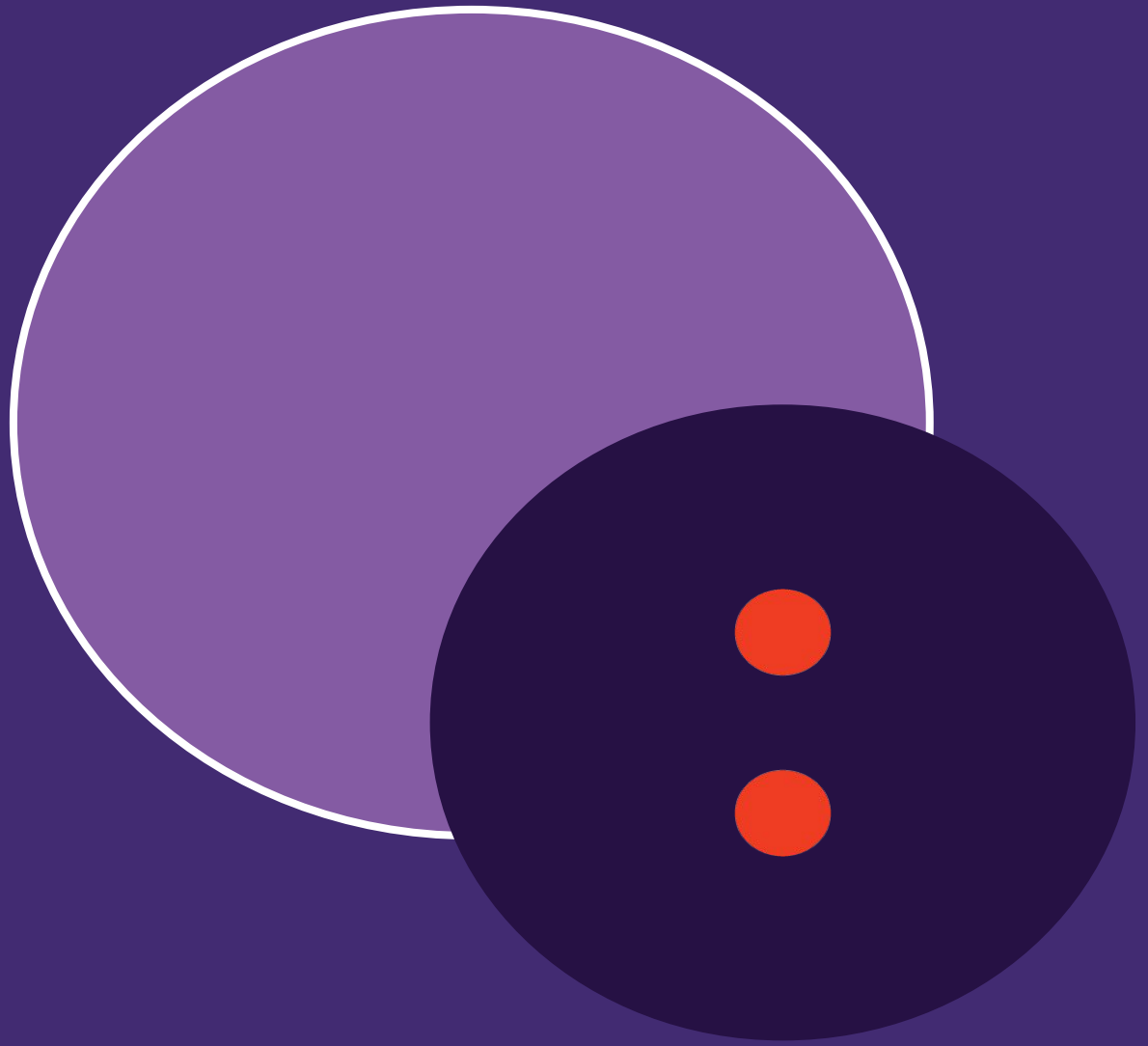
TEXT MINING

inference and prediction based on textual data

GRAPH ANALYSIS

finding patterns within a network to show connections

Text mining



Chat question

- What data would you need to answer this question?
- What kinds of relationships might you expect to see?
- What domain knowledge is most important for interpreting the results?

"A lot of people seem to be posting about this event. How can we measure how they feel about it?"

Text mining example

- **Text mining** helps analyze **unstructured text data** to extract meaning and identify patterns.
- For example, researchers used text mining to analyze tweets about different smartphone brands, like #iPhone and #Samsung, to understand consumer sentiment and trends.
- This allowed them to extract valuable information from the unstructured text data of Twitter.

Document info

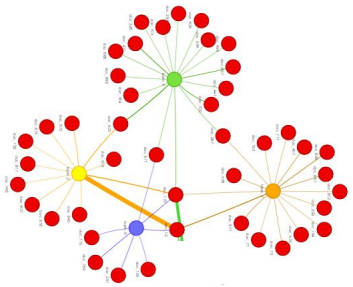
ID	Title	Position	Key word
123	The Catcher in the Rye	6	Testing
334	Pride and Prejudice	4	behind
432	Brave New World	2	all
5854	Alice's Adventures in Wonderland	5	wonder
153	The Adventures of Huckleberry Finn	34	experience
64	To Kill a Mockingbird	7	sky
737	1984	1	willingly
847	Wuthering Heights	10	bio
790	Jane Eyre	12	society
908	The Lord of the Rings	3	real
8664	The Sun Also Rises	9	every

Text mining branches

Text mining

Entity extraction

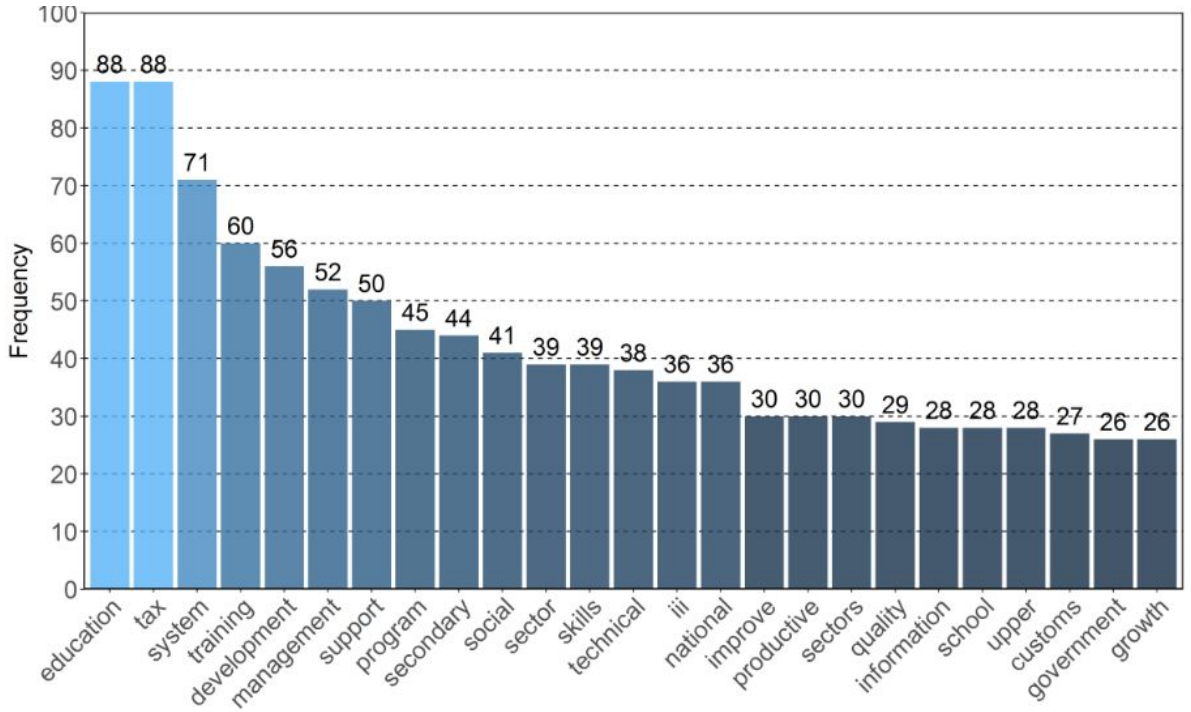
Document classification

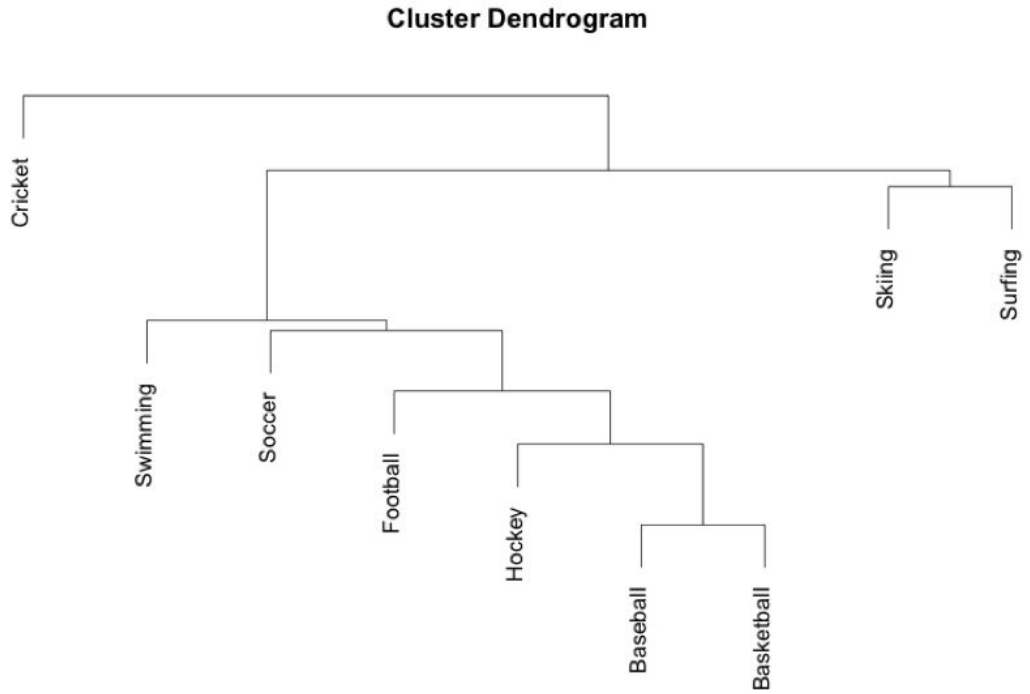


Sentiment analysis

Entity extraction

- Use entity extraction when you want to get an overview of the themes and topics in documents.
- Measure word frequency and word co-occurrences.

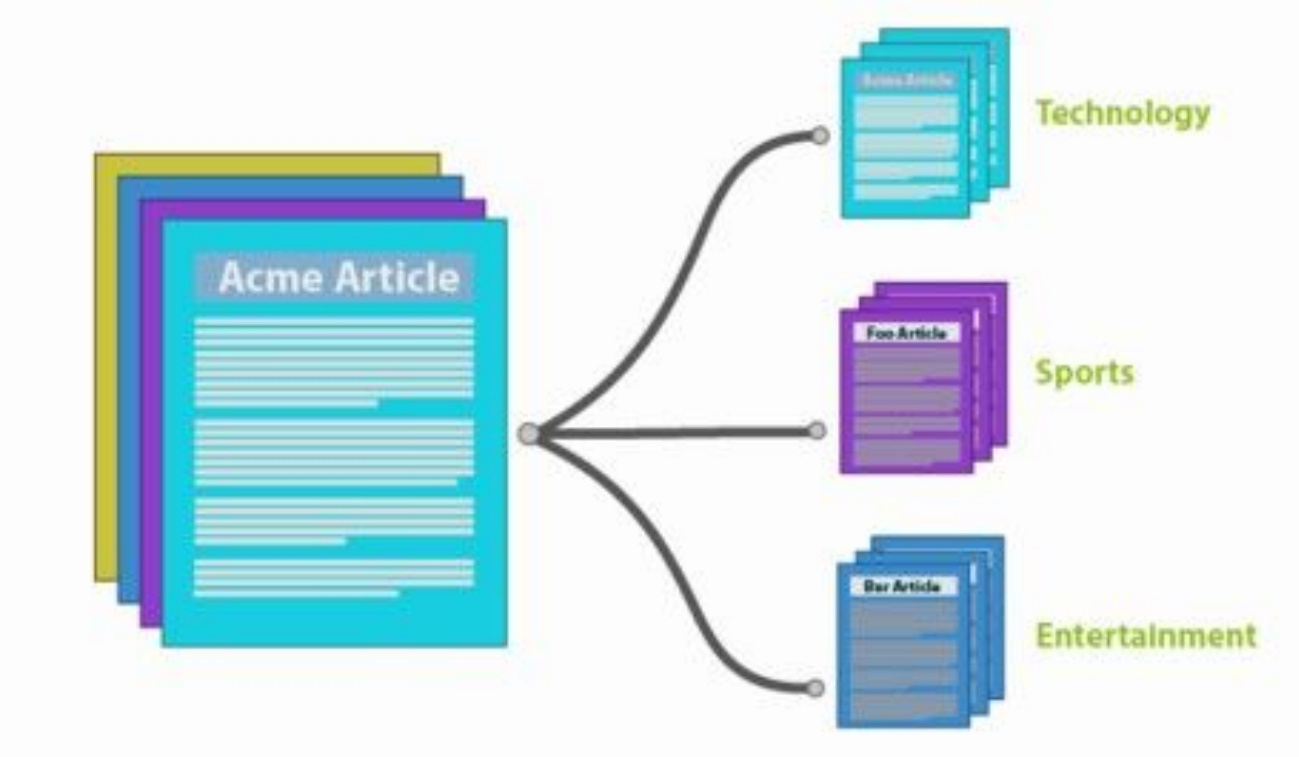




Document classification

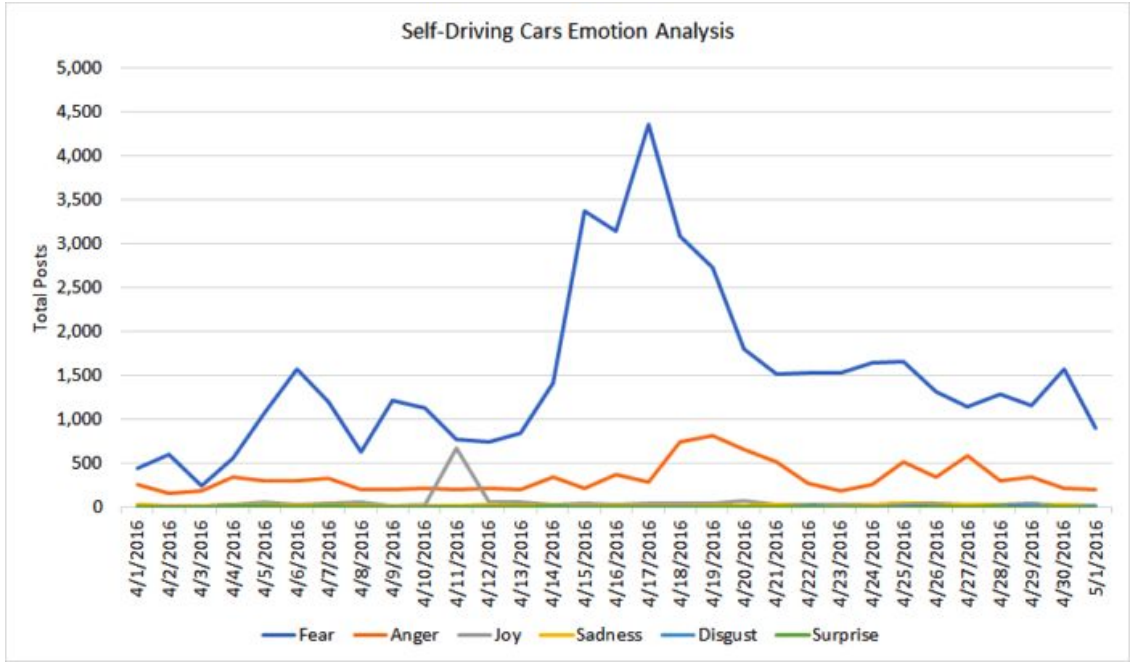
- Use document classification wher you want to sort through documents and identify groups o similar articles.

- Based on similarity of topics / other metrics.



Sentiment analysis

- Use sentiment analysis when you want to understand the emotions and overtones of documents.
- Use reference dictionaries to identify positive / negative words.
- Natural language processing (a similar branch) doesn't focus specifically on sentiment, but rather on the meaning of the document.



What events might have driven the trends in emotion depicted above?

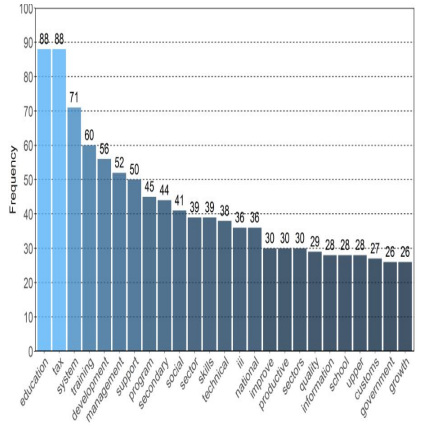
Text mining process

Scrape / collect

Clean & organize

Index	Word	Freq	%
A	Apple	5	20
B	Book	7	28
C	Cat	13	52

Visualize

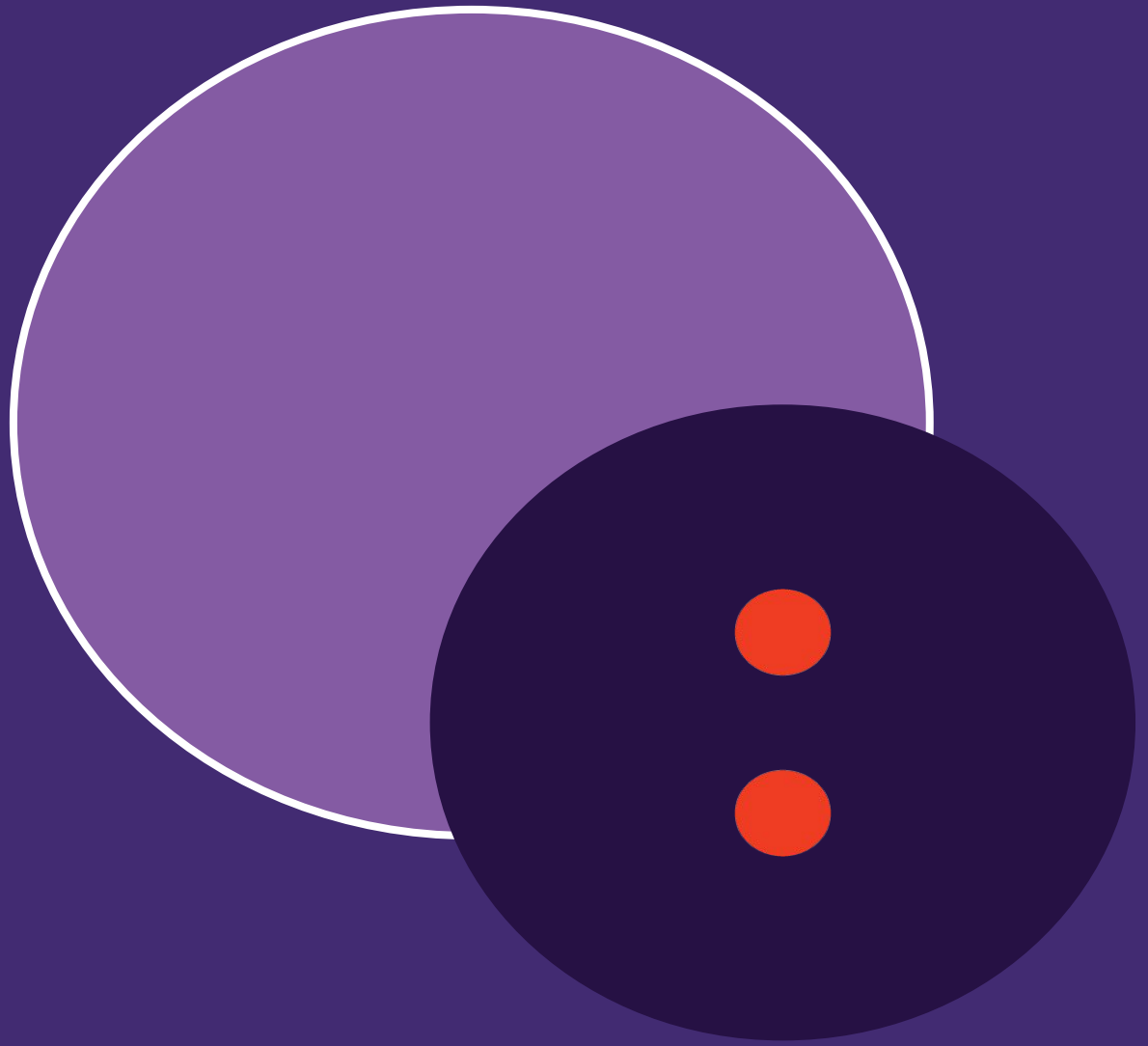


Analyze

Questions you should ask

- How does the model take sarcasm / irony / colloquialisms into account?
- Is there an existing library of reference words that can assist you in text mining?
- Does that reference library include misspellings, alternate versions of words, symbols, different parts of speech or compound terms?
- How do the topics change over time?

Graph analysis



Chat question

- What data would you need to answer this question?
- What kinds of relationships might you expect to see?
- What domain knowledge is most important for interpreting the results?

"In the event of a flood in the region, which roads are the likeliest evacuation routes?"

Graph analysis

- **Graphical analysis** uses visual representations like graphs and charts to **identify patterns, trends, and relationships between variables** within a dataset.
- For example, Google Shopping uses your searches and activity to identify related products you may be interested in.
- This uses patterns and relationships to make personalized recommendations.

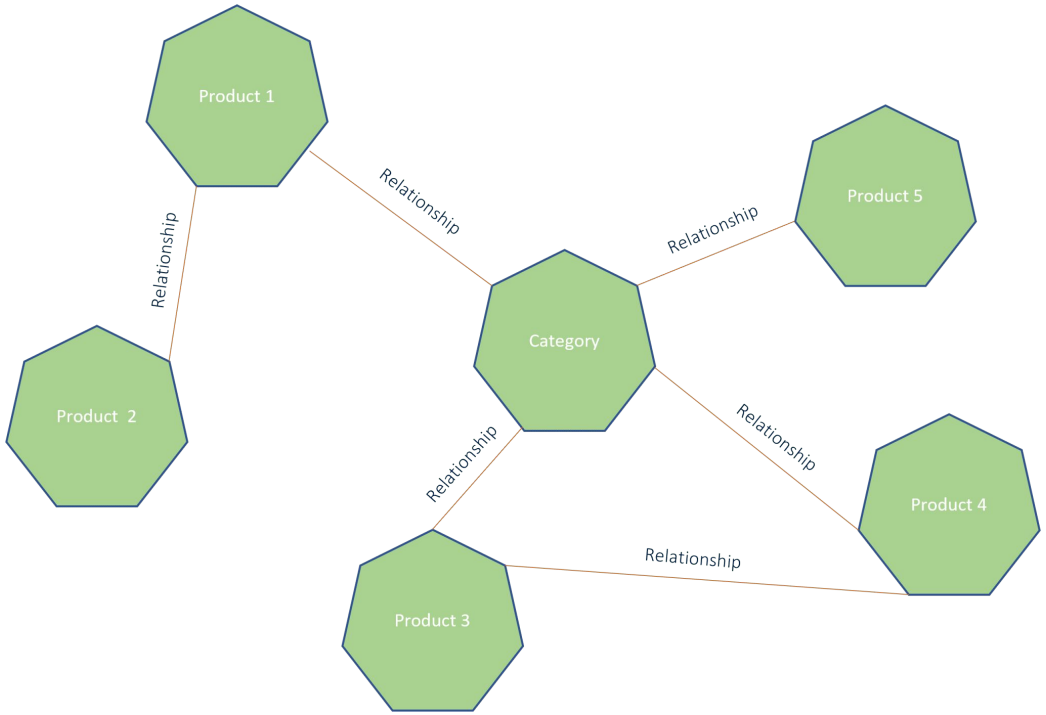


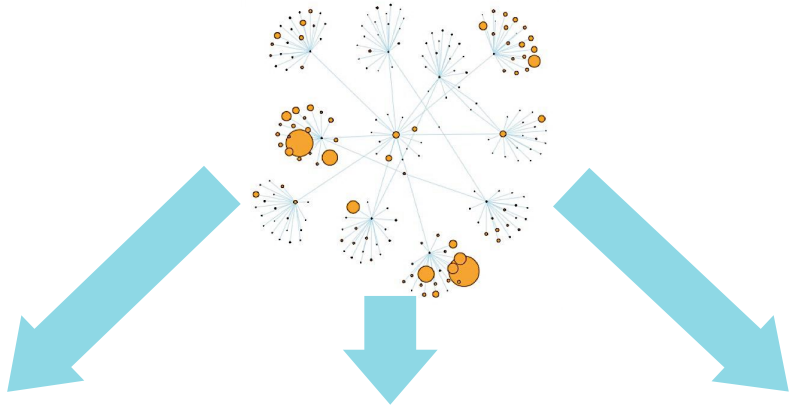
Image source: Kopp Online Marketing Consultant

Questions graph analysis answers

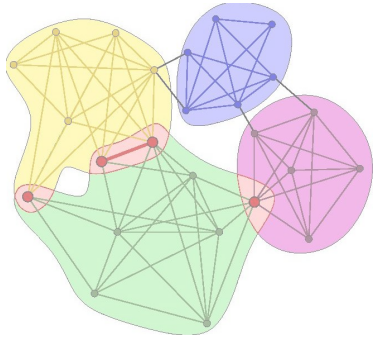
- What communities exist within a target population?
- How will a message / disease spread through a population?
- Which individuals are most trusted in a community?

Types of graph analysis

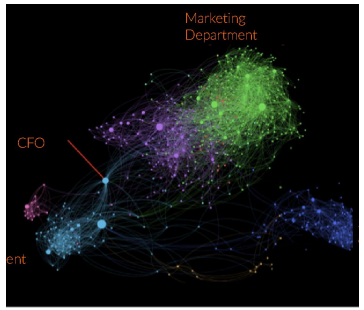
Graph analysis



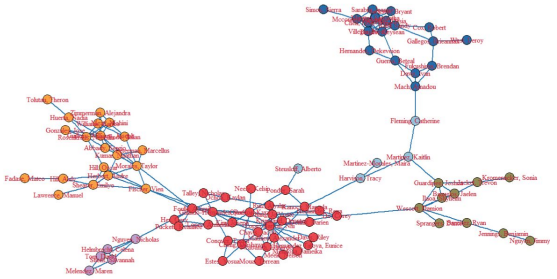
Community detection



Centrality metrics



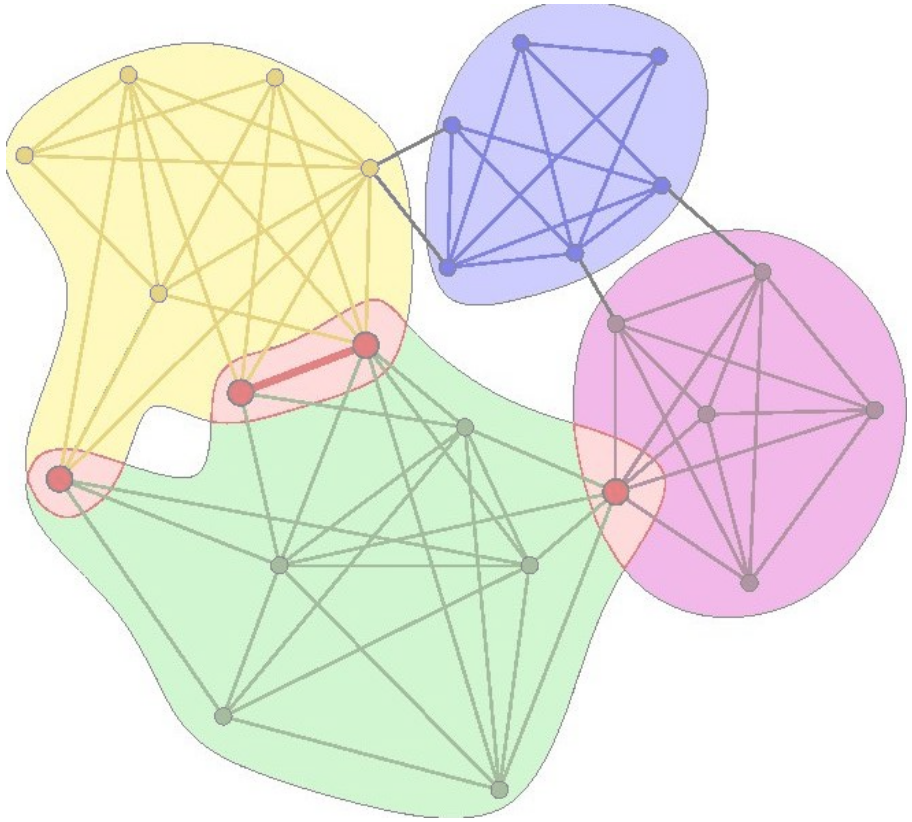
Social media



Community detection

- Use community detection when you want to dive into your network to find new communities and groups.

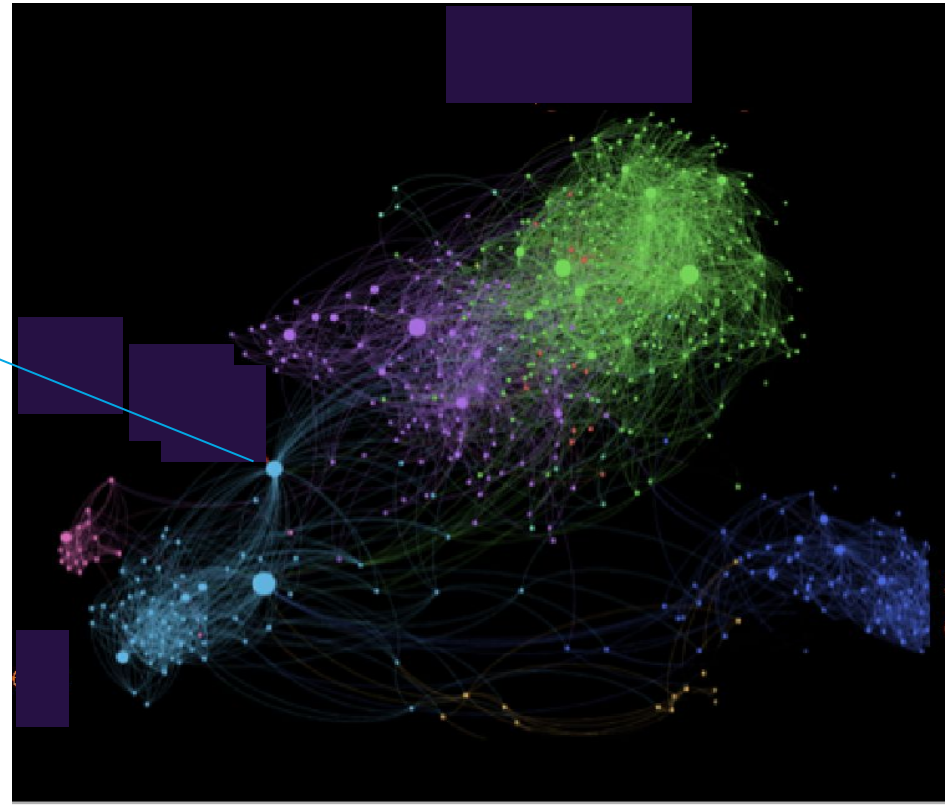
- Identifies groups of individuals / nodes that belong together; can detect latent connections and communities.



Centrality metrics

- Use centrality metrics when you want to look at an overview of a network and identify key nodes.
- Identifies the most important nodes, most central nodes, shortest paths, etc.

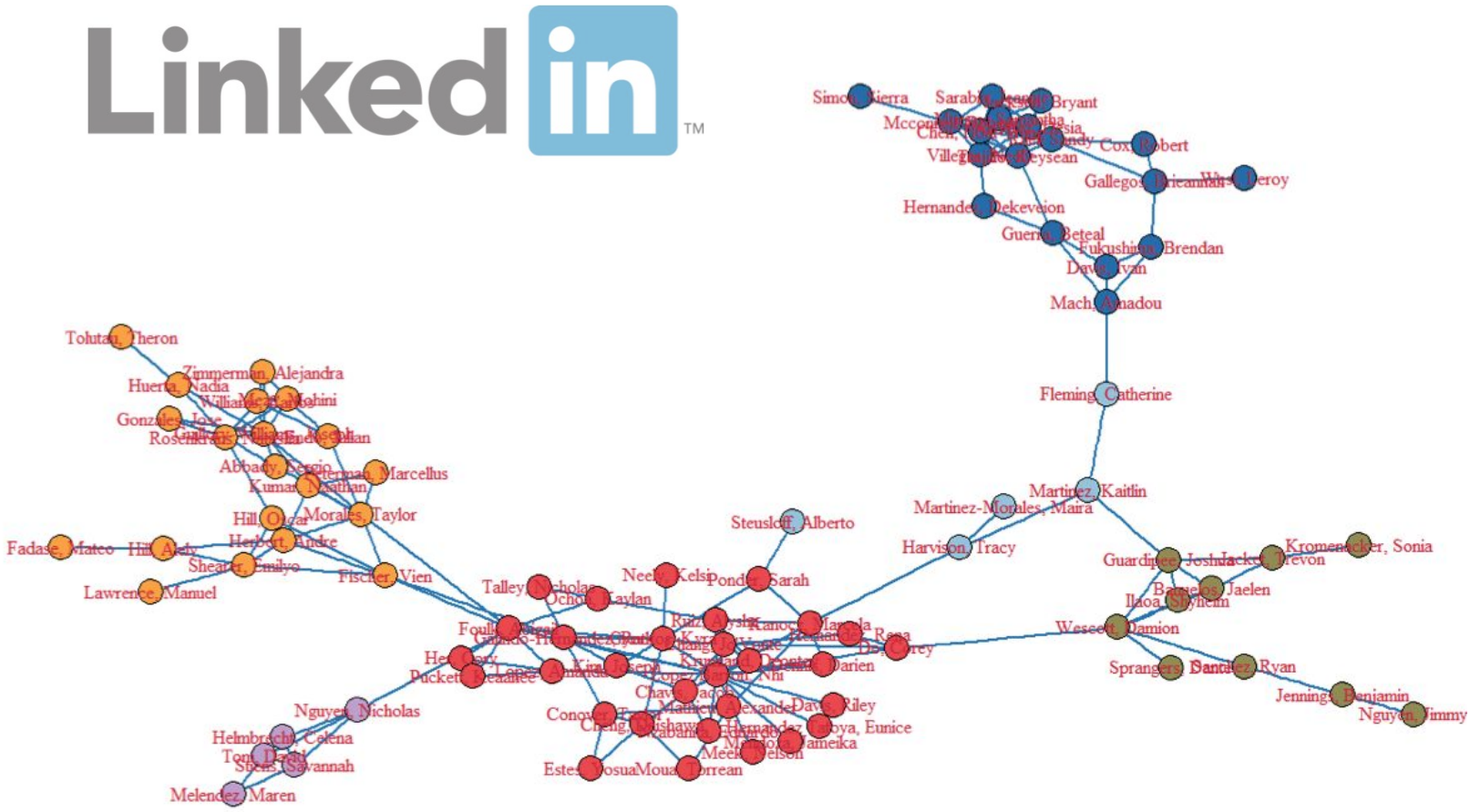
This email network shows how a company communicates.



Social media

- Use social media when you are using data from social media platforms.

- Identifies how an idea travels across social media platforms and how individuals are connected.



Questions you should ask

- What aspect of the relationship are you most interested in (i.e., who is the most connected, who has the strongest connections, who is most important)?

- Does the data you're using account for a large amount of a relationship? How much is in the numbers versus not collected?

- What metrics did you use to evaluate the proximity between nodes / communities?

Summary

In this module, we covered:

- Foundational ML methods
- Advanced ML methods

In the next module, we will cover:

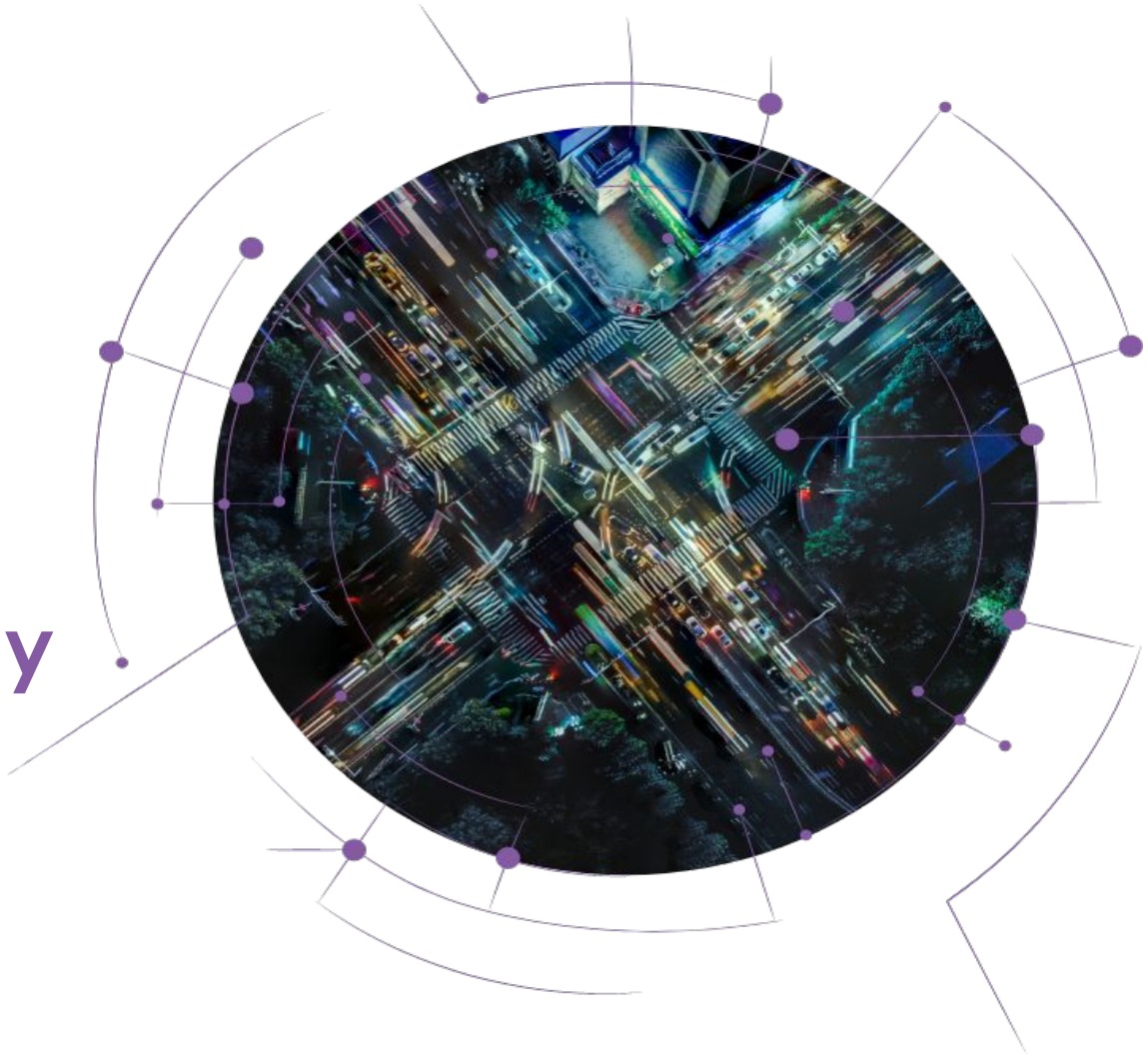
- AI methods
- Refining your data project
- Intro to data visualization
- Best practices in data viz

End of Day 3

DATASOCIETY:

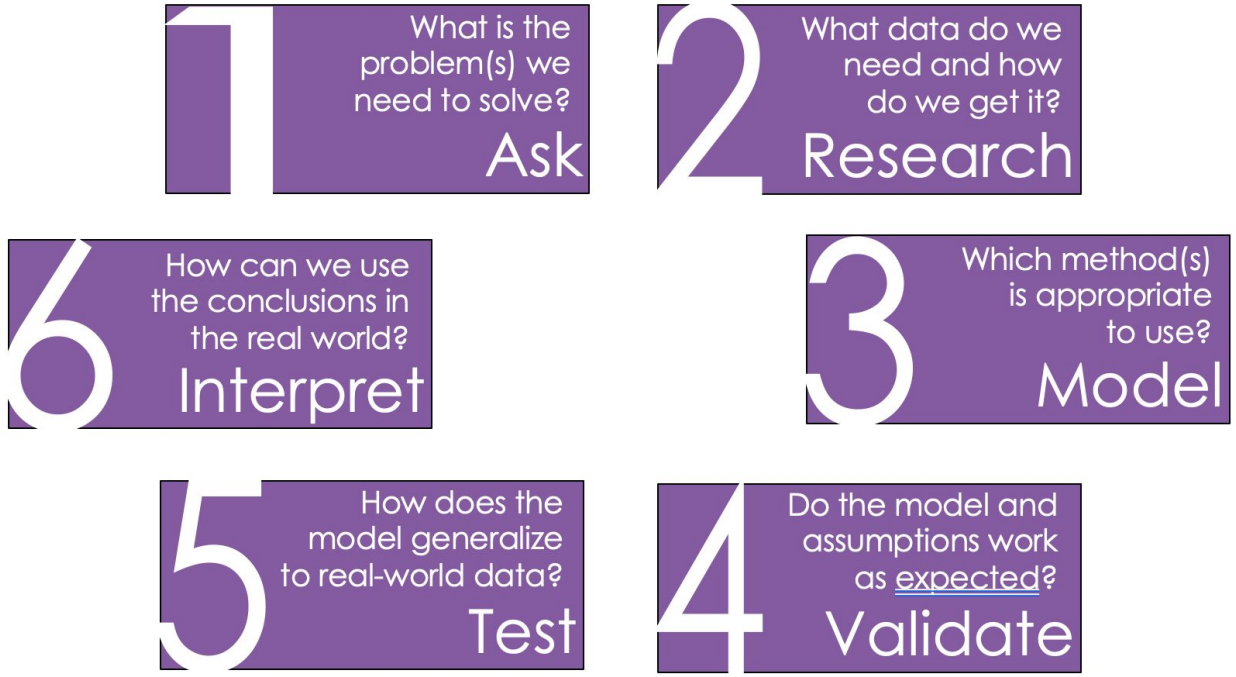
Fundamentals of Data Literacy

Day 4



Recap

- In our last session, we talked about a few different machine learning methods for carrying out data science projects.
- We also mentioned that there were some cases where we need more advanced techniques.
- Any questions before we get going?



How machines learn

Agenda

- AI methods
- Refining your data project
- Intro to data visualization
- Best practices in data viz

What are neural networks?
What are the limitations to using neural networks?

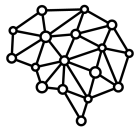
Agenda

Day 4

- **AI methods**
- Refining your data project
- Intro to data visualization
- Best practices in data viz

AI methods

- There are three methods that play a key role in **building AI technologies**.



NEURAL NETWORKS

Modelled on the human brain, these algorithms train on large amount of data and have built in mechanisms to reduce modelling errors

DEEP LEARNING

Comprised of neural networks that have more than 2 hidden layers, it is commonly used in text and image prediction

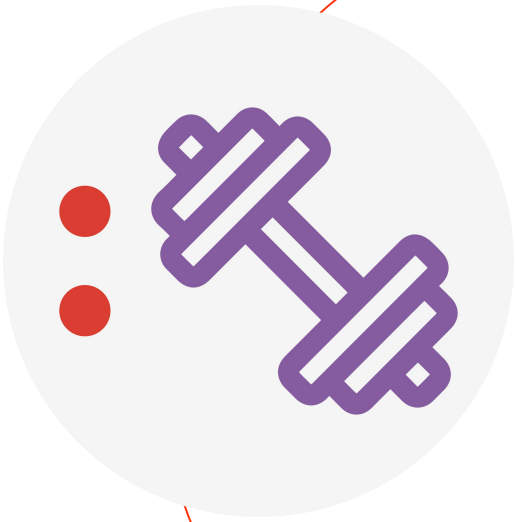
TRANSFORMERS

Designed to process sequential Natural Language Processing data, these neural networks have an added encoder and decoder layers

Activity: field trip

- Visit https://quickdraw.withgoogle.com/.
- Click the "Let's Draw!" button and play a round (6 drawings).
- At the end of the round, visit the data to see why guesses were made. Also, make a note of how many of your drawings were guessed correctly.

Note: A clickable link is available on page 16 of the participant guide.



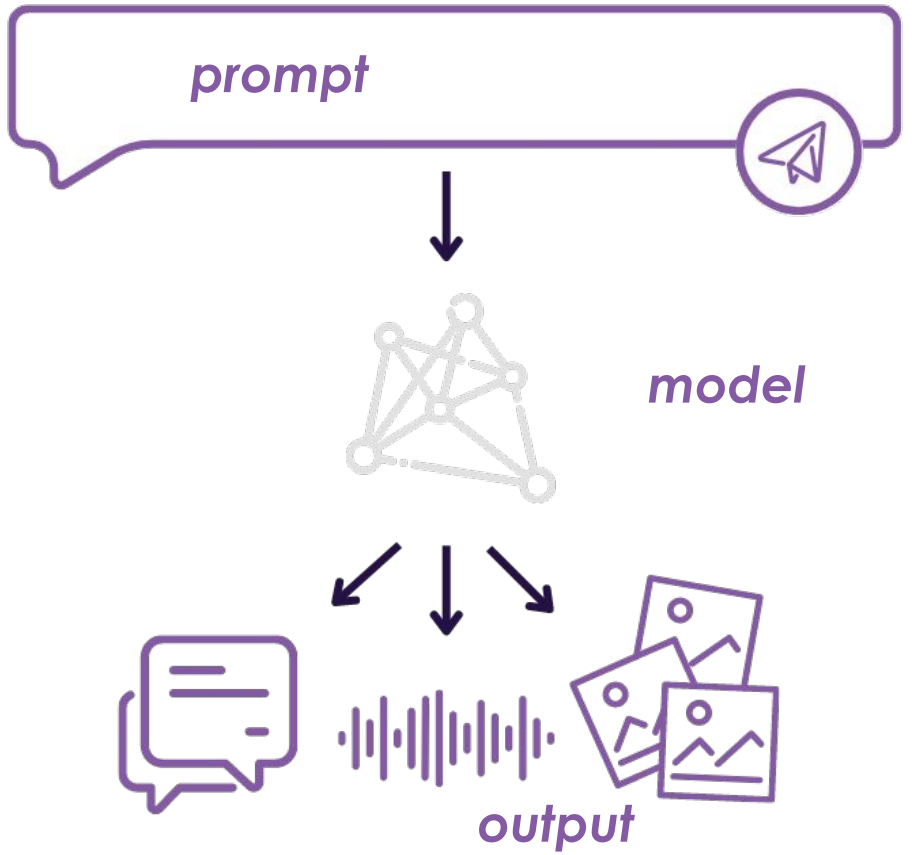
AI method for facial recognition

- Imagine a computer trying to recognize a face.
- Unlike our brains, which perceive faces holistically, a computer sees only a grid of pixels, each with a specific brightness value.
- A neural network's task is to analyze this grid of numbers, decipher the patterns within, and ultimately identify the person in the image.

Source: Designed by Freepik

Generative AI

- **Generative AI** is a category of artificial intelligence algorithms that **produce new, realistic outputs** in the form of text, images, audio, video, or synthetic data.
- Typically, it starts with a simple prompt in which the user describes the output they want.
- Then, various algorithms generate new content according to what the prompt was asking for.



Types of content generation

TEXT

ChatGPT
Google Gemini
Microsoft Copilot

IMAGE

Adobe Firefly
DALL-E-3
Mid Journey
Stable Diffusion

AUDIO

Murf AI
MuseNet
Play HT
Soundraw

VIDEO

Lumen5
Synthesia

Key takeaways

- AI technologies are leveraged to drive automation, efficiency, innovation, and strategic advantage.
- Most data and AI projects combine a few methods to extract the full picture.
- The two big components that drive the decision for which method to use are: the **question** you are asking, and the **data** you have.

Agenda

Day 4

- AI methods
- **Refining your data project**
- Intro to data visualization
- Best practices in data viz

Chat discussion: data project methods

- Let's revisit the data projects you've been thinking about one more time

- Are any of the following methods useful to your data project?
  - Clustering, Classification, Regression
  - Text Mining, Graph Analysis
  - Neural Networks, Deep Learning, Transformers

- How might your project be modified or enhanced to take advantage of one of these methods?

Break

Agenda

- Neural networks
- Refining your data project
- Intro to data visualization
- Best practices in data viz

What are neural networks?
What are the limitations to using neural networks?

Agenda

Day 4

- AI methods
- Refining your data project
- **Intro to data visualization**
- Best practices in data viz

What is data visualization?

- Data visualization is any attempt to make data more easily digestible by rendering it in a visual context.

- Common data visualizations include tables, charts, graphs, and dashboards.

Explore or explain

- We can use data visualization to review new data to discover patterns, to spot anomalies, to test hypotheses, and to check assumptions.

- We can also use data visualization to transform raw data into a compelling story or takeaway for an external audience.

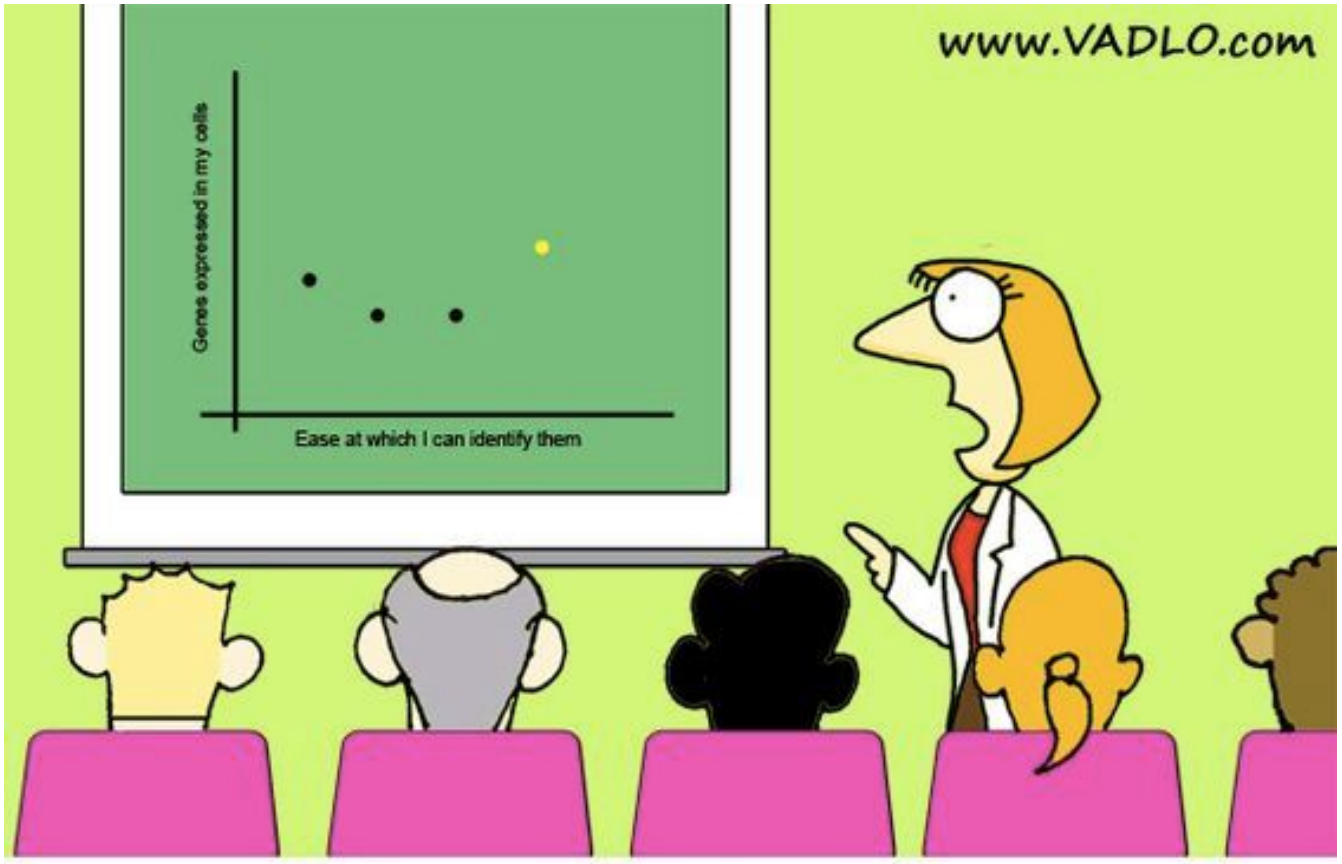
Chat question

Is data visualization an important part of your job?

What types of data visualization does your organization produce?

Getting it right

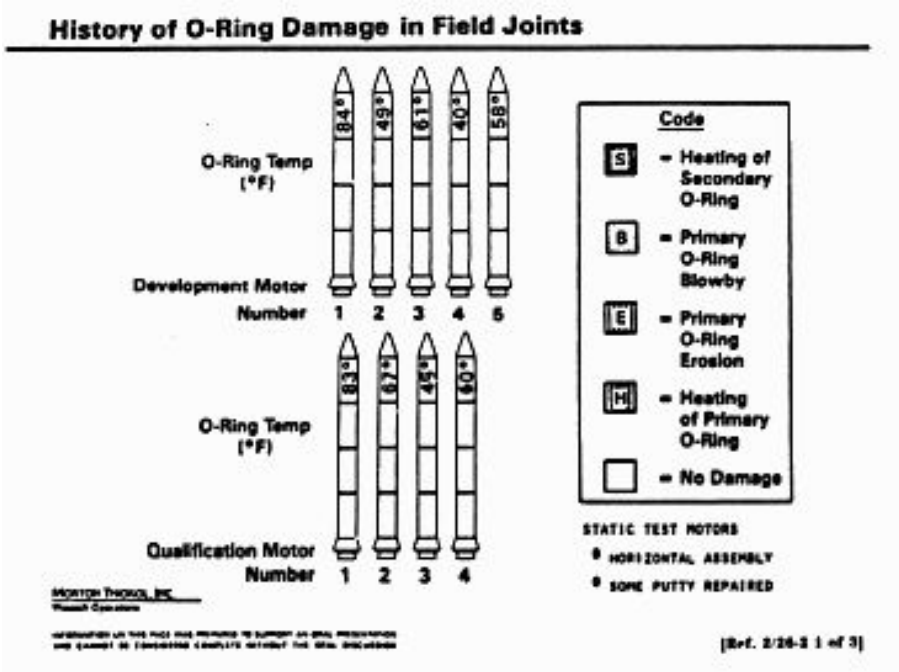
- Using visualizations incorrectly can cause you to lose your audience, lose the value in your data, and ultimately lead to poor decision making.



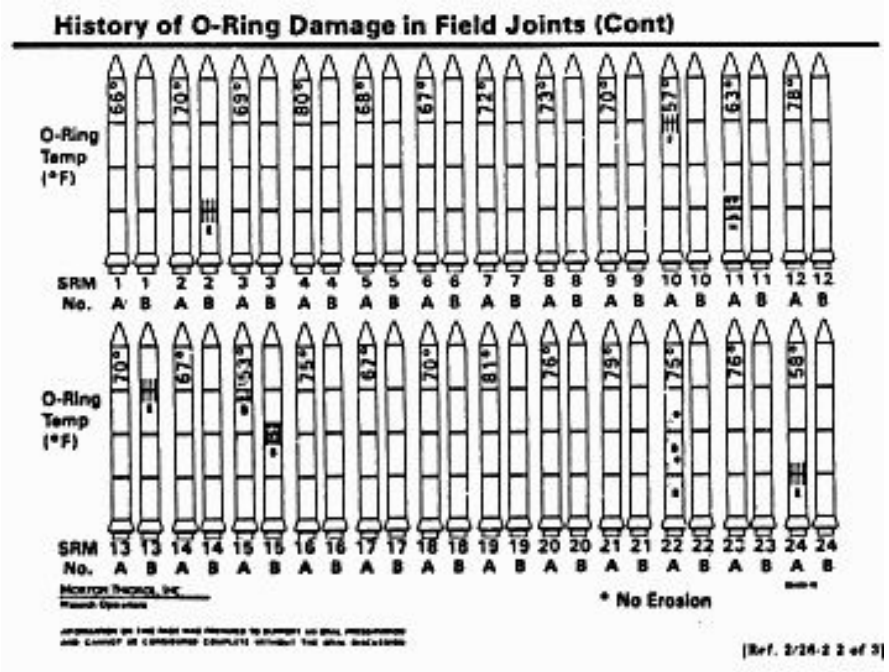
"Same graph as last year,
but now I have an additional dot."

Example: The Challenger

- On January 27, 1986, concerned engineers presented data and the following charts to try to illustrate the damage cold temperatures would have on the O-rings of the Challenger space shuttle.



Source: Presidential Commission on the Space Shuttle Challenger Accident, vol. 5 (Washington, DC: US Government Printing Office, 1986.) pp. 895-896.



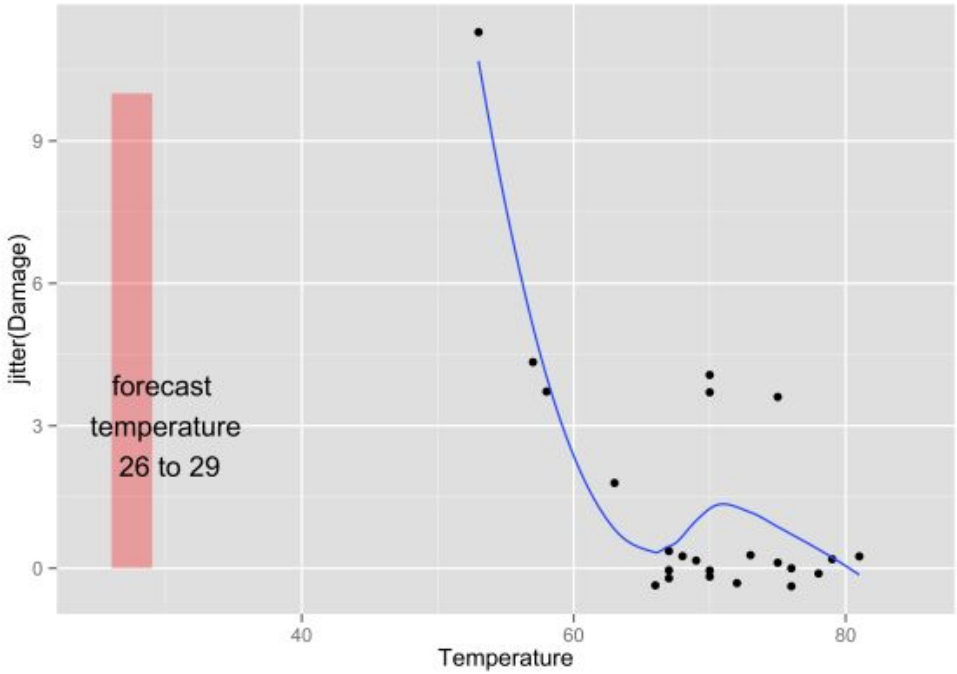
Example: The Challenger

- January 28, 1986, the Challenger space shuttle exploded within seconds of takeoff.

- Data visualization legend Edward Tufte argues that the shuttle's engineers failed to communicate dangers because their data wasn't presented in an easily digestible form.

The chart below shows O-ring damage on the y-axis and temperature on the x-axis.

Is it easier to see the issue?



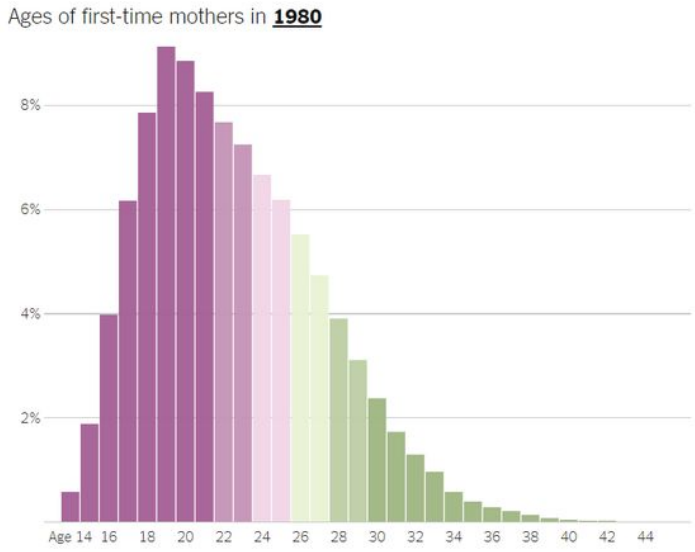
https://vizdatar.wordpress.com/2015/05/06/space-shuttle-challenger-explosion-2/

Representing distribution

- Boxplots and histograms are useful for showing the values of a single variable and the frequency of their occurrence.

https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/boxplot

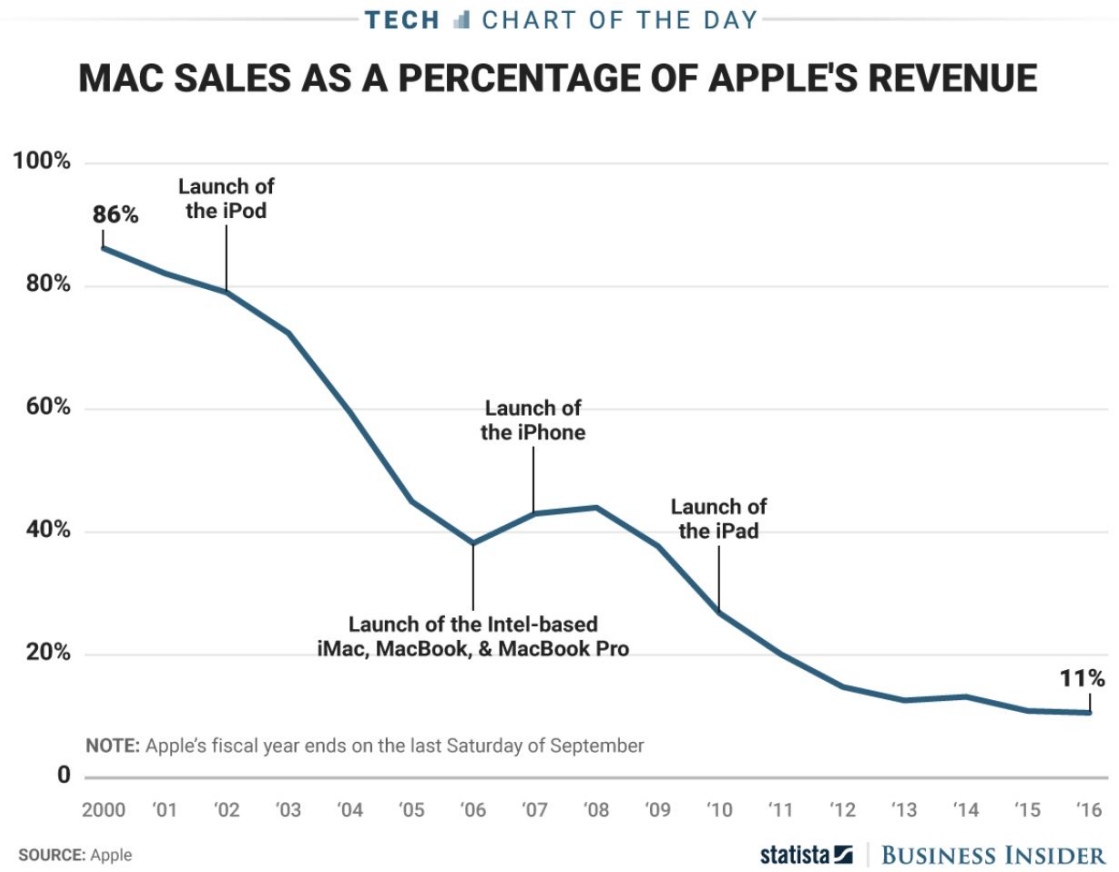
- A boxplot displays median, higher/lower quartiles and maximum/minimum.
- A histogram groups numeric data into bins, displaying the bins as segmented columns.



https://www.nytimes.com/2018/11/22/learning/whats-going-on-in-this-graph-nov-28-2018.html

Representing time series

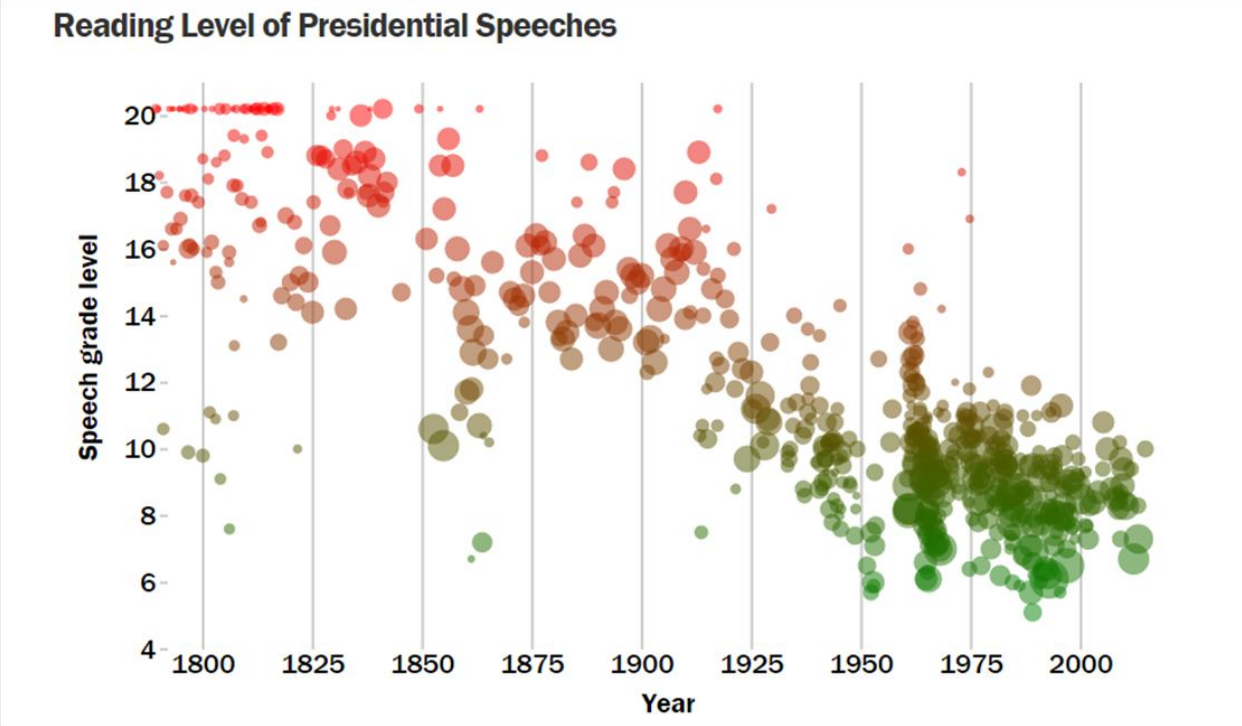
- A line chart displays information as a series of data points called markers.
- The markers are connected by straight line segments to show trend.
- An area chart is a line chart with the area below the lined filled with colors or textures.



http://www.datavizdoneright.com/2017/04/mac.html

Representing association

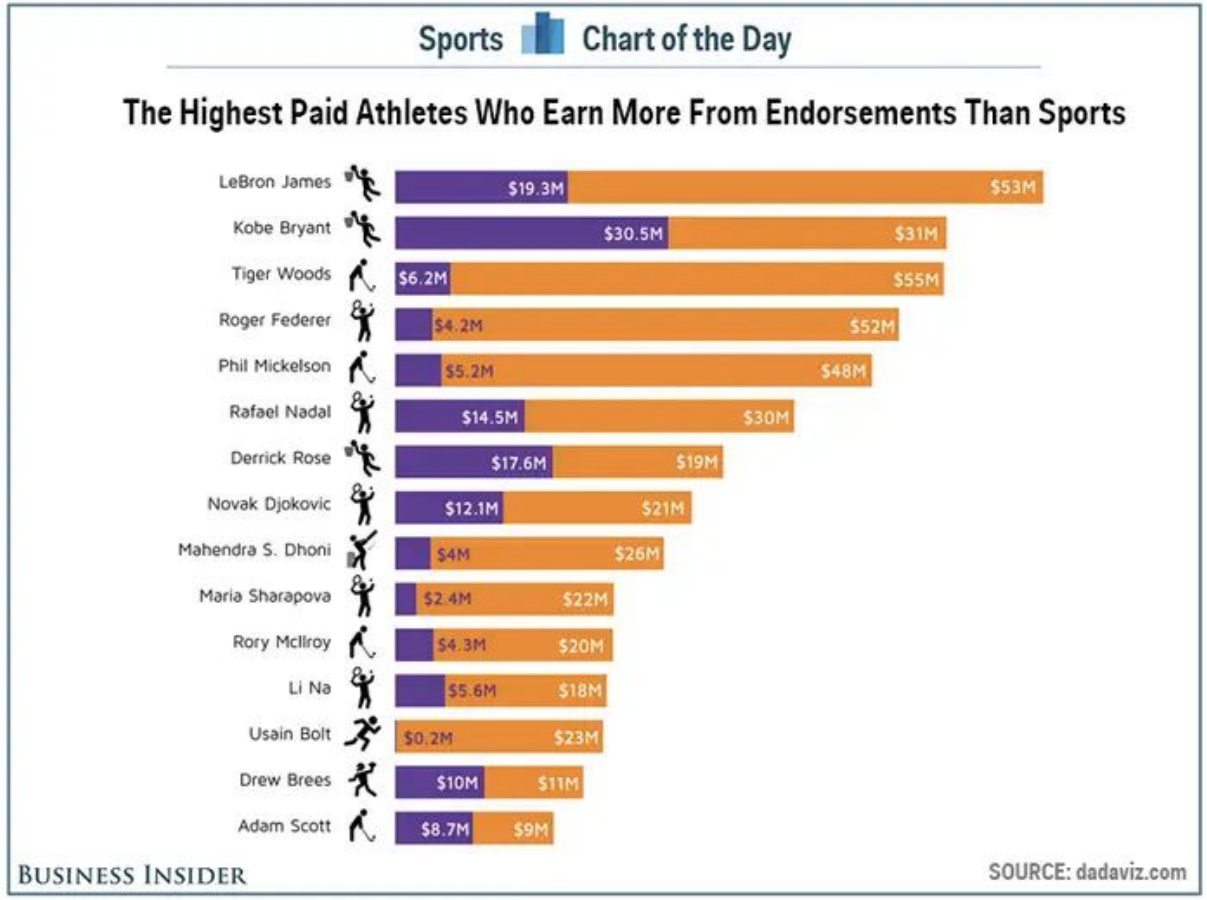
- A scatterplot is a type of diagram that uses coordinates to display values for two variables for a set of data.
- Additional data can be encoded by varying the color or shape of each marker.
- Bubble charts encode another variable through the size of the markers.



http://dadaviz.com/i/1276

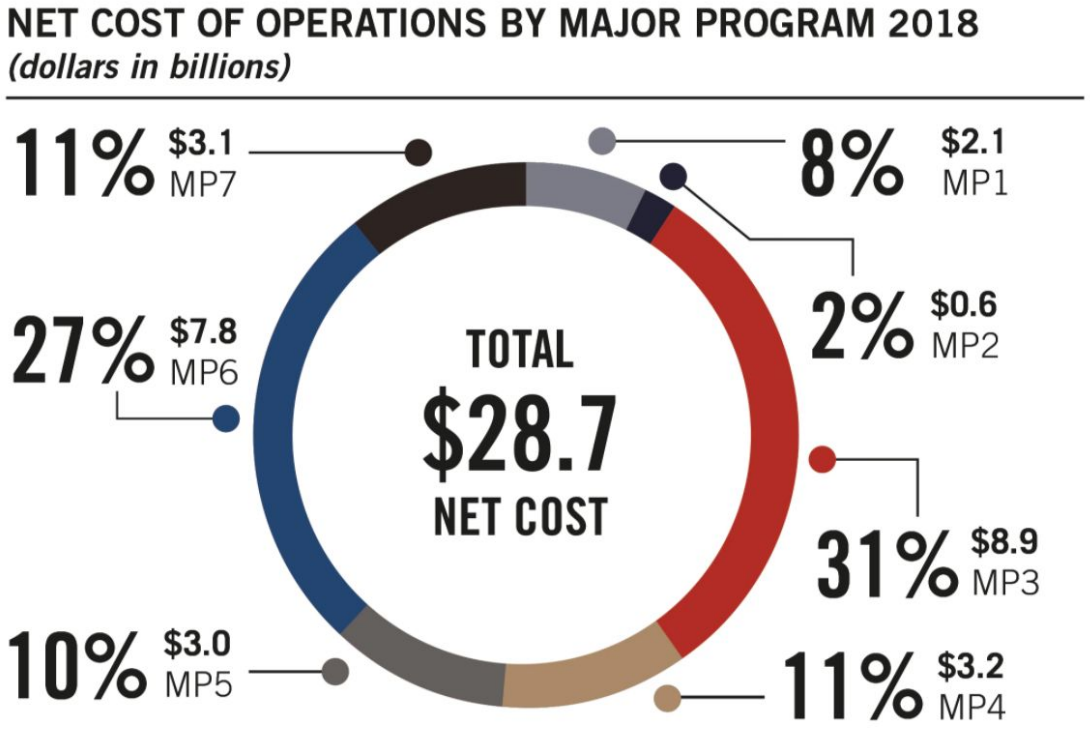
Comparing categories

- A bar chart is a chart with rectangular bars with lengths proportional to their values.
- One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value.
- Vertical bar charts are also called column charts.



Comparing part to whole

- A pie chart is divided into sectors, illustrating numerical proportion.



Cases per 100k people

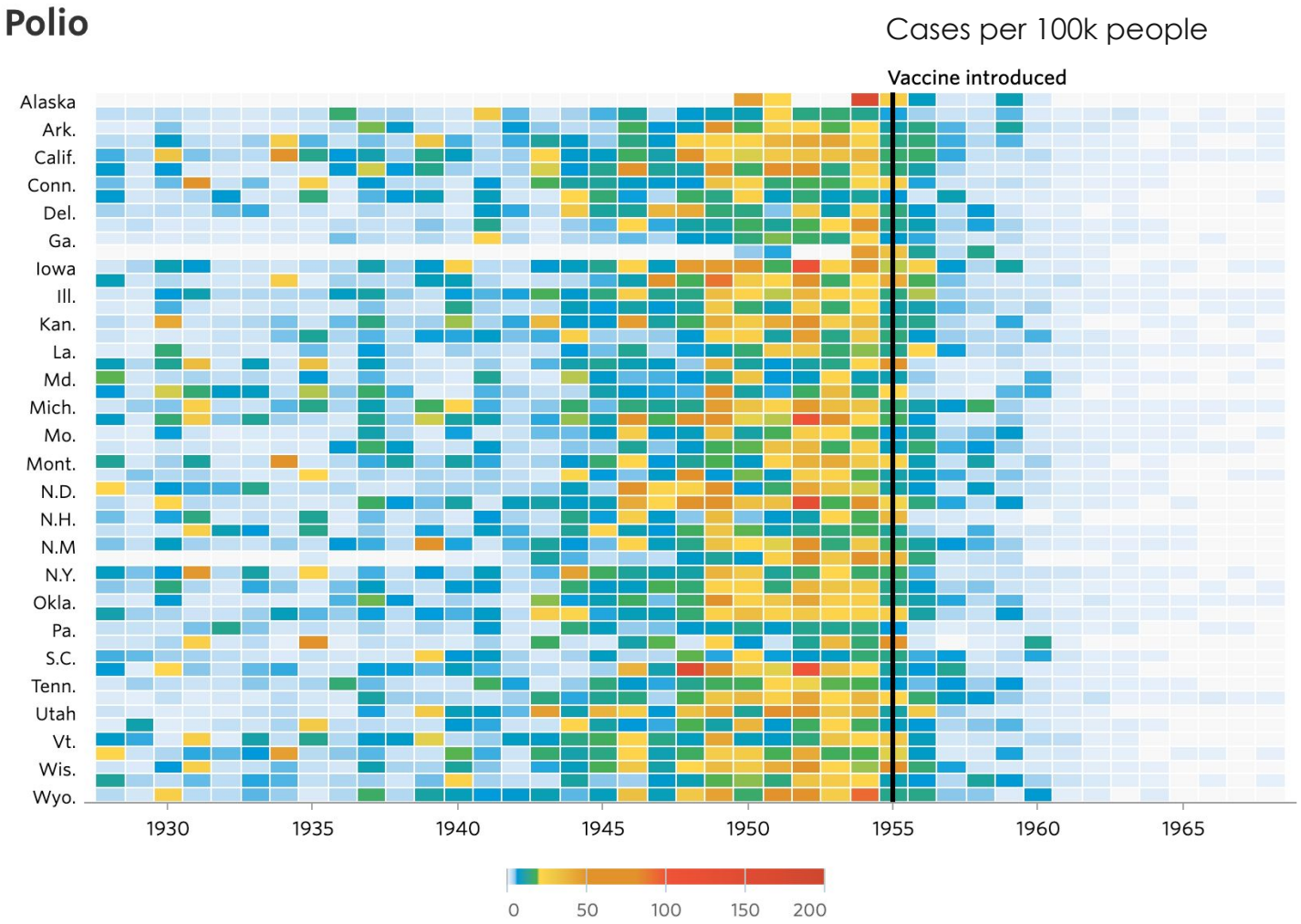
- A doughnut chart is a pie chart with a blank center to display data.

https://www.state.gov/reports/fy-2018-department-of-state-agency-financial-report/section-i-managements-discussion-and-analysis/

Representing density

- In a heatmap, individual values are contained in a matrix with variations in coloring to show magnitude or concentration.

- Heatmaps allow you to see where the greatest number of a variable of interest is distributed.

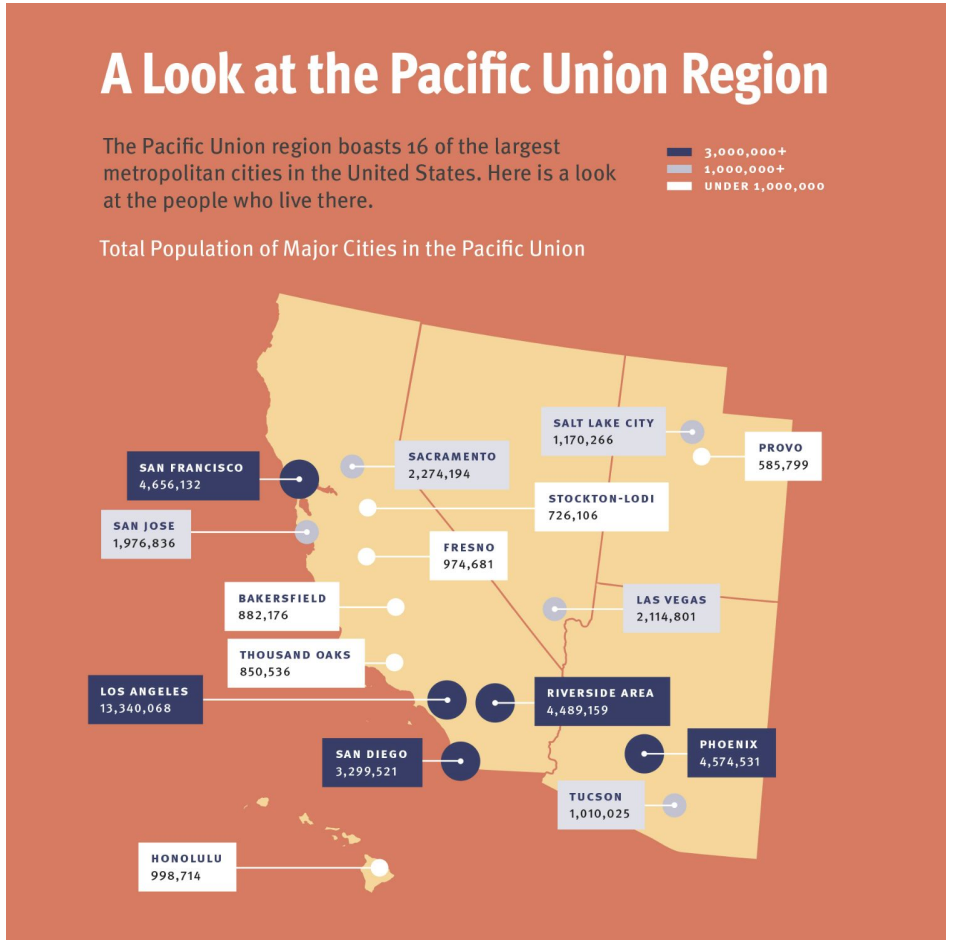


http://graphics.wsj.com/infectious-diseases-and-vaccines/

Representing territory

- Maps present geographically-related data in a clear and intuitive manner.

- Maps are often combined with points, lines, bubbles, and more.



https://dribbble.com/shots/2881738-Infographic/attachments/595499

Just a few numbers

- Don't overcomplicate!
- Simple text works well when there is just a number or two to share.

…we spent only $75,000 of our $125,000 budget…

…therefore, it is not surprising that only 29 percent of the applications were accepted…

…product A ($12.99) was much more affordable than product B ($59.99)…

Unique data

- Don't overcomplicate!

- Tables are great when communicating to a mixed audience who will look to a particular row of interest or when you need to show different units of measure.

Valid Passports in Circulation (1989-2020)

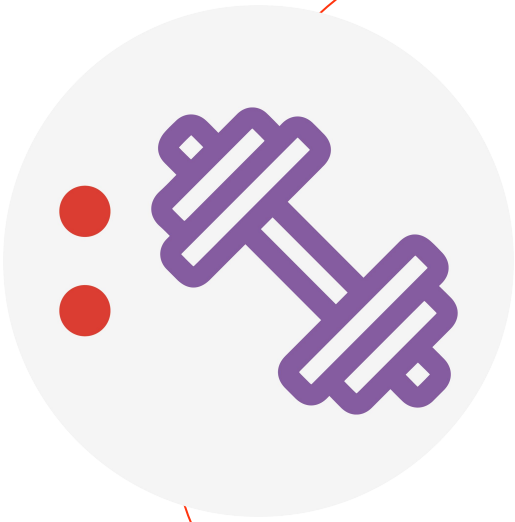
Enter Text To Filter Table Below

Year	Valid U.S. Passports
2020	143,116,633
2019	146,775,089
2018	137,588,631
2017	136,114,038
2016	131,841,062
2015	125,907,176
2014	121,512,341
2013	117,443,735
2012	113,431,943
2011	109,780,364
2010	101,797,872
2009	97,597,368
2008	92,038,623
2007	82,100,668
2006	70,598,794
2005	64,772,634
2004	60,890,770
2003	57,642,868
2002	55,169,571

https://travel.state.gov/content/travel/en/about-us/reports-and-statistics.html

Activity: choosing data visualizations

- Turn to page 17 of your participant guide to find the **what would you viz?** activity.
- Read the description of each dataset. Then choose which of the available visualizations you would use to best represent the data.
- Check how you did using the answer key on page 19



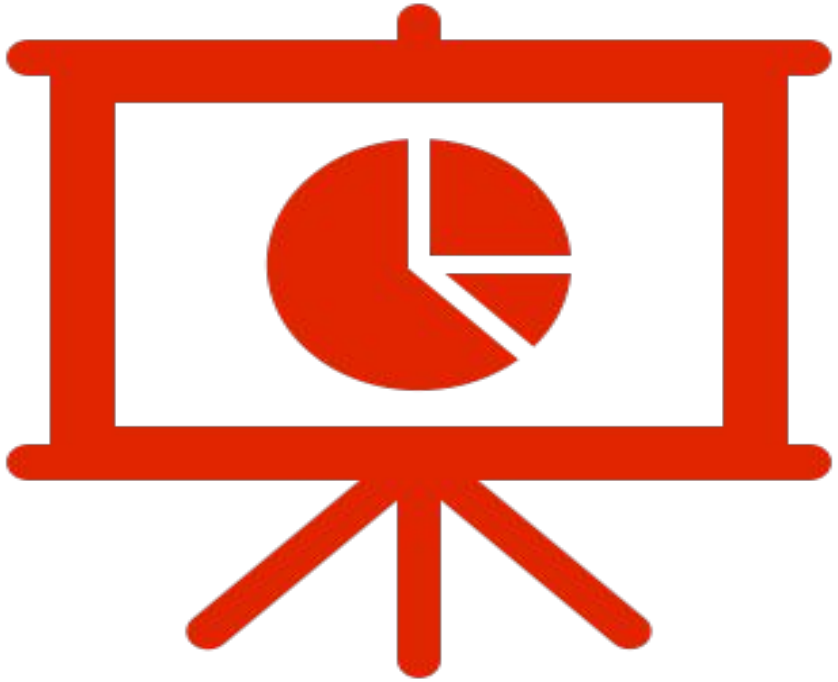
Agenda

Day 4

- AI methods
- Refining your data project
- Intro to data visualization
- **Best practices in data viz**

Designing compelling visuals

- Picking the right chart type isn't enough.
- There are choices to be made about the elements you include and how they are formatted.
- Data visualization is an art, informed by science.



Visual design theory

- Our eyes "load" information while the brain "processes" it.
- We give the most attention to what looks good and struggle when our working memory is overwhelmed.
- For information to be effective, it should not provide more data than what the human brain can process.

Example: buying oranges

- You want to buy oranges at a new supermarket.
- Our eyes scan the layout of the supermarket, while the brain processes the various sections.
- The brain then instructs the eyes to zone in on the fruit section by sending signals about how fruits look from memory.
- The eyes then break the entire scanned area into parts and scan each part to spot the fruit section.
- The process is repeated until oranges are located.

Designing compelling visuals

- Our eyes and brains work the same way with data visualizations as they did in the oranges example.
  - Use visual clues to make data visualizations easier for the audience.
- However, every piece of information in a visualization also creates cognitive load on the viewer, asking them to use their brain power to process it.
  - Reduce visual clutter to lower the cognitive load and help transmission of the message.

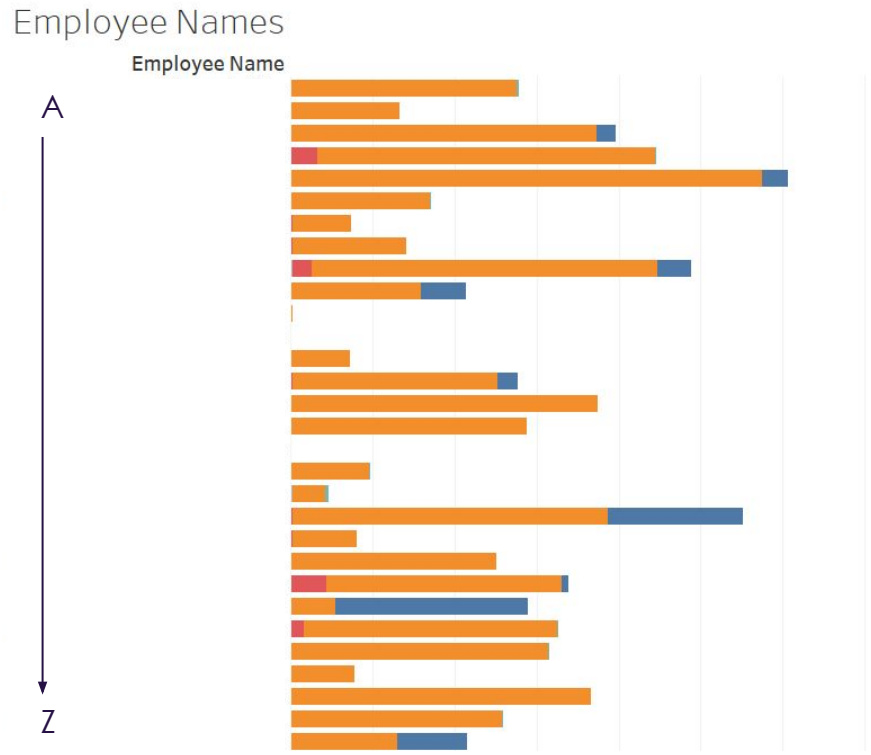
Theory

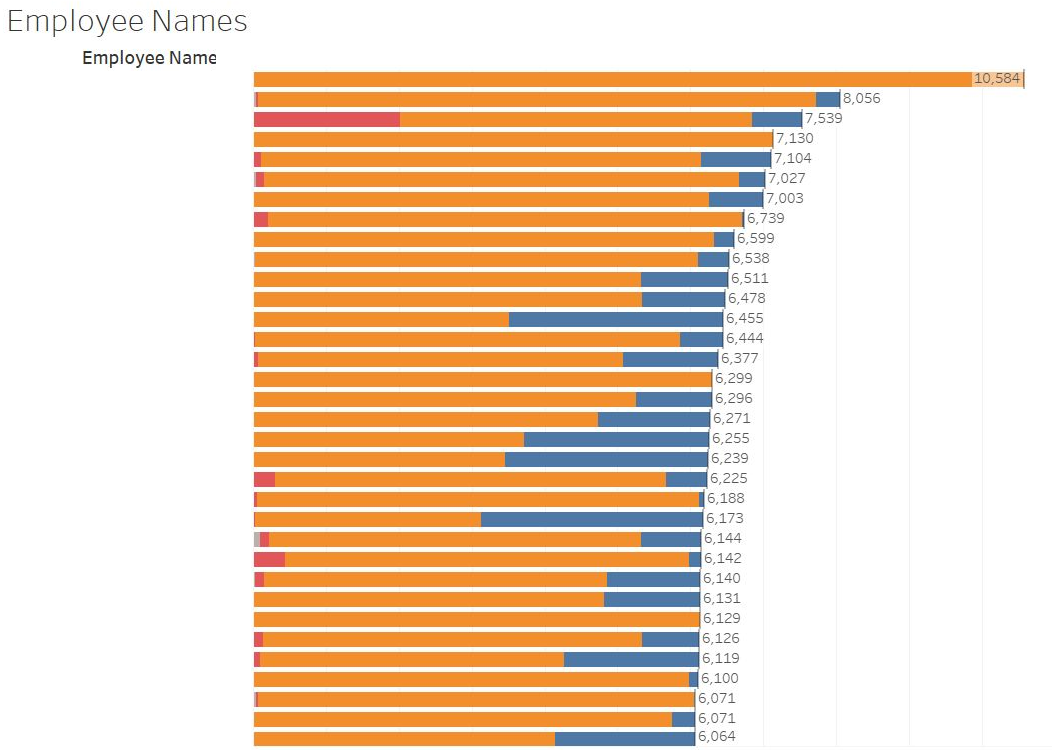
- The visual design tips we'll review today draw on theory such as:
  - the building blocks of visual design described by the Interaction Design Foundation
  - the four categories of preattentive visual attributes described in Colin Ware's book, Information Visualization: Perception for Design
  - the Gestalt Principles of visual perception, which describe how people group similar elements, recognize patterns, and simplify complex images when we perceive objects

Tip 1

Make position meaningful

Data should be sorted and placed in the visual in a meaningful way.





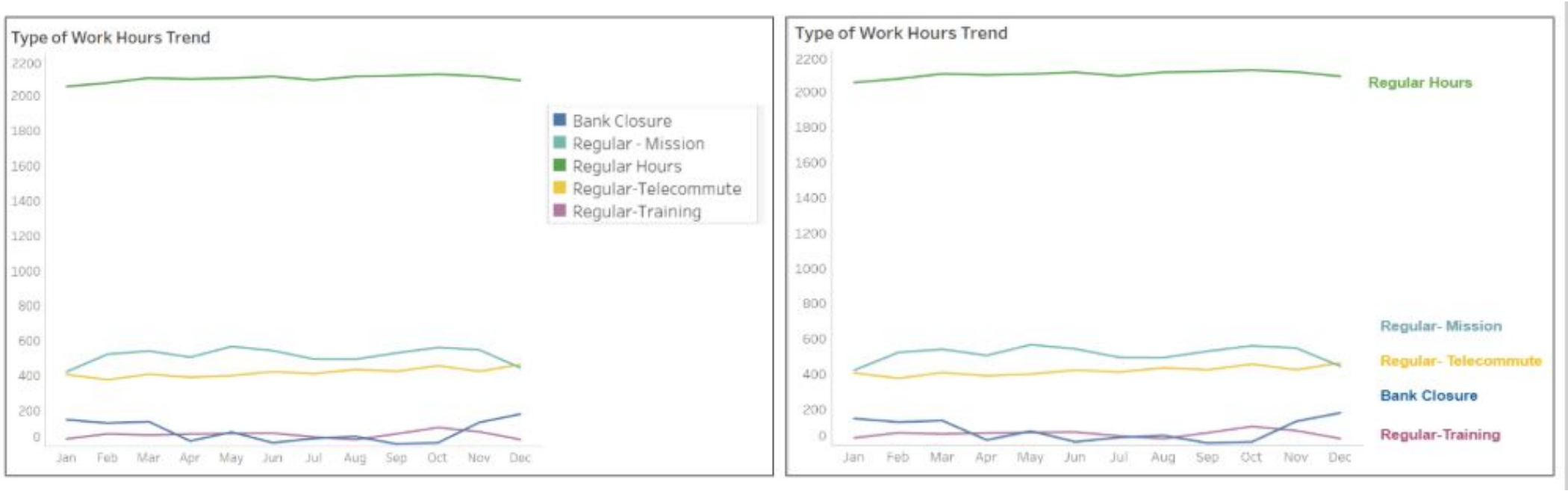
The left chart is sorted alphabetically; the right by value.

When would you use one over the other?

Tip 2

Group related items

- Things that are closer appear to be more related than those that are spaced farther apart.
- In fact, proximity overrules the similarity of other factors (e.g., shape, color).

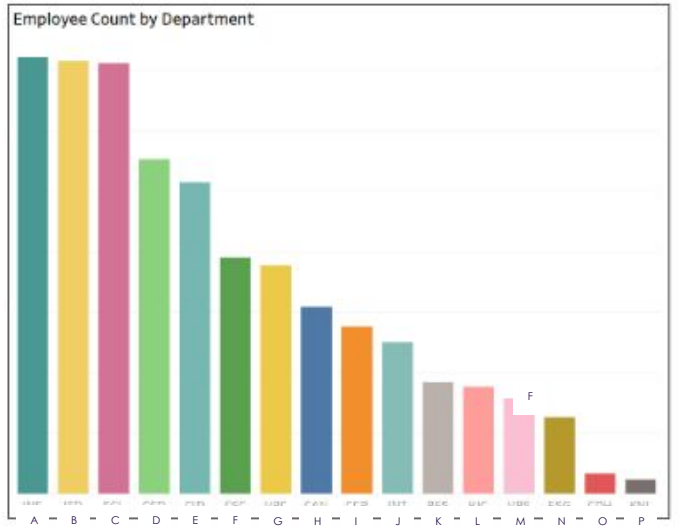


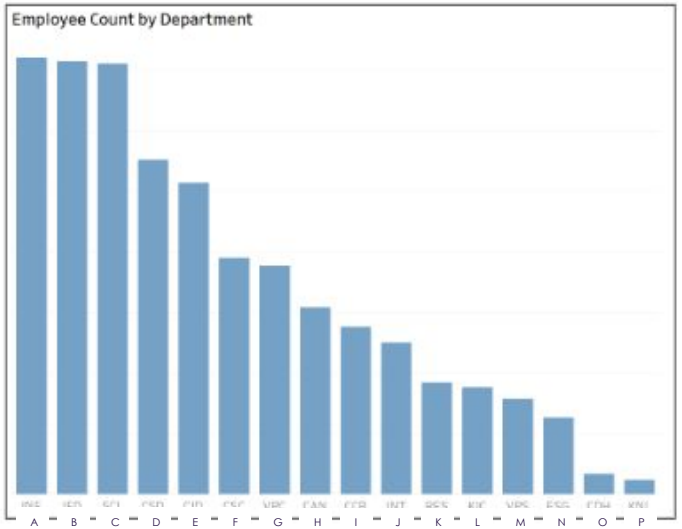
Tip 3

Distinguish different items

- The mind groups together things that look to be similar and assumes they have the same function.

- We can use this principle for:
  - distinguishing different sections
  - differentiating links from regular text
  - showing that elements with certain characteristics serve one purpose and others different

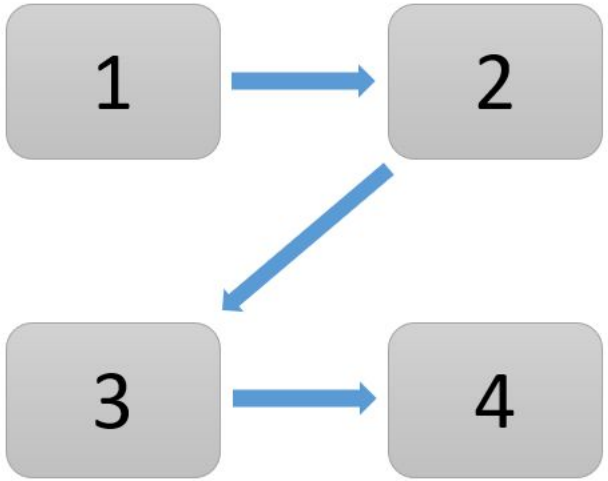




Tip 4

Use natural positioning

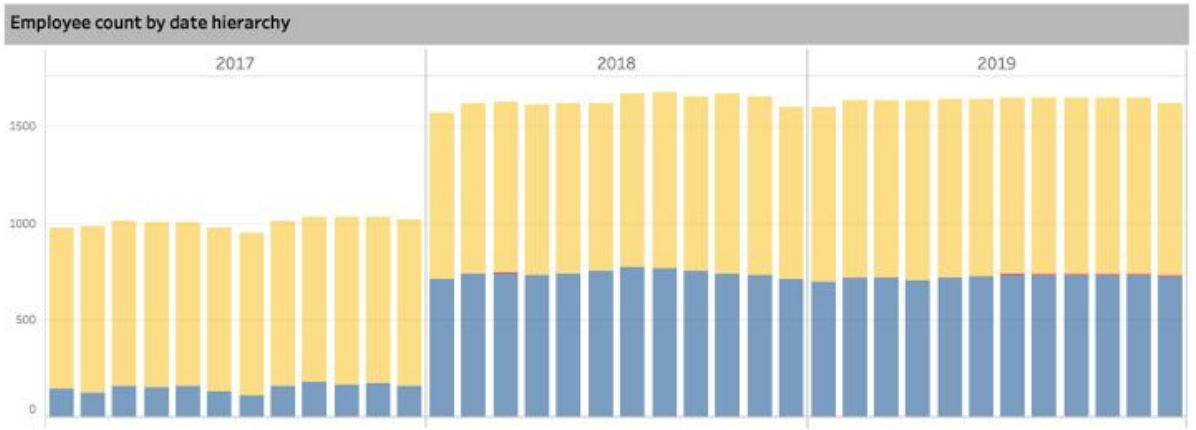
- People usually tend to start at the top left of the visual and scan in zig-zag motions across the page forming a Z-pattern.
- Aim to position elements in a way that will feel natural for users to consume.
- Also, remember that the top of the page is the most precious.

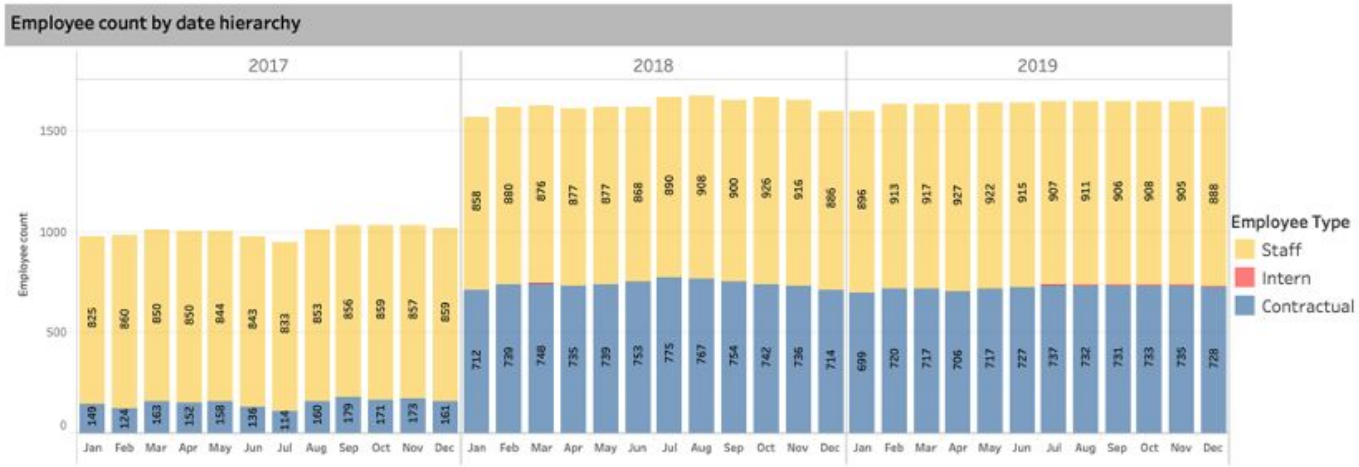


Tip 5

Use labels and legends

- Labels can be used to show value of datapoint.
- Legends can be used to identify the size, color or any other distinguishing feature in the visual.



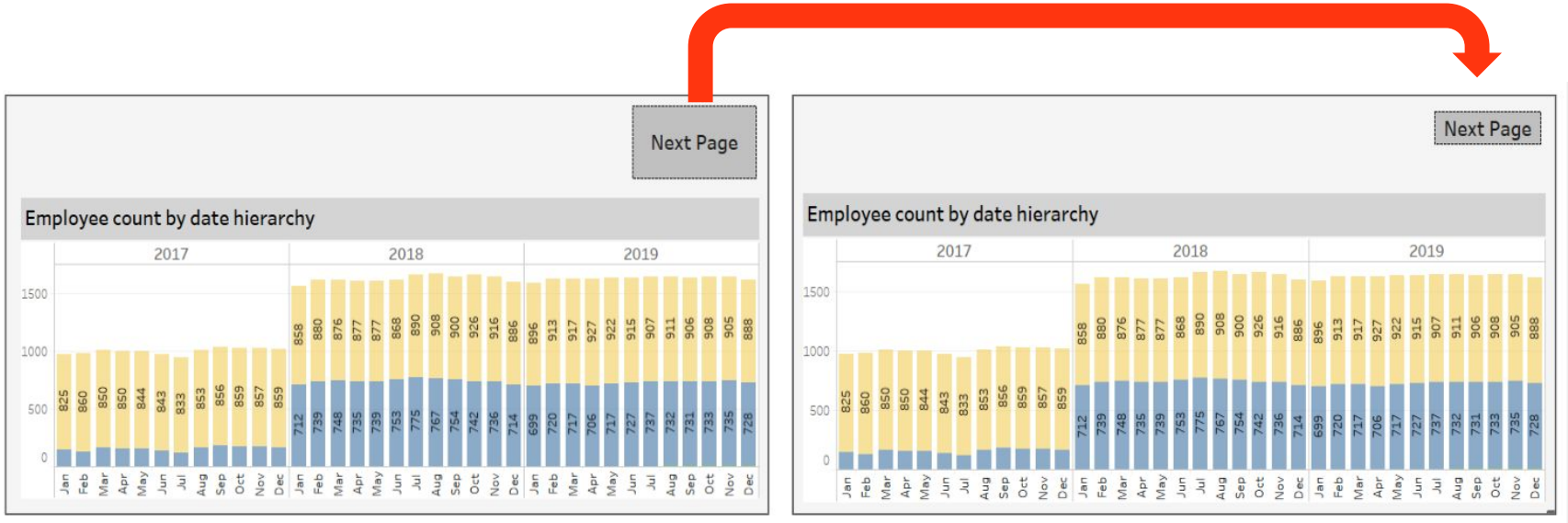


The labels and legends used in the bottom chart makes it easier to understand.

Tip 6

Use size to show importance

- Relative size represents relative importance.
- Visuals of almost equal importance should be sized similarly.
- If there's one really important thing, it must be BIG.



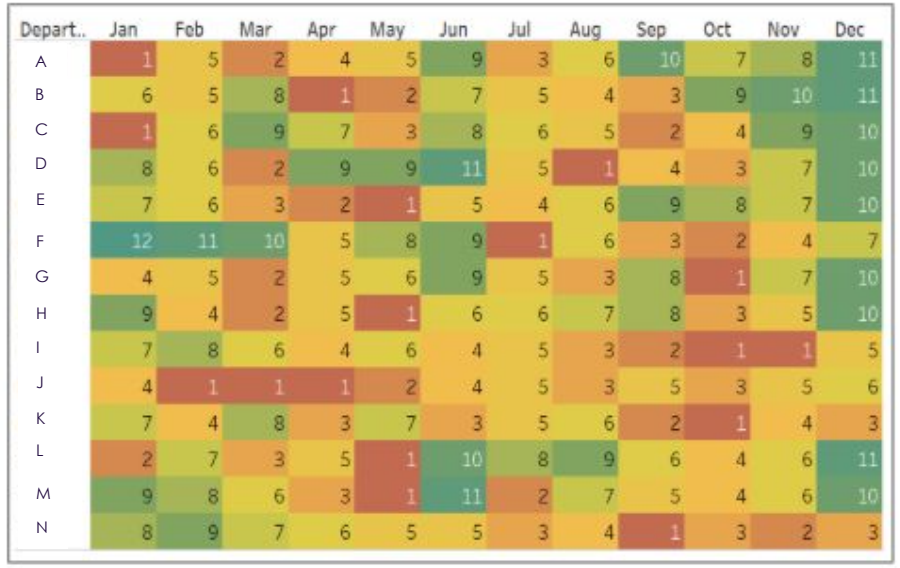
Resizing the "Next Page" button deemphasizes its importance.

Tip 7

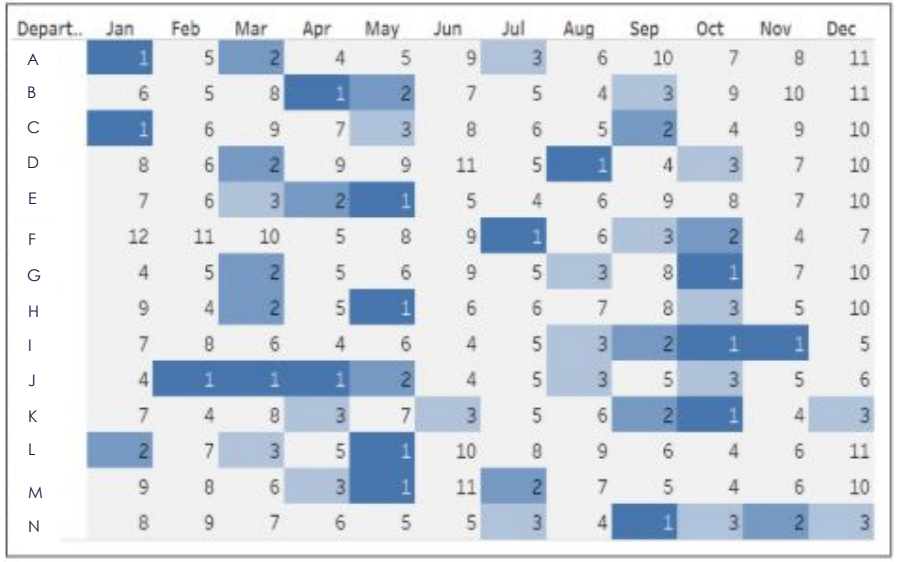
Use color to grab attention

- Color is another powerful tool used to draw the audience's attention
- However, the following must be kept in mind:
  - Use it sparingly: too much variety prevents anything from standing out
  - Use it consistently: a color change can be used to visually reinforce change in topic or tone

Too many colors are used in the image on the left, making it difficult to identify which are the busiest months.



Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	1	5	2	4	5	9	3	6	10	7	8	11
B	6	5	8	1	2	7	5	4	3	9	10	11
C	1	6	9	7	3	8	6	5	2	4	9	10
D	8	6	2	9	9	11	5	1	4	3	7	10
E	7	6	3	2	1	5	4	6	9	8	7	10
F	12	11	10	5	8	9	1	6	3	2	4	7
G	4	5	2	5	6	9	5	3	8	1	7	10
H	9	4	2	5	1	6	6	7	8	3	5	10
I	7	8	6	4	6	4	5	3	2	1	1	5
J	4	1	1	1	2	4	5	3	5	3	5	6
K	7	4	8	3	7	3	5	6	2	1	4	3
L	2	7	3	5	1	10	8	9	6	4	6	11
M	9	8	6	3	1	11	2	7	5	4	6	10
N	8	9	7	6	5	5	3	4	1	3	2	3



Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	1	5	2	4	5	9	3	6	10	7	8	11
B	6	5	8	1	2	7	5	4	3	9	10	11
C	1	6	9	7	3	8	6	5	2	4	9	10
D	8	6	2	9	9	11	5	1	4	3	7	10
E	7	6	3	2	1	5	4	6	9	8	7	10
F	12	11	10	5	8	9	1	6	3	2	4	7
G	4	5	2	5	6	9	5	3	8	1	7	10
H	9	4	2	5	1	6	6	7	8	3	5	10
I	7	8	6	4	6	4	5	3	2	1	1	5
J	4	1	1	1	2	4	5	3	5	3	5	6
K	7	4	8	3	7	3	5	6	2	1	4	3
L	2	7	3	5	1	10	8	9	6	4	6	11
M	9	8	6	3	1	11	2	7	5	4	6	10
N	8	9	7	6	5	5	3	4	1	3	2	3

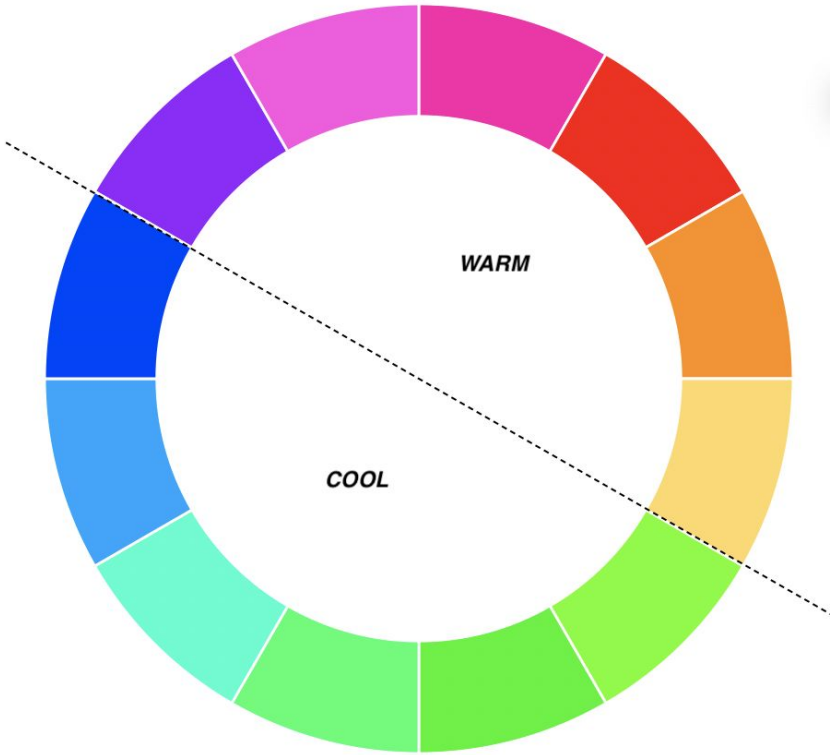
Tip 8

Use color to evoke emotion

- Color evokes emotion, so choose the one that helps reinforce the emotion you want to arouse in your audience.

Warm colors represent energy

Cool colors represent calmness



Tip 9

Encode data with color

- Use color schemes to encode data as sequential, diverging, or categorical.

Sequential	Diverging	Categorical
when the order matters	to highlight minimums, maximums, and midpoints	for discrete data values representing distinct categories

Sequential color schemes

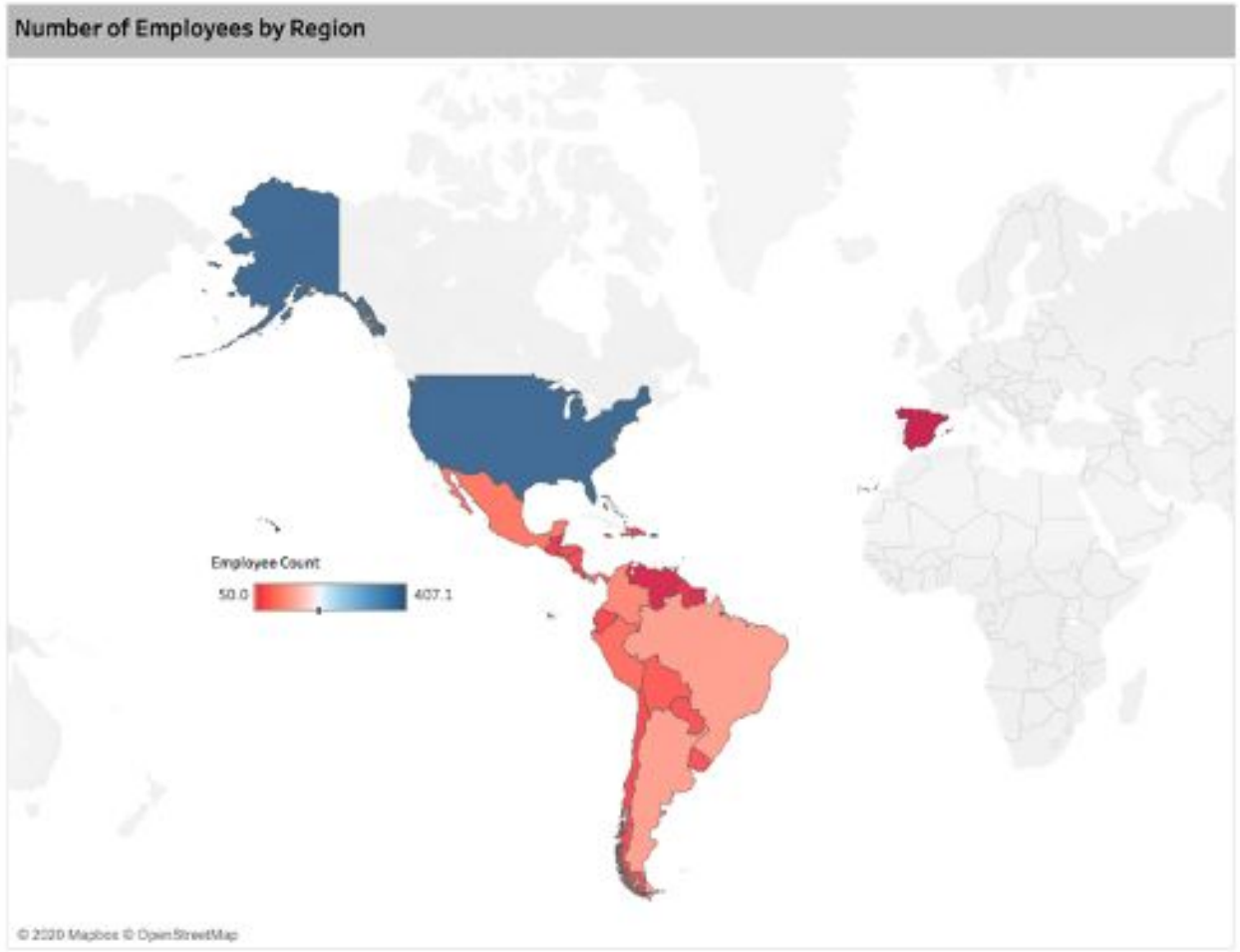
- Use a sequential color scheme when the order matters.

- These schemes range between two colors—usually a lighter shade to a darker one—by varying one or more parameters such as saturation.

Diverging color schemes

- Use a diverging color scheme to highlight minimums, maximums, and midpoints.

- These schemes range between three or more colors with the different colors being quite distinct—usually having different hues.



Categorical color schemes

- Use a categorical color scheme for discrete data values representing distinct categories.

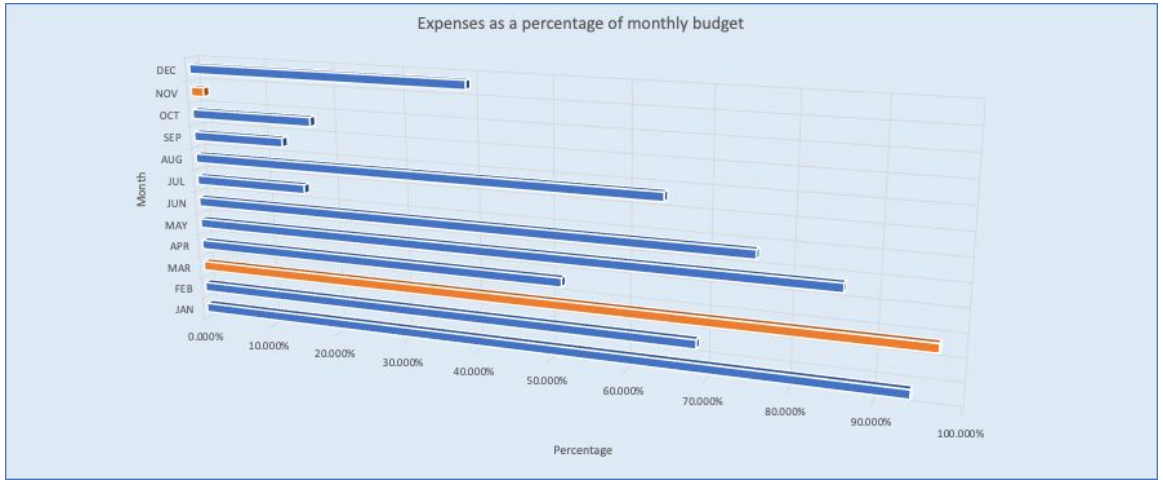
- These schemes use different hues with consistent steps in lightness and saturation.

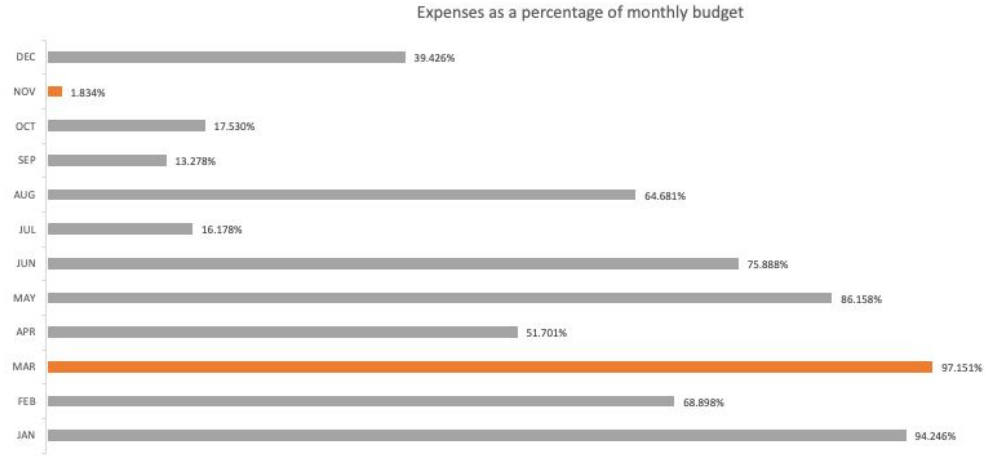
Tip 10

Reduce chart clutter

Small changes can have a big effect on a visualization's impact.

1. Remove special effects
2. Lighten the background
3. Remove chart borders
4. Remove gridlines
5. Direct label
6. Clean up axis titles and labels
7. Use consistent colors

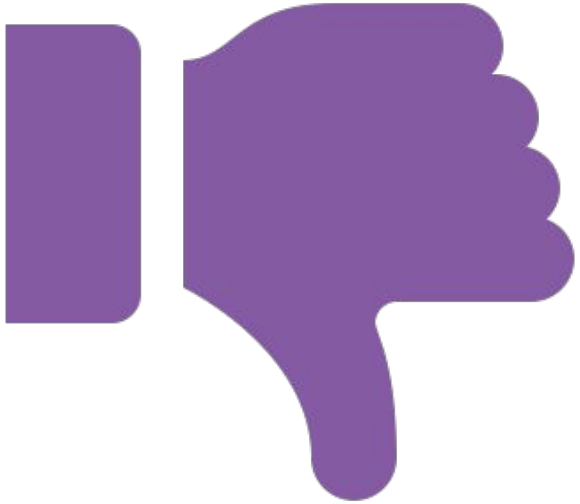




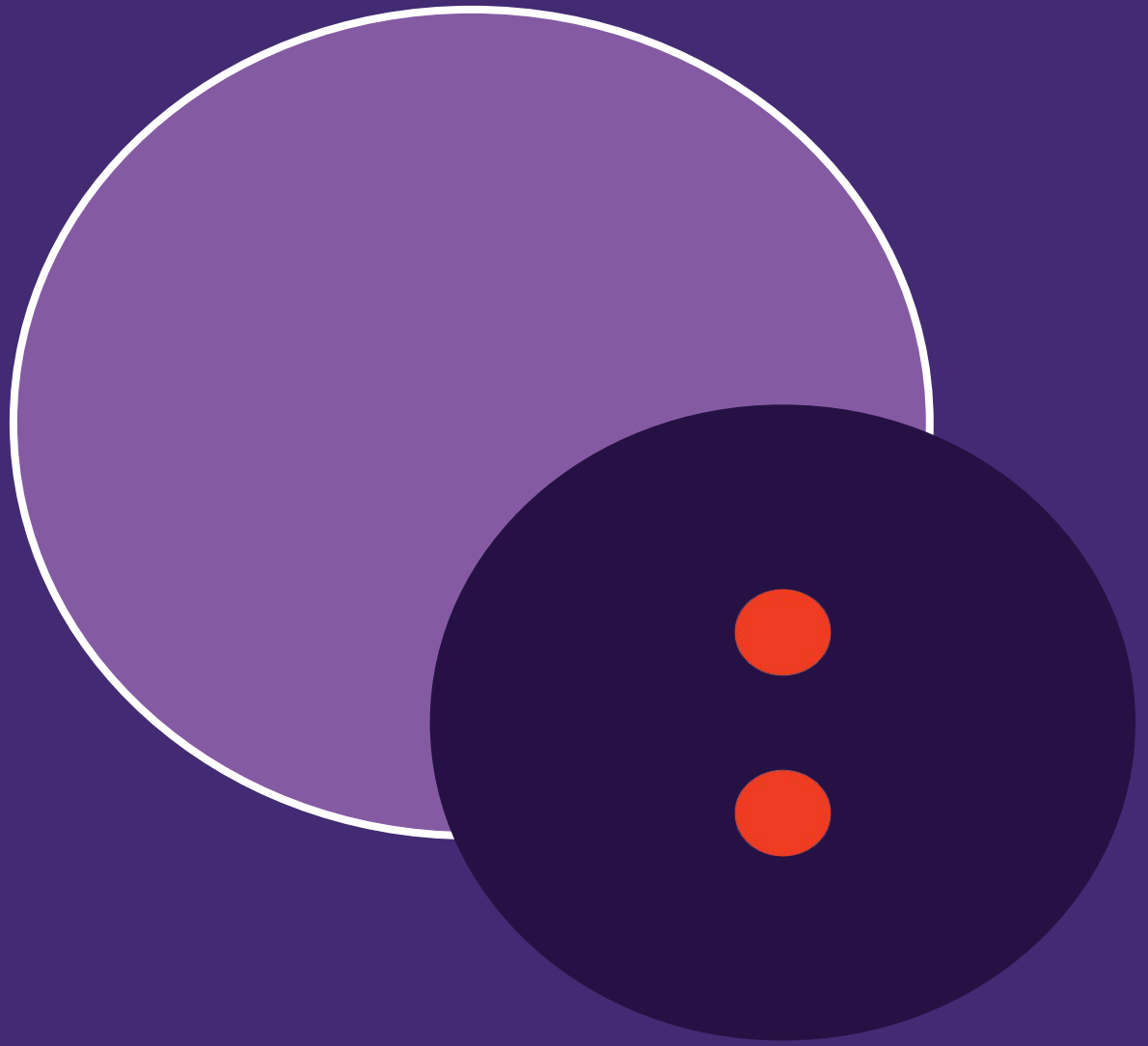
Misleading stats & visual distortions

- Sometimes charts and statistics look presentable but could be misleading.

- Unreliable data comparisons erode credibility and eventually dissuade viewers from using the analysis.



Misleading statistics



Misleading statistics

- "Bill Gates walks into a bar and everyone inside becomes a millionaire…on average."

- In 2011, the average income of the 7,878 households in Steubenville, Ohio, was $46,341. But if just two people, Warren Buffett and Oprah Winfrey, relocated to that city, the average household income in Steubenville would rise 62 percent overnight, to $75,263 per household.

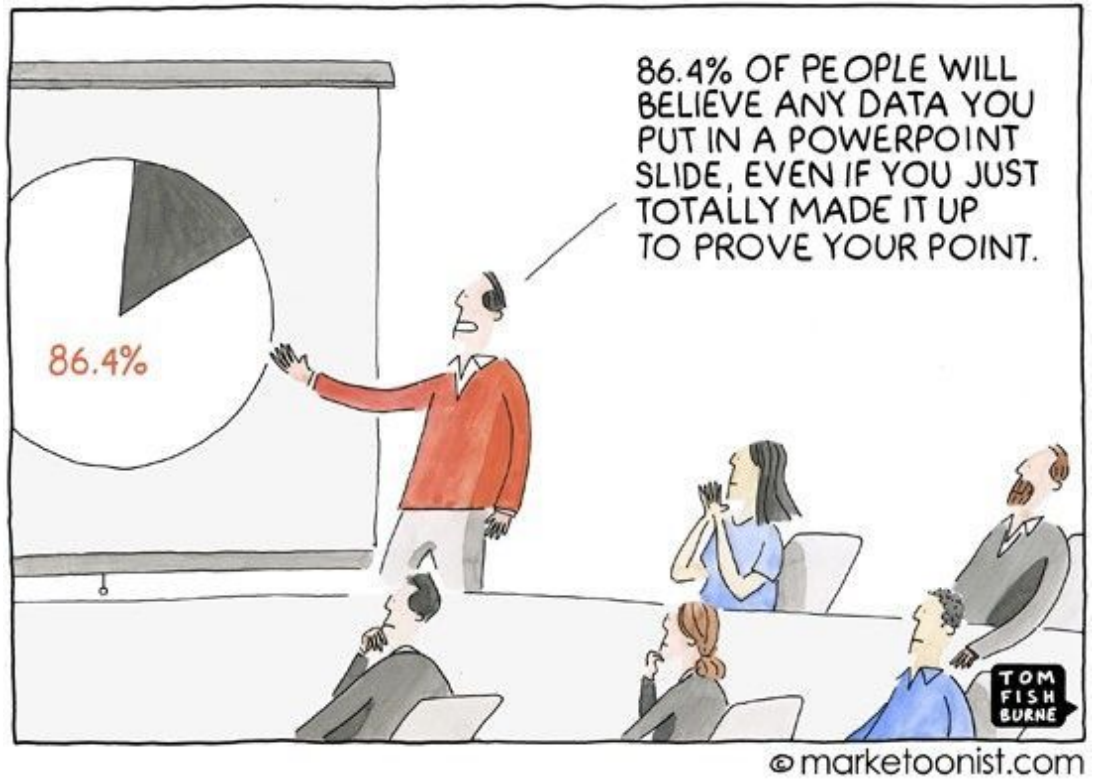
What's wrong with these statements?

https://www.nytimes.com/2013/05/26/opinion/sunday/when-numbers-mislead.html

Misleading statistics

- Numbers don't have to be fabricated to be misleading.

- Misleading statistics are the misusage—purposeful or not—of numerical data.



Misleading statistics

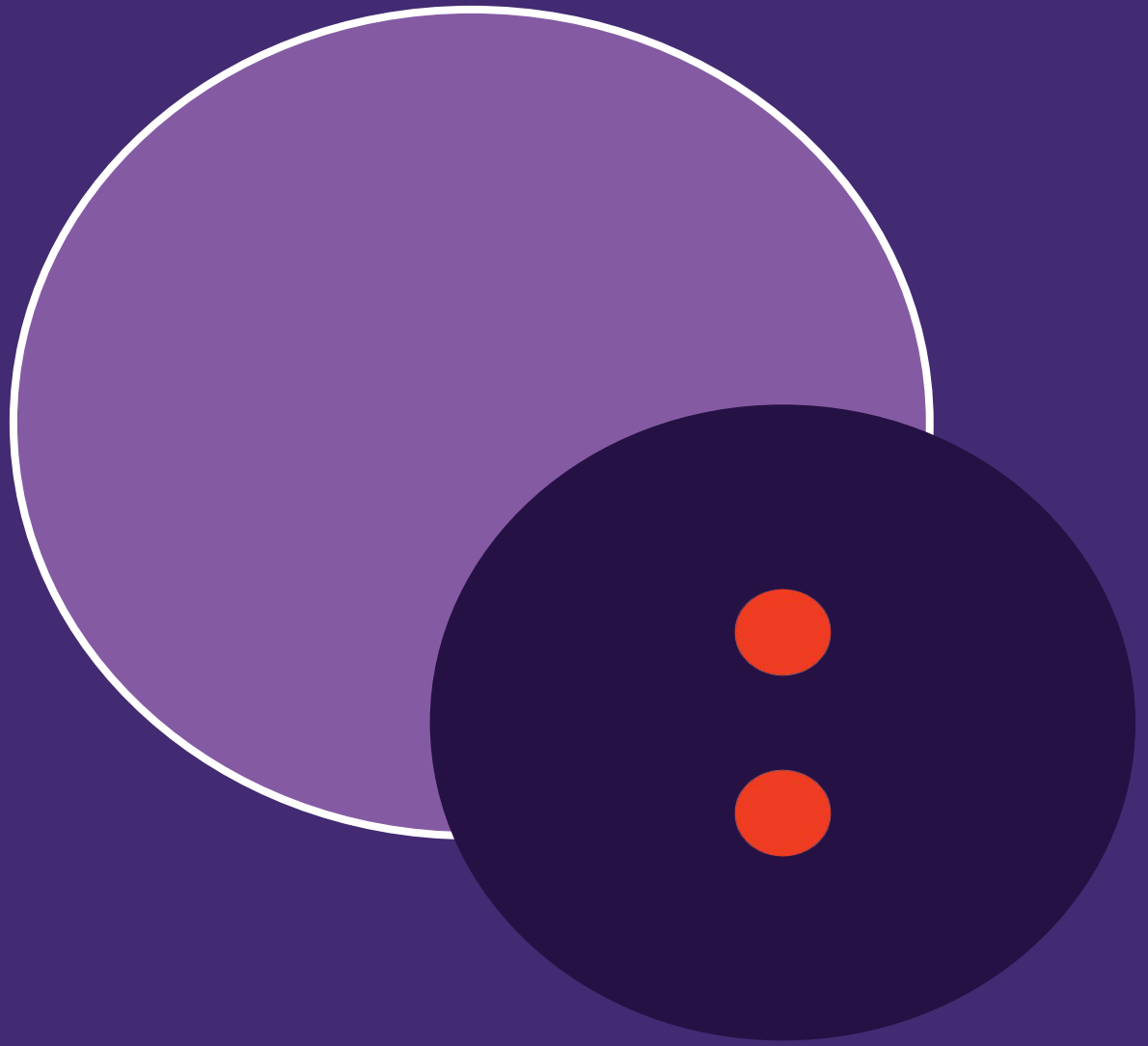
- Misleading statistics can be created through issues with:
  - data collection
  - data processing
  - data presentation

Data collection	• Small sample sizes • Biased sampling • Loaded questions
Data processing	• No/poor data normalization • Ignoring important features
Data presentation	• Hiding context • Omitting certain findings • Visual distortions

How to avoid being misled?

- Do some math. Are there any obvious mistakes?
- Check the source. Is it creditable and current?
- Question the methodology. Is there bias? Is the result statistically significant?
- Conduct research. What does Google tell you?

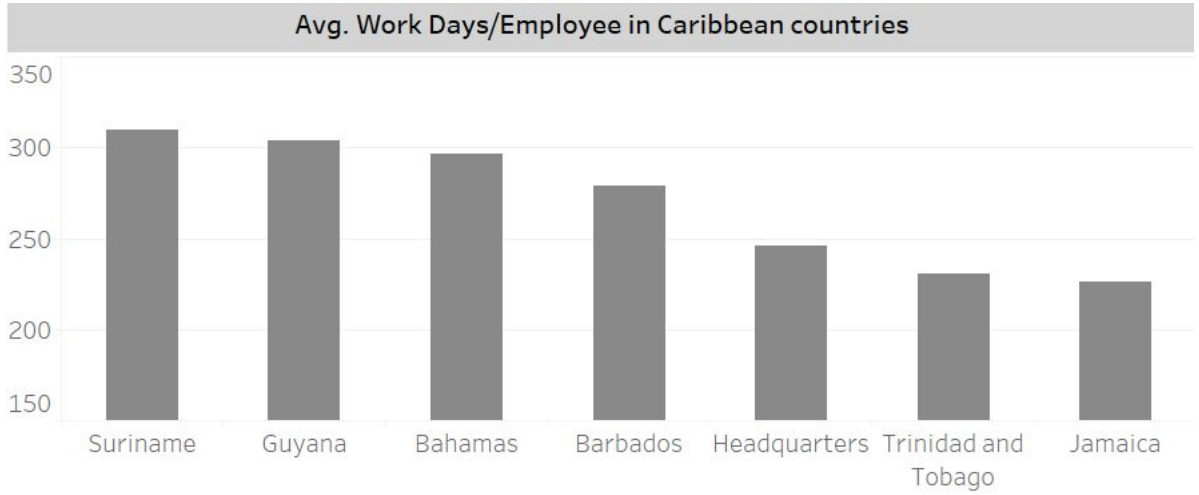
Visual distortions

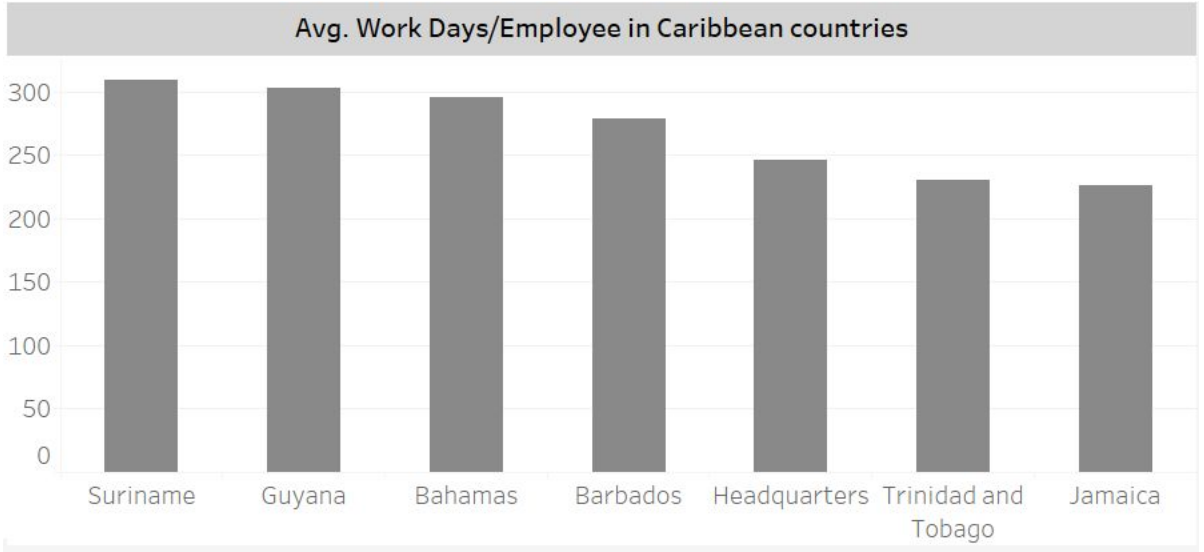


Visual distortions

- Look at the top graph.
- At first, Jamaica seems to have half the average workdays per employee that Suriname does.
- In reality, the difference is much less.

What's the difference between the two charts?

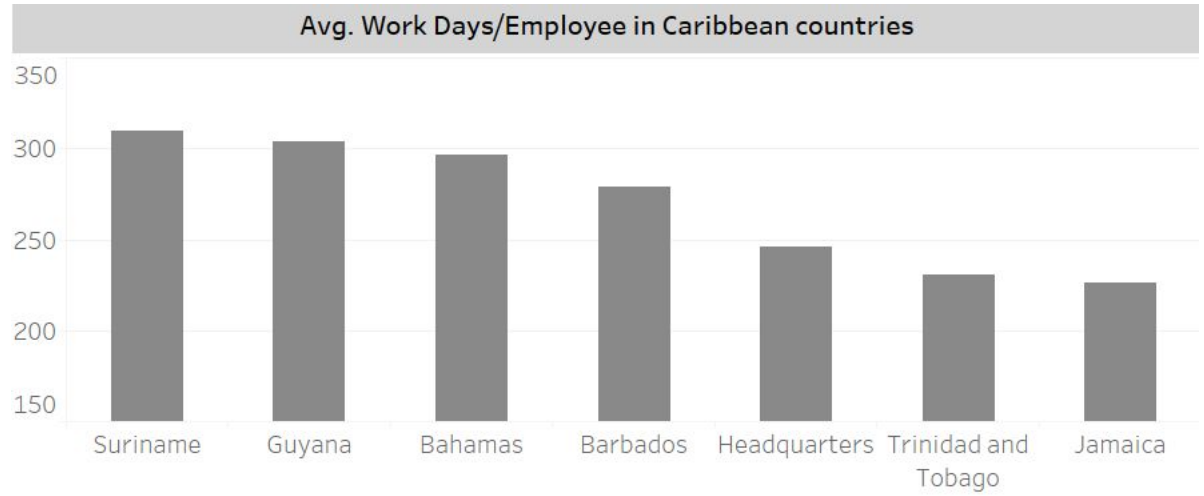


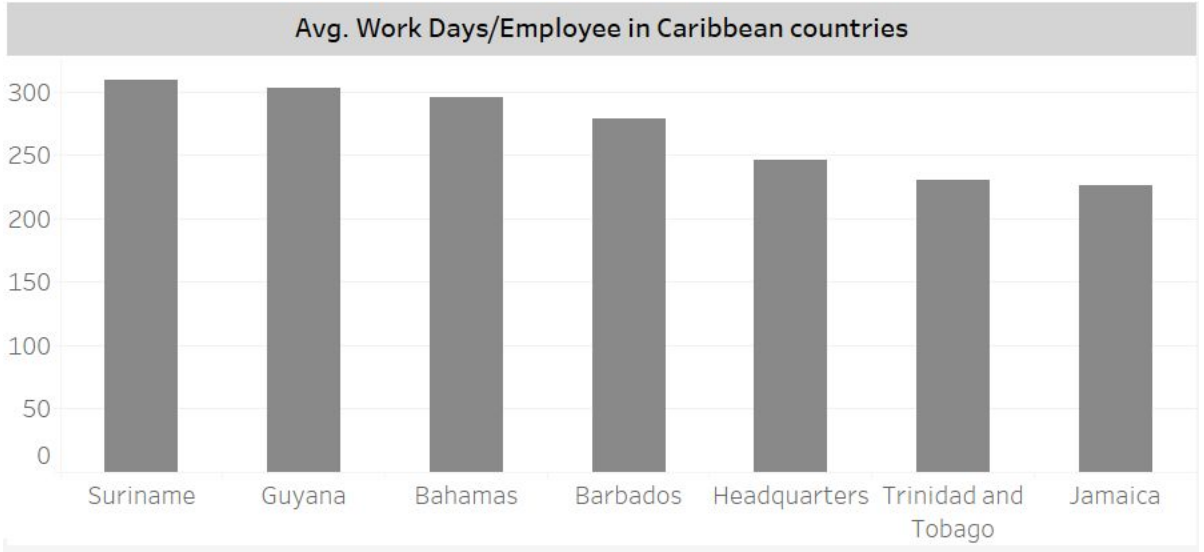


Truncated graphs

- One of the most common manipulations is omitting baselines or beginning the y-axis of a graph at an arbitrary number instead of 0.

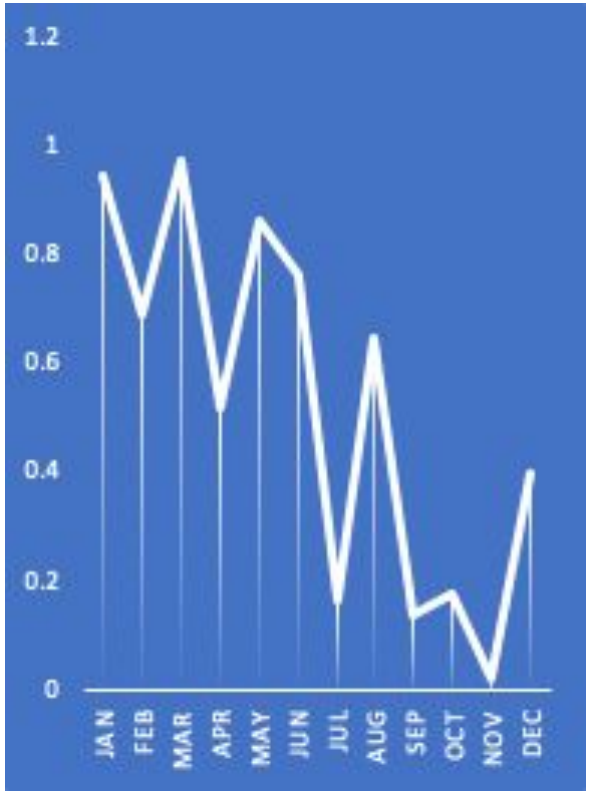
- This creates the impression that there is a significant difference between data points, when in fact, there is relatively little disparity.





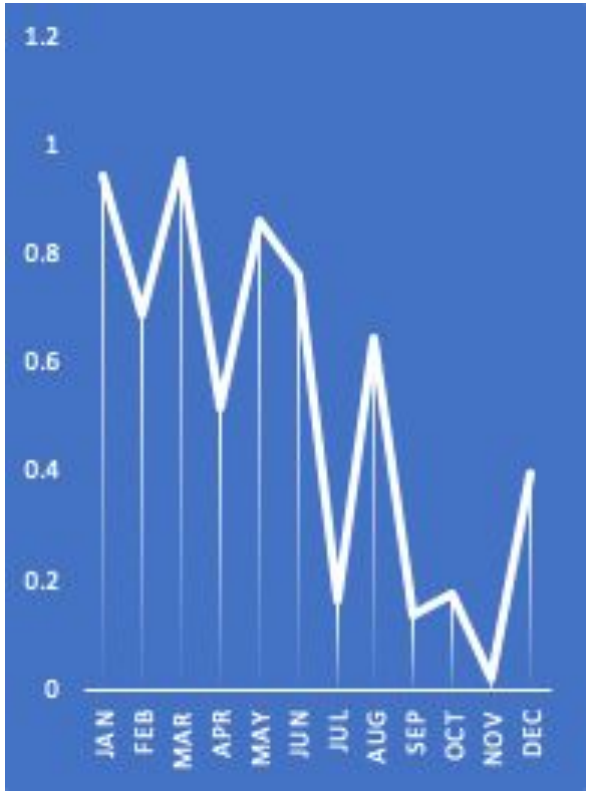
Visual distortions

What distortion has been used in these charts to change how the data appears?



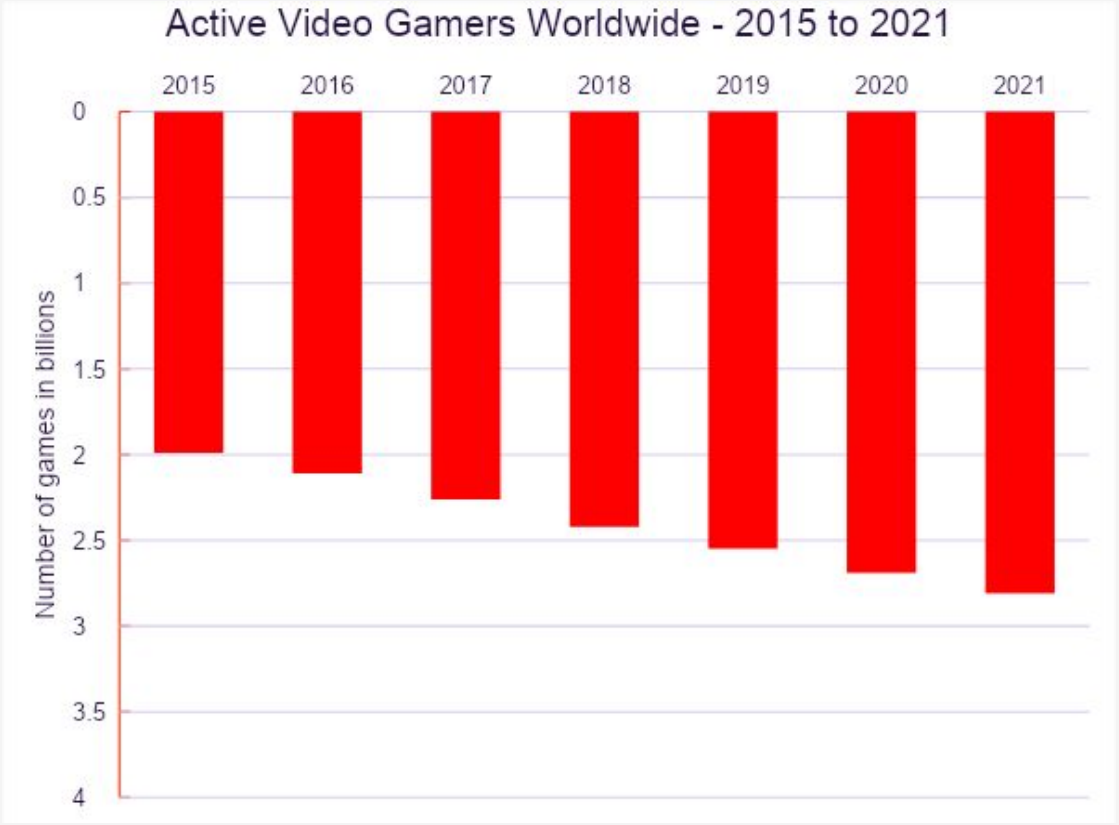
Exaggerated scaling

- Exaggerating the scale of a line graph can easily minimize or maximize the change shown.



Visual distortion

How might this chart be misleading?

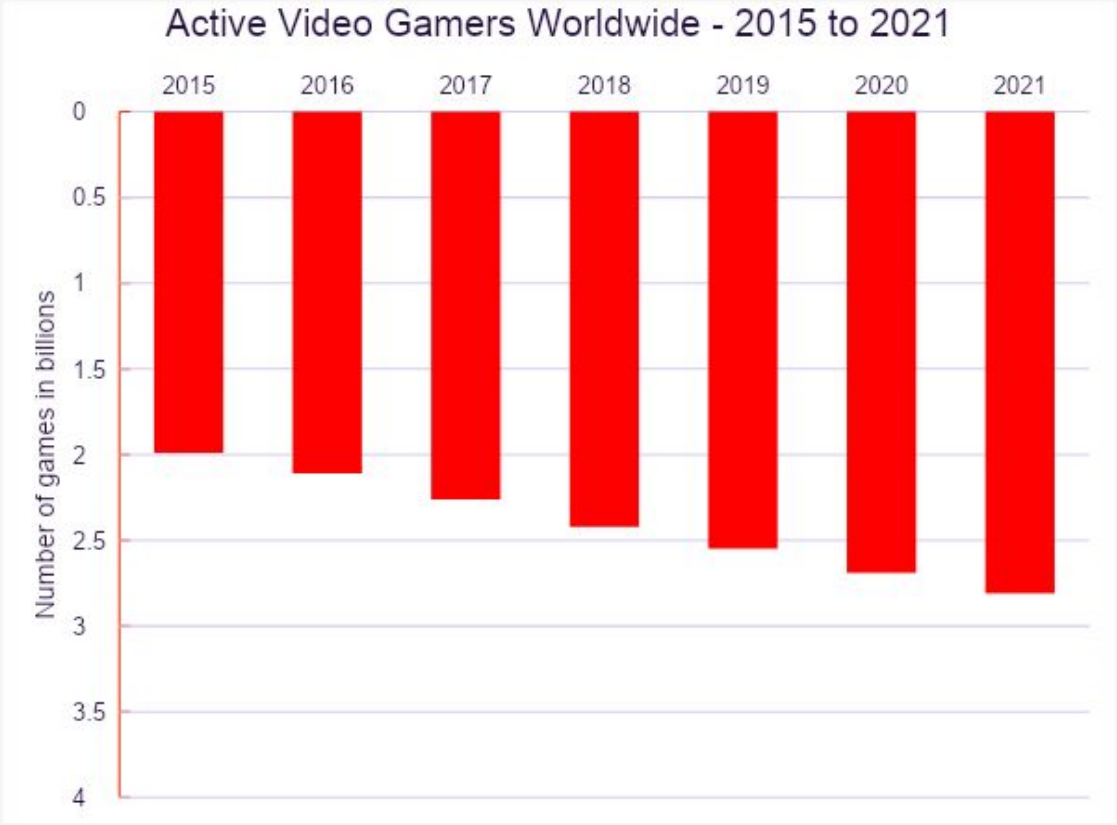


https://financesonline.com/number-of-gamers-worldwide/

Ignoring convention

- Deviating from convention (such as green is positive and red is negative) can create confusion and misinterpretation of the facts.

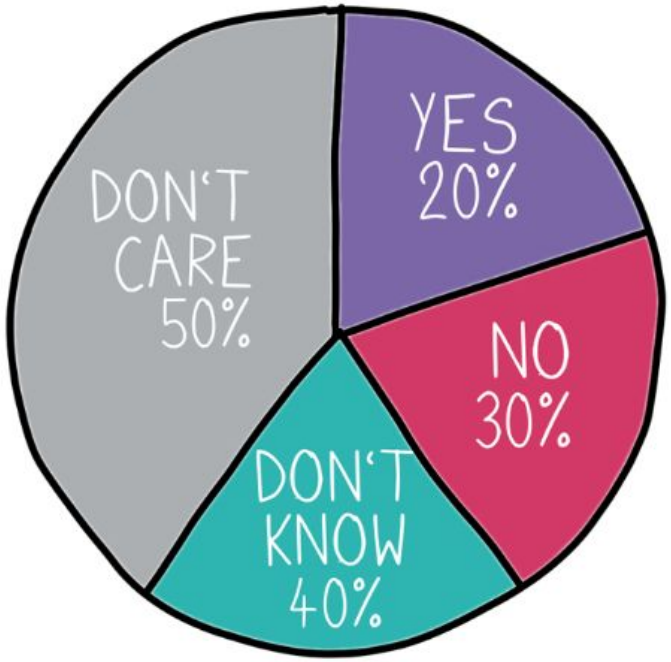
- In this example, the axis also moves downward, making an increase in gamers look like an decrease, at a quick glance.

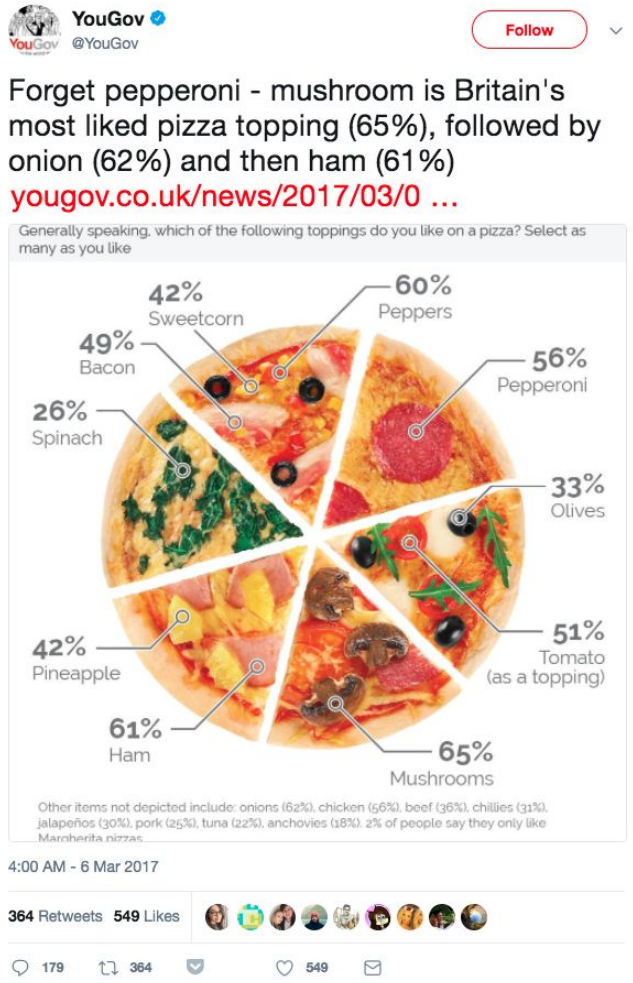


https://financesonline.com/number-of-gamers-worldwide/

Visual distortion

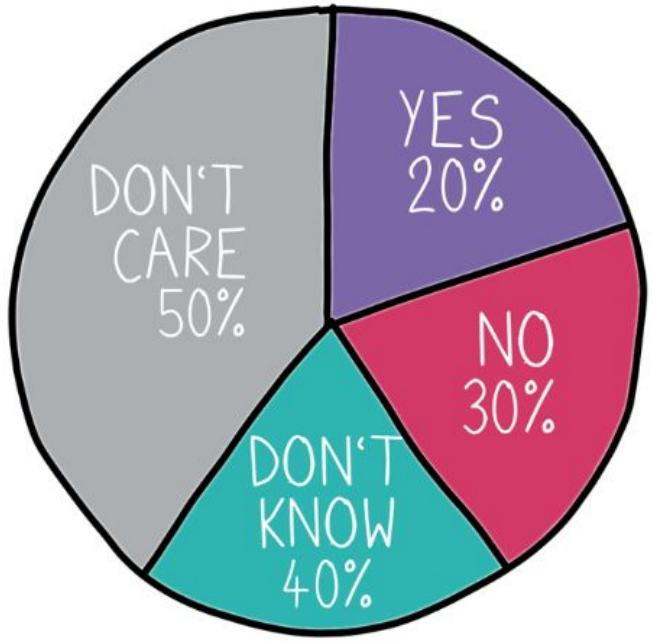
What do you notice about these pie charts?

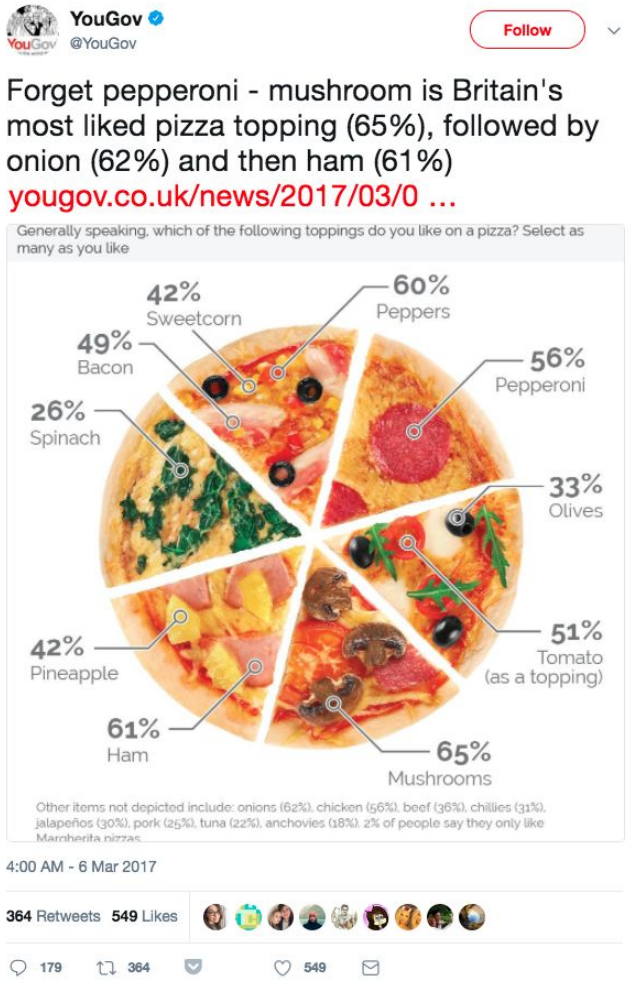




Numbers don’t add up

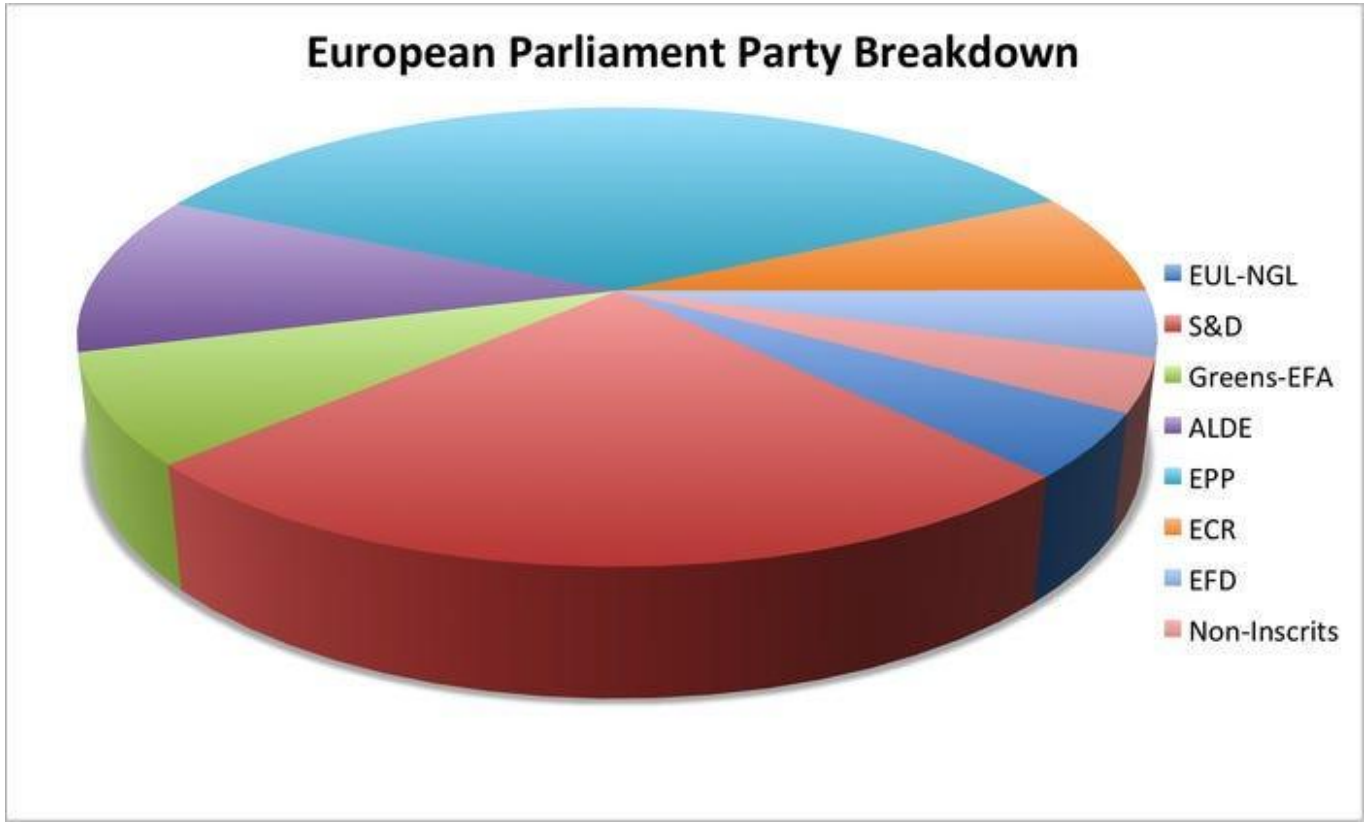
- With pie charts, the sum of each slice must add up to the whole. When the numbers don't add up, you know there's an issue.





Visual distortion

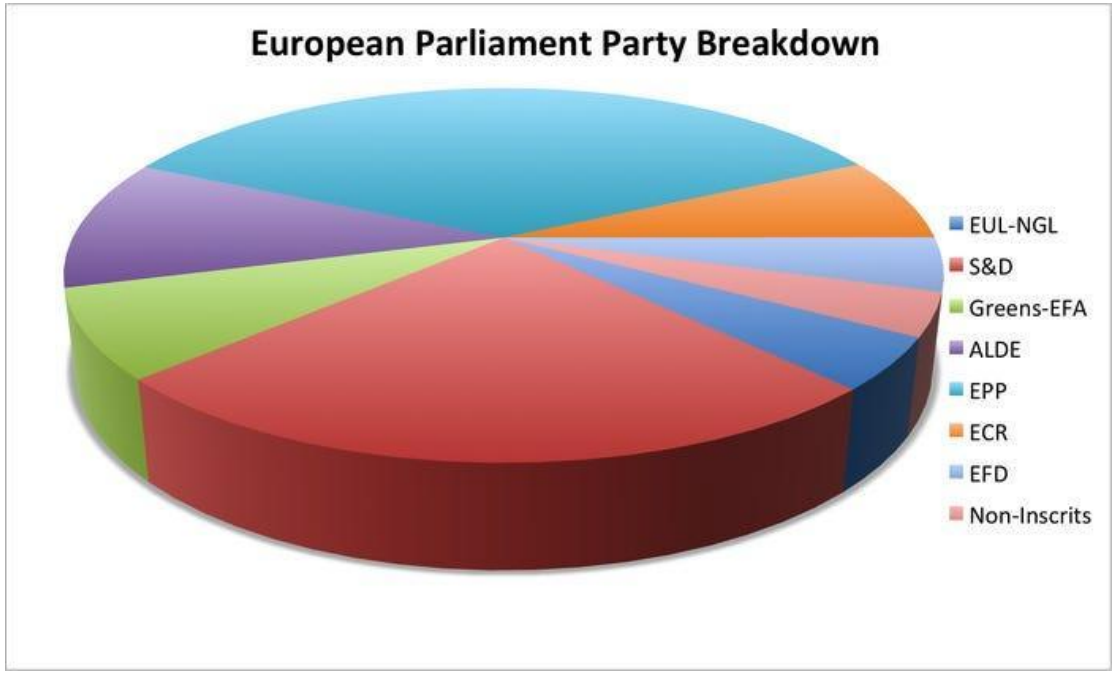
Does the S&D or the EPP party have more representation in parliament?

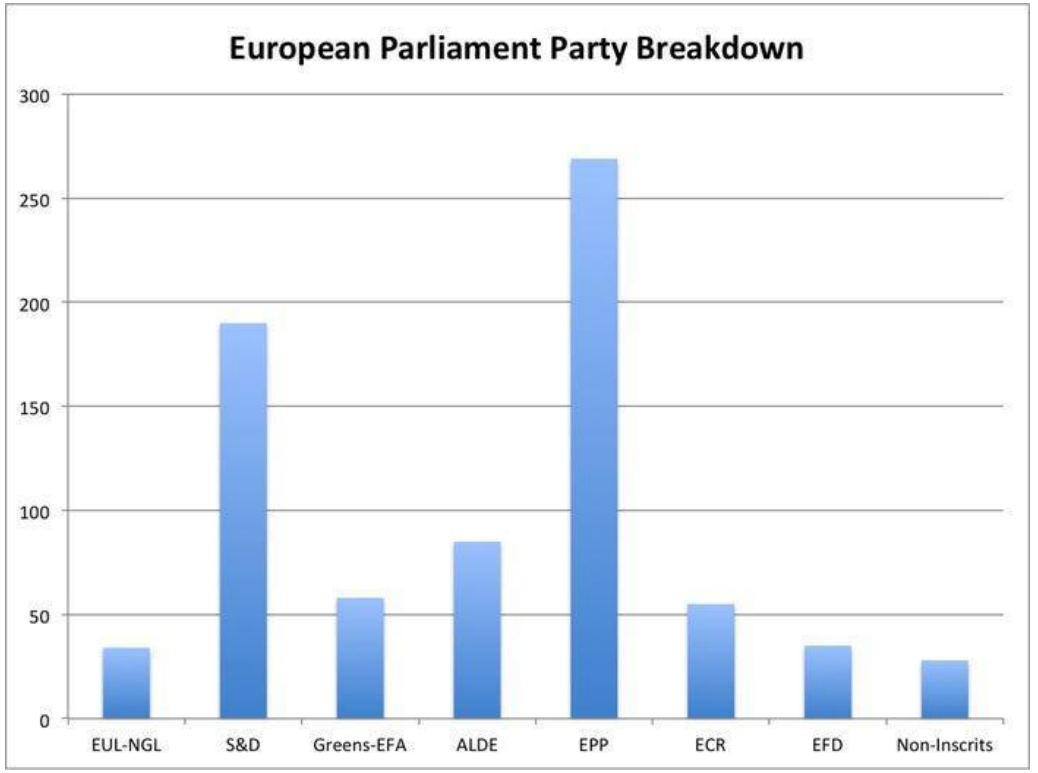


https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

3D distortion

- 3D pie charts can be used to distort and cause a misinterpretation of the data.
- The same data is represented in both charts below.

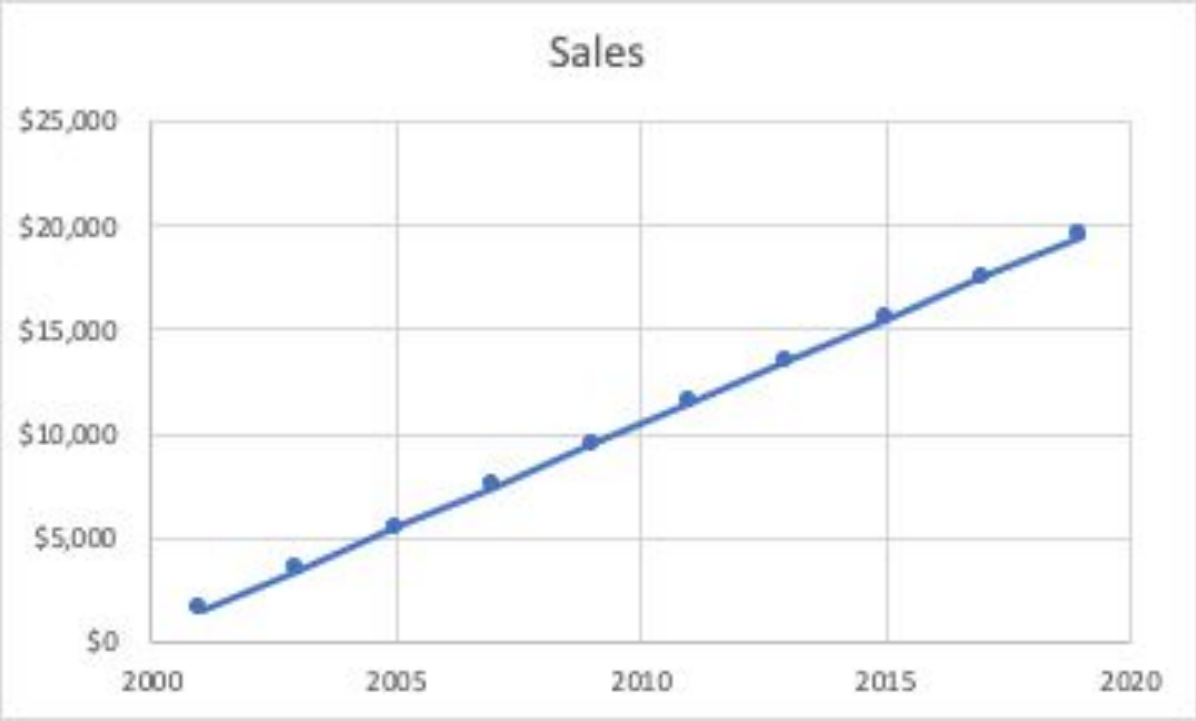


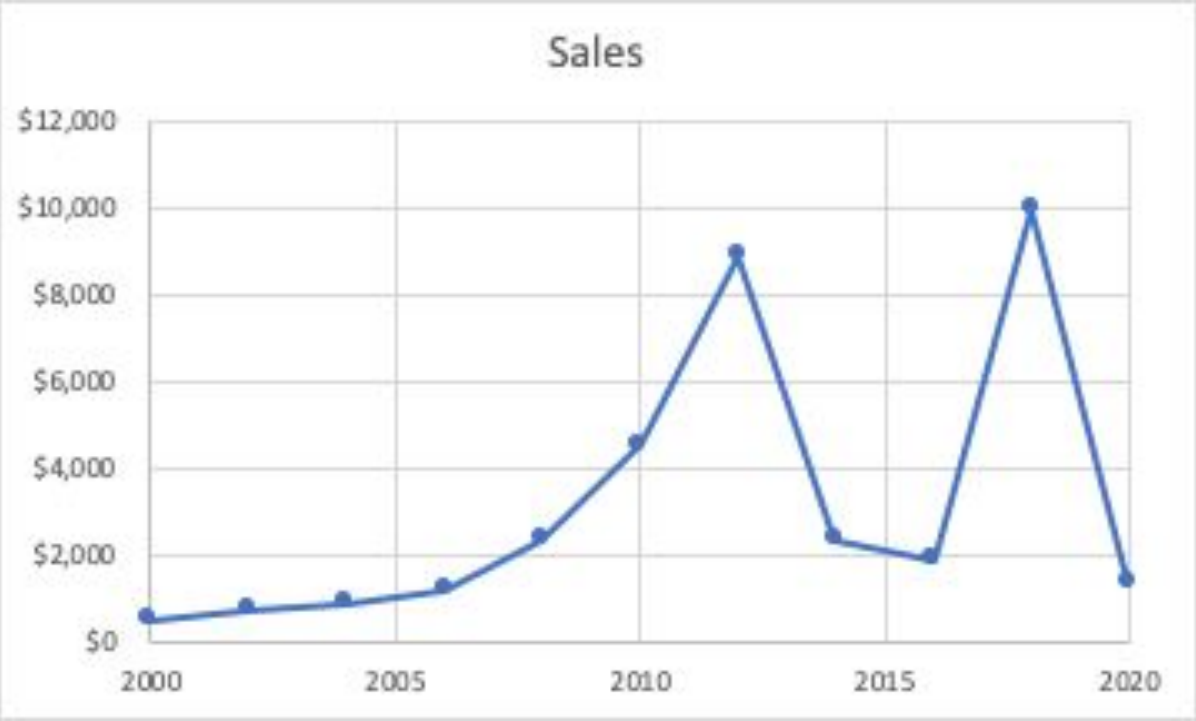


https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

Visual distortion

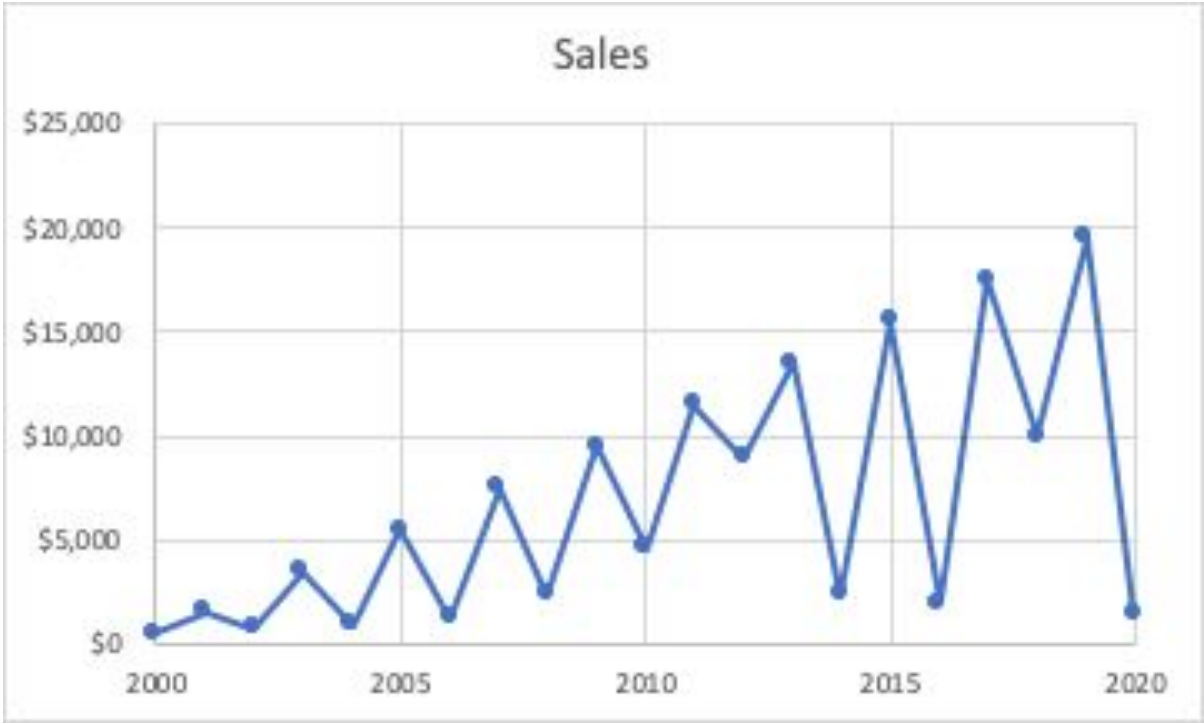
Which company has a better sales trajectory?



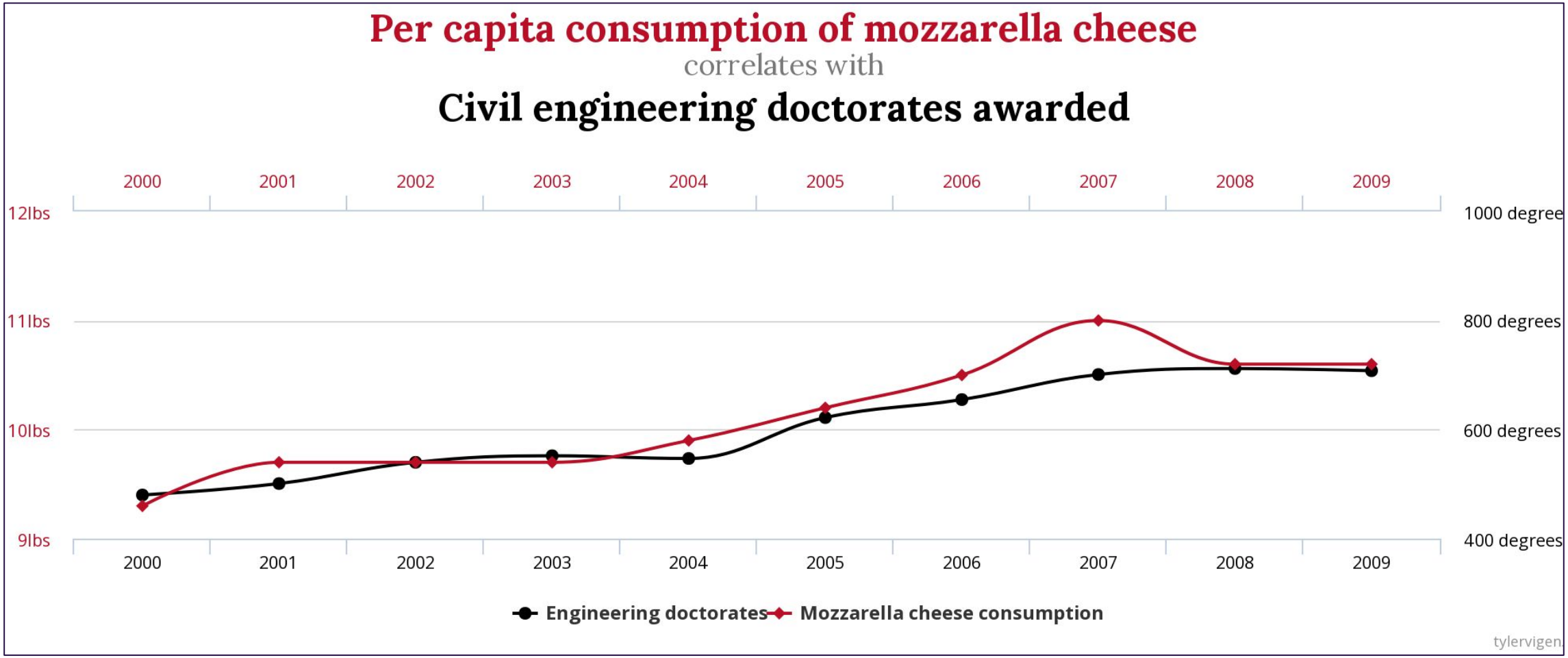


Improper extraction

- Surprise! It's the same company. One graph showed only odd years and the other only even.
- To align to a particular narrative, some may choose to visualize only a portion of the data.
- This is more common in graphs that have time as one of their axes.



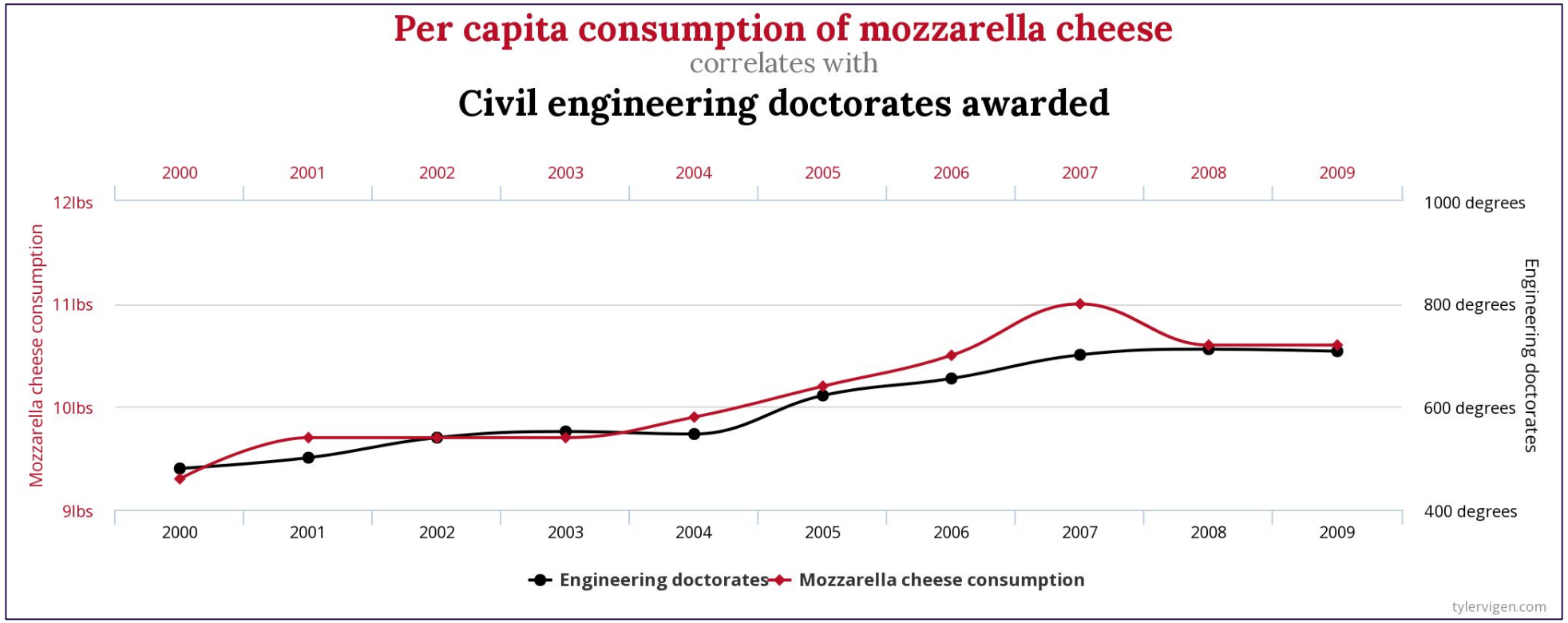
Correlating causation



What story does this visualization tell?

Correlating causation

- Data visualizations can create causal links by the way that data is presented to the viewer.
- However, correlation does not equal causation.



Recap

To avoid being misled, look for:

- misleading statistics
- truncated graphs
- exaggerated scaling
- ignored conventions
- numbers that don't add up
- 3D distortion
- improper extraction
- correlating causation

Thank you!

hello@datasociety.com

1100 15th St. NW, Floor 4
Washington, D.C. 20005

(202)600-9635

datasociety.com