DATASOCIETY:

DATA SCIENCE FOR MANAGERS

Who we are

Data Society's mission is to integrate Big Data and machine learning best practices across entire teams and empower professionals to identify new insights.

We provide:

- High-quality data science training programs
- Customized executive workshops
- Custom software solutions and consulting services

Since 2014, we've worked with thousands of professionals to make their data work for them.

About the course

- Instructor introduction

- Schedule:
  - 4 sessions
  - 11 am – 2 pm
  - 1 or 2 short breaks each session

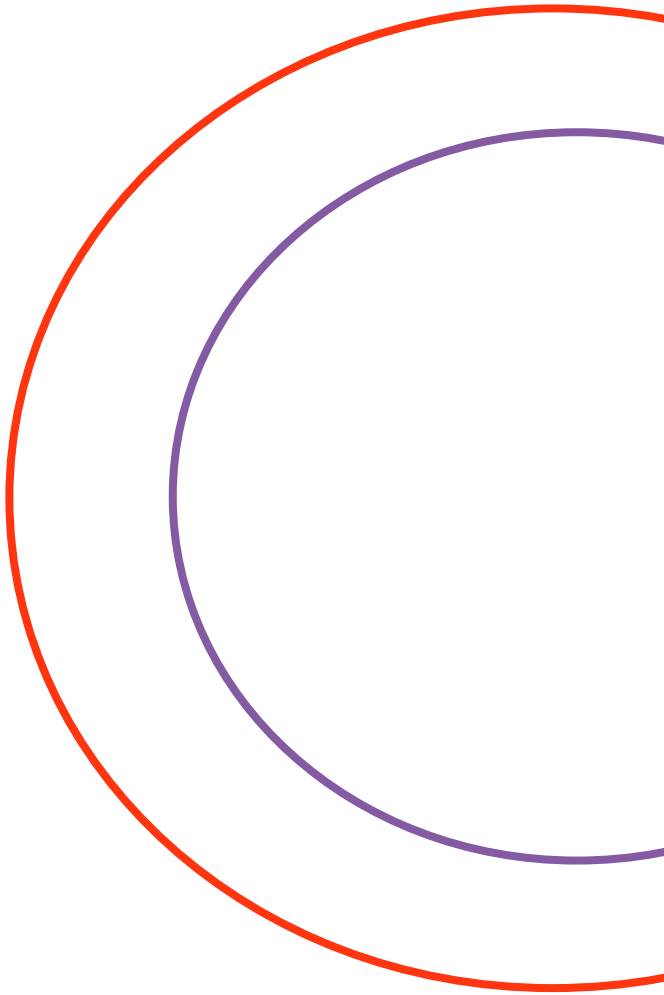
Best practices for virtual learning

1. Find a quiet place, free of as many distractions as possible. Headphones are recommended.

2. Remove or silence alerts from cell phones, e-mail pop-ups, etc.

3. Participate in activities and ask questions.

4. Give your honest feedback so we can troubleshoot problems and improve the course.



Class materials

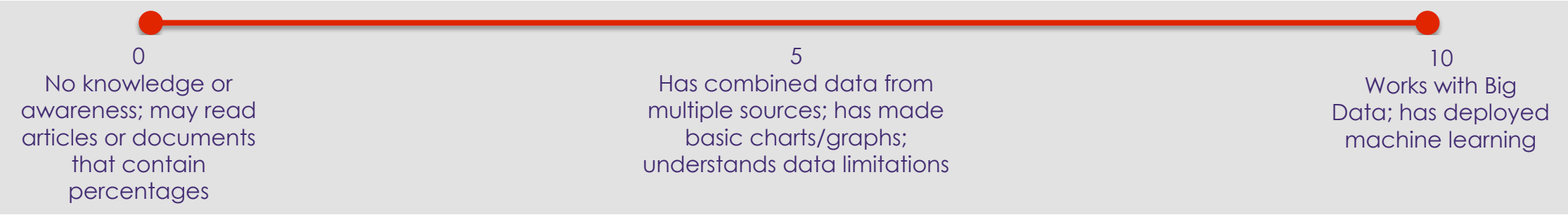
You should have received the following materials:

- Slides

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

Polling question

What you rate your current data literacy level on a scale of 0 -10?



Agenda

Day 1

- Data and its uses
- Data analytics overview
- Data governance
- Data ethics

Day 2

- Building a data-driven culture
- Data tools
- Data teams
- The data science process
- Putting together a project

Day 3

- Foundational data science methods
- Advanced data science methods

Day 4

- Data visualization
- Misleading statistics & visual distortions
- Data storytelling

Agenda

Day 1

- Data and its uses
- Data analytics overview
- Data governance
- Data ethics

- What is data and why should we use it?
- How can data be used in ways that bring value?

What is data?

Definition of *data*

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available
— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published
— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

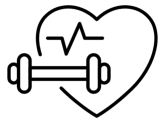
Merriam Webster

Data in our daily life

- Using data to make informed decisions isn't just for business but also personal reasons in our day-to-day lives.

- Let's look some examples how data is collected and used routinely:

Checking reviews before buying a new product



Using fitness tracker to measure your heart rate, calories burnt and to track your progress

Tracking productivity during the day using an application on your phone or laptop

Comparing two car rental organization to find the best deal

Types of data

Structured

y1	x1	x2	x3

Semi-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>
```

Quasi-structured

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.

Unstructured

Sources of data

- HR (performance data, salary/compensation, hiring, 360 view, etc.)
- Network data (application logs, webserver logs, firewall alert logs, e-mails, etc.)
- Clickstream
- ERPs (Enterprise Resource Platforms) - Oracle SAP, etc.
- CRMs (Customer Relationship Management) - SalesForce, Hubspot, etc.
- Webserver
- Contracts/proposals/procurement

External sources of data

- Publicly-accessible APIs
  - e.g., api.data.gov

- Other open data sources
  - e.g., data.worldbank.org

- Large businesses (e.g., Wal-Mart, Best Buy, Trip Advisor, Expedia, Google, and Spotify) are increasingly giving people access to their data

- Data is sometimes available for purchase (e.g. weather data)

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 include:
  - High volume. Typically, the size of big data is described in terabytes, petabytes, even 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.

Why use data?

Data may be collected, retained, and used for several reasons:

- Compliance: avoiding penalties
- Automation: economic efficiencies
- Analytics: insights

What can using data do?

1. Find a needle in haystack

2. Prioritize work for high impact

3. Provide early warning / detection

4. Speed up decisions

5. Optimize resources

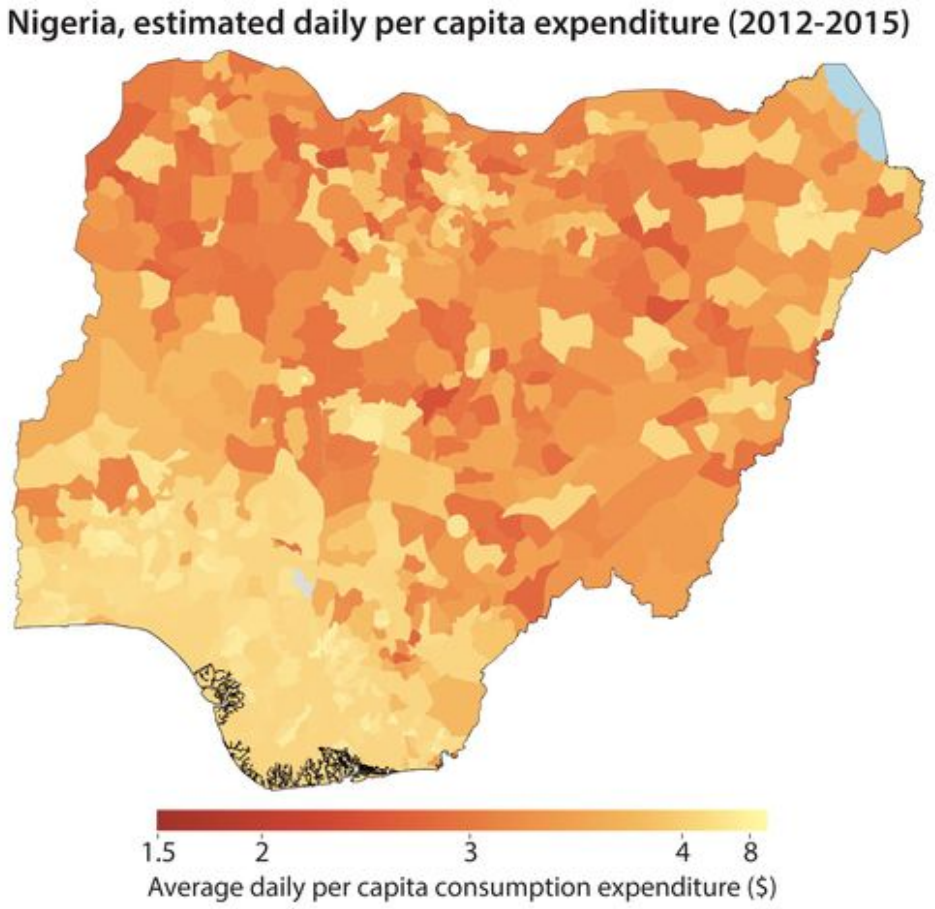
6. Enable experiments

Find a needle in a haystack

- Stanford is using satellite imagery and predictive analytics to estimate consumption expenditures and asset wealth.

- This could transform efforts to track and target poverty in developing countries with existing, public data.

http://sustain.stanford.edu/predicting-poverty/



Prioritize work for high impact

- Consultants in Philadelphia developed a model for prioritizing building inspections based on a location's:
  - Distance to nearby vacant properties
  - Distance to certain crimes
  - Distance to infestation reports

- Benefits could include generating better daily inspection routes or providing more information to inspectors on existing routes.

http://urbanspatialanalysis.com/portfolio/proof-of-concept-using-predictive-modeling-to-prioritize-building-inspections/

Provide early warning / detection

- When individuals and groups are planning criminal activity, they often signal their intentions online via open-source social media.

- Tactical Institute uses cognitive analytics to monitor social channels 24x7, analyze billions of comments and posts, home in on threats, and identify perpetrators before they can act.

- They then provide real-time notification of threats issued so that clients can take pre-emptive action before the threat is executed.

Tactical Institute

https://www.ibm.com/case-studies/tactical-institute

Speed up decisions

- Recruiting chatbots are used to automate the communication between recruiters and candidates. They are useful:
  - when there is a high number of applicants
  - to ensure that similar questions are asked of all candidates
  - for answering frequently asked questions effectively

- JobAI is a German recruiting chatbot. Their platform offers jobseekers the opportunity to contact companies, inform themselves, and apply via familiar messenger apps such as WhatsApp and Telegram.

https://jobai.de/

Optimize resources

- BNY Mellon developed and deployed more than 220 automated computer programs in 2016 and 2017.
- These "bots" carry out repetitive tasks such as formatting requests for dollar funds transfers and responding to data requests from external auditors.
- The bank estimates that its funds transfer bots alone are saving it $300,000 annually.
- Bots that reply to information requests on financial statements from auditors enabled the bank to cut down its response time to 24 hours from 6 to 10 business days

https://www.reuters.com/article/us-bony-mellon-technology-ai-idUSKBN186253

Enable experiments

- The NYC government reduced the number of people who fail to appear (FTA) in court using data to evaluate options.
- The cost of a one-time court summons' redesign corresponded to a 13% drop in FTAs.
- When paired with a text message costing $0.0075 per message, there was a 36% decrease.

OLD

NEW

https://www.sciencemag.org/news/2020/10/new-york-city-uses-nudges-reduce-missed-court-dates

Polling question

The most relevant use of data for my organization is:

- Finding a needle in haystack
- Prioritizing work for high impact
- Speeding up decisions
- Optimizing resources
- Enabling experiments
- Providing early warning/ detection

Chat question

What hurdles might you face trying to implement a data analytics project in your organization?

Agenda

Day 1

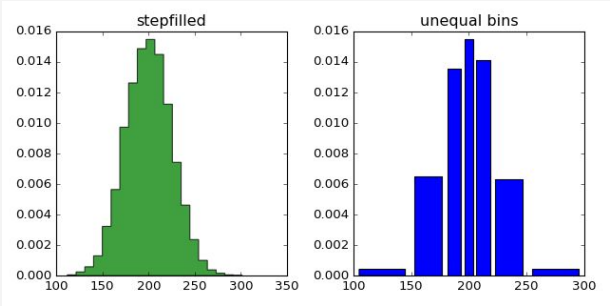
- Data and its uses
- Data analytics overview
- Data governance
- Data ethics

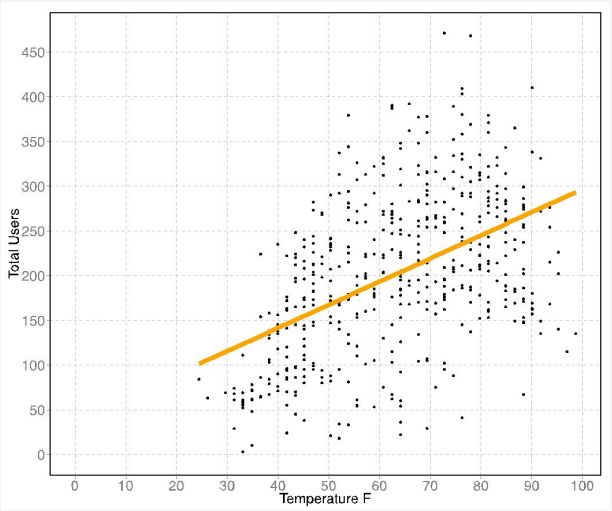
- What is data analytics and how can it be used?
- What are the principles of data science?

What is data analytics?

- Data analytics 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.

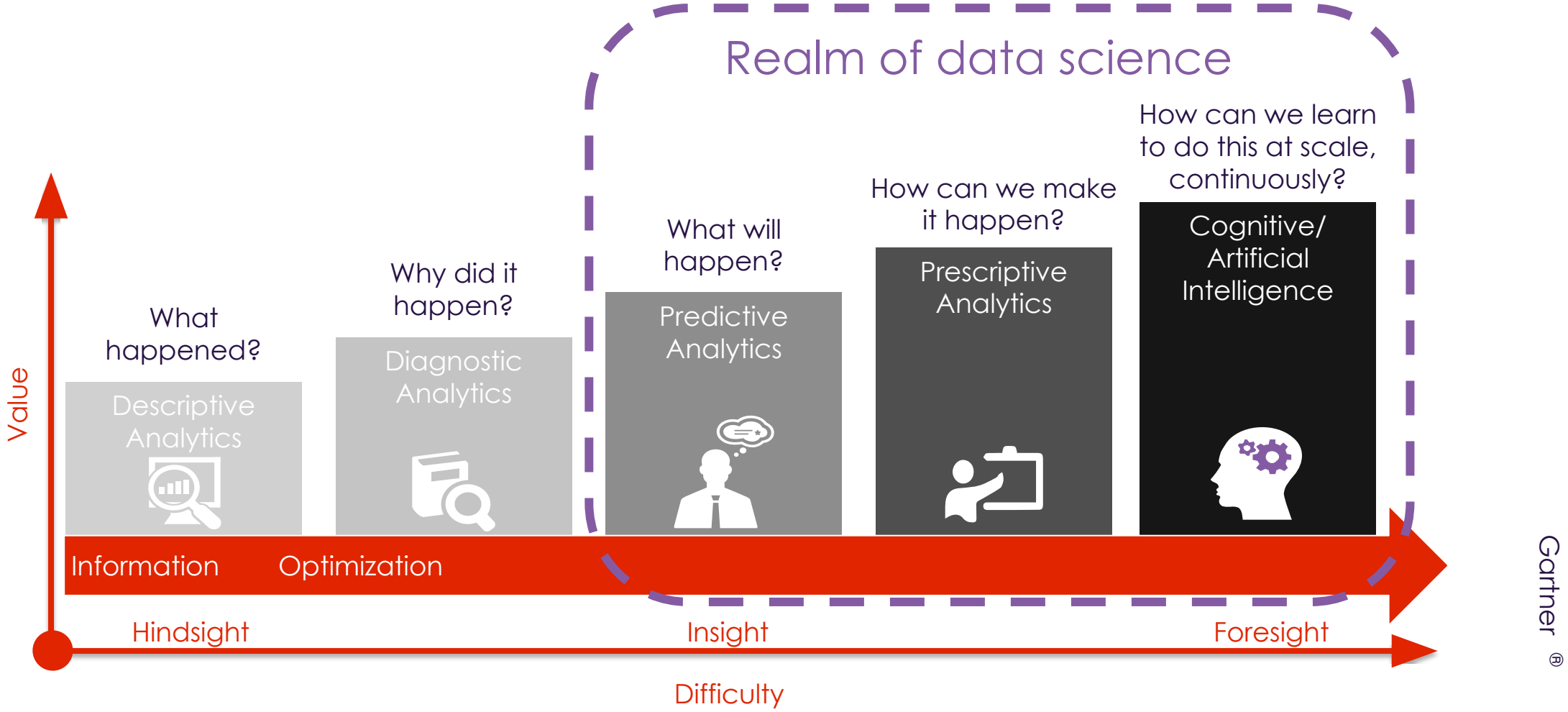




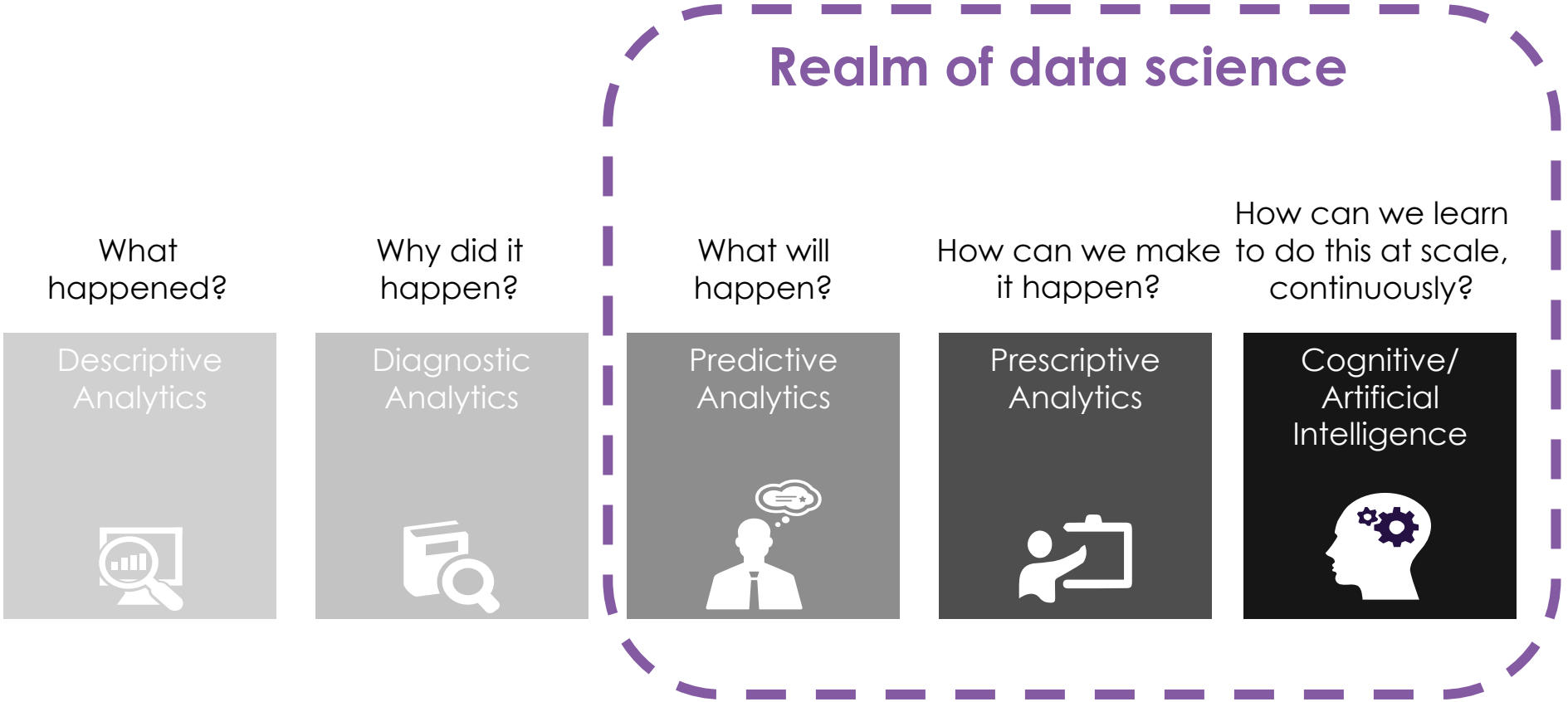
Chat question

How do you use data analytics within your organization currently?

Data analytics maturity model



Model revisited



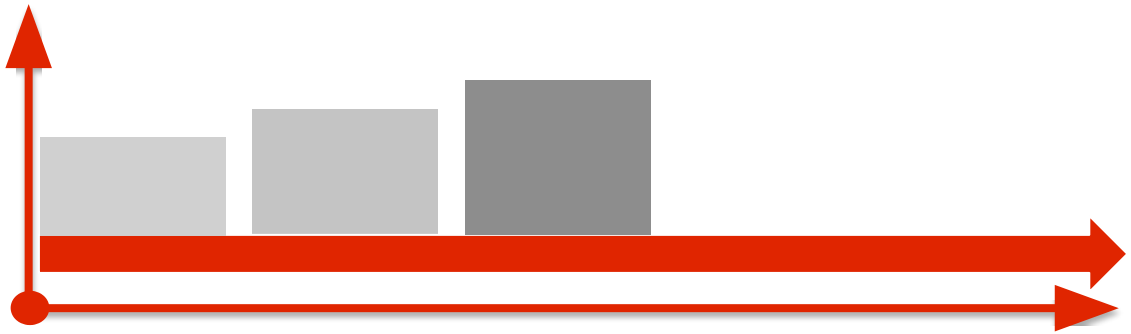
Stage 1: descriptive analytics

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

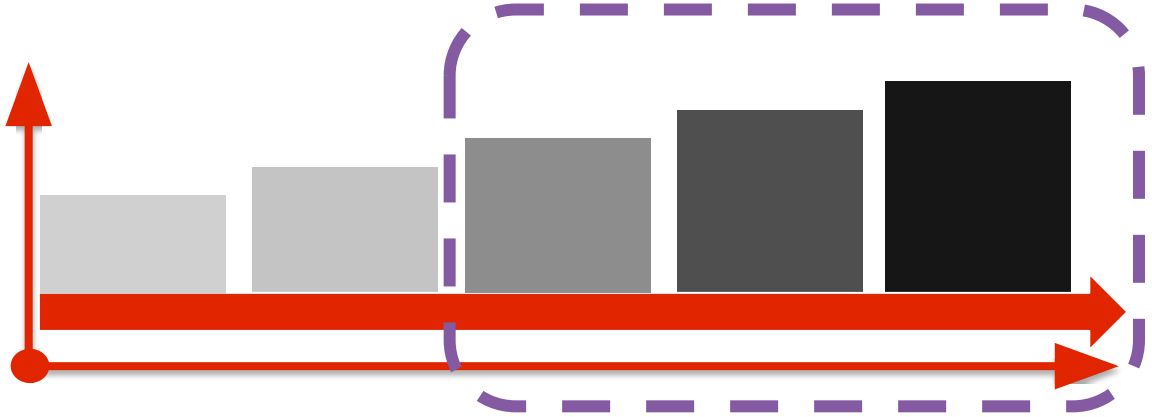
What questions does it answer?	Why did something happen in the past?
How valuable is it?	Provides insights into a particular problem, and can help you identify some root causes for past trends and behaviors
How labor intensive is it?	Requires detailed data, but doesn't have to be overly intensive

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

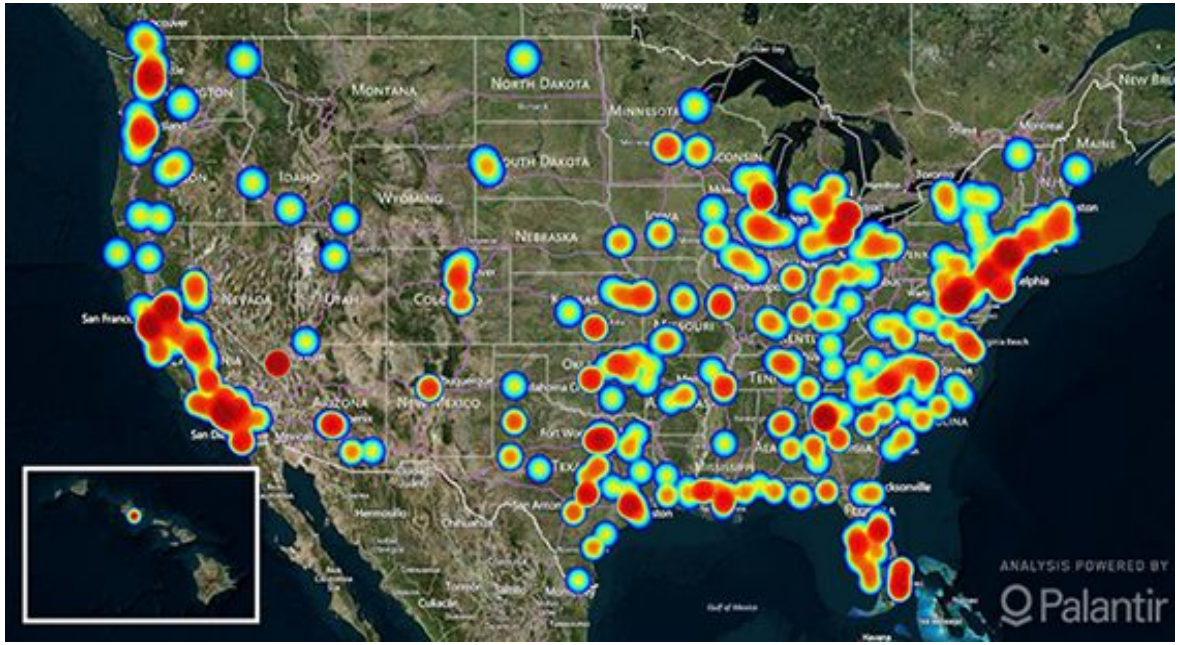
Stages 4, 5: prescriptive analytics, AI



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

Example: fighting human trafficking

- Polaris has made a connection between massage parlors and human trafficking.
- Once they find one owner of an illicit massage business by tracing business records, they often find that he owns several more businesses in the area.
- They are now able to use data to identify illicit activities and alert law enforcement.



https://www.datanami.com/2016/10/07/data-analytics-fight-human-trafficking/

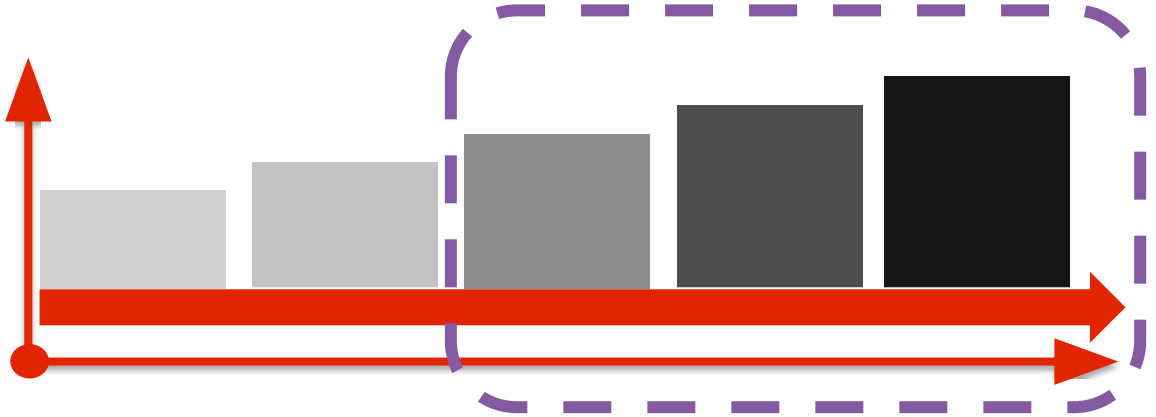
Polling question

What type of analytics is demonstrated when Polaris uses data to identify possible illicit activities and alert law enforcement?

- Descriptive
- Diagnostic
- Predictive
- Prescriptive

How do we move forward?

- To reach the realm of data science organizations require:
    - quality data
    - an innovative environment
    - resources, with the requisite knowledge and technical skillsets to use them



Break

Agenda

Day 1

- Data and its uses
- Data analytics overview
- Data governance
- Data ethics

- What is data governance? Why is it important?
- What do Federal managers need to know about data governance?

Quality data is “clean”

Clean data is:

- Valid
- Accurate
- Consistent
- Complete
- Uniform

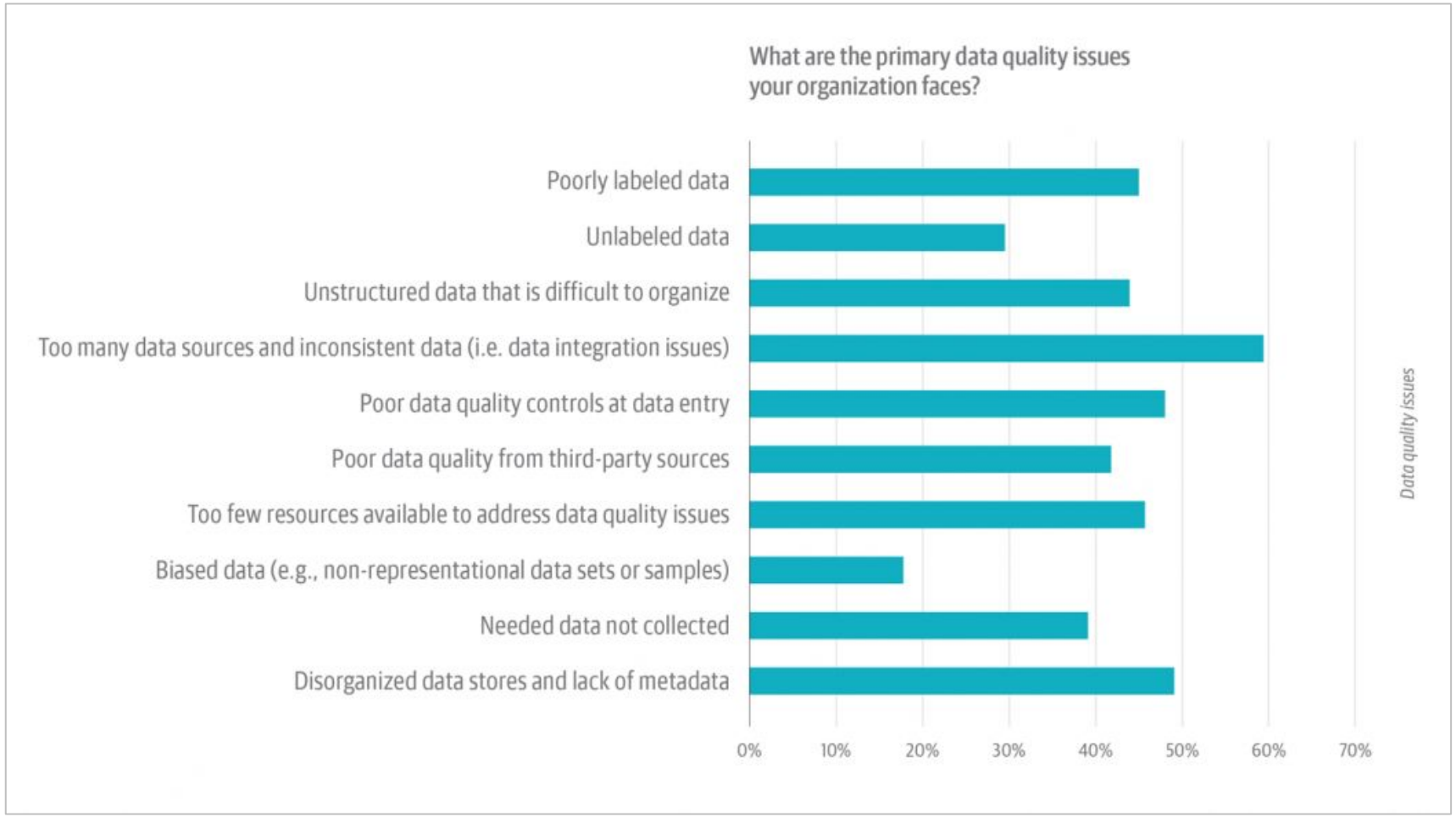
Clean data is **not**:

- Corrupt
- Incorrect
- Duplicate
- Incomplete
- Wrongly formatted

Acquiring quality data is hard

2019 O'Reilly survey of more than 1,900 leaders and data professionals

https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/



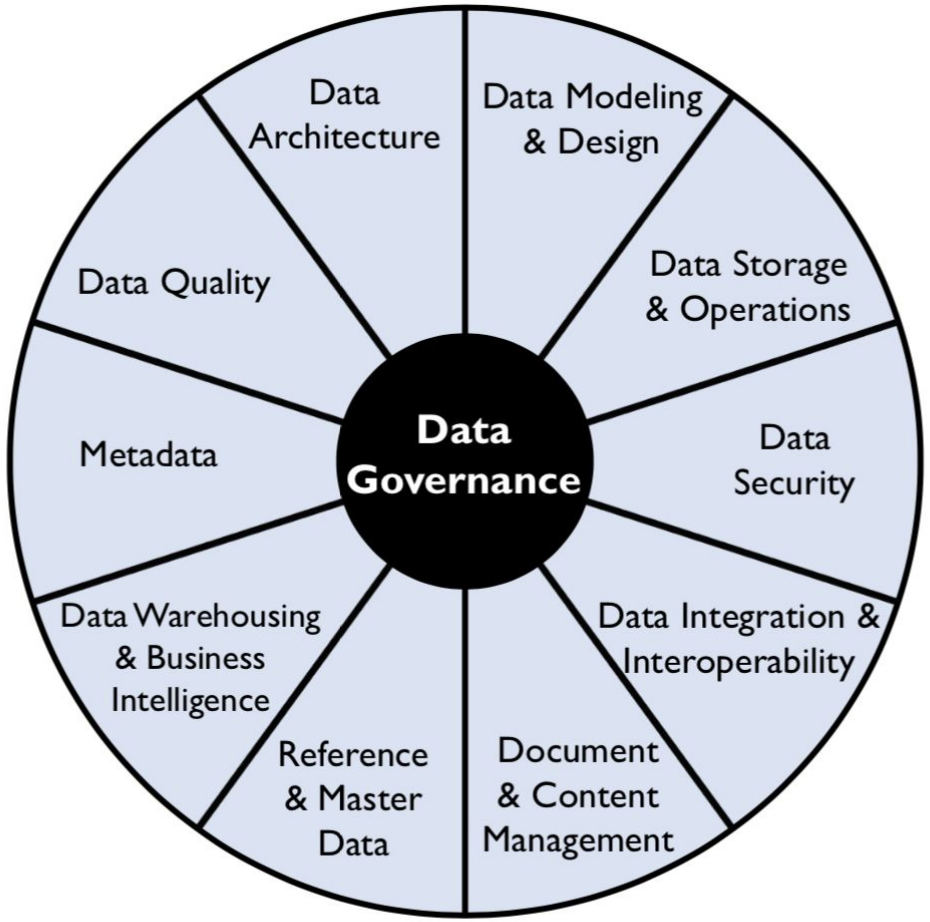
Controlling data

- Data is arguably the most important asset that organizations have.
- Controlling it through data governance practices and processes helps to ensure that data is usable, accessible, and protected.
- Effective data governance helps to **avoid data inconsistencies and errors** in data, plays a role in the organization's ability to **comply with laws and regulations**, and **increases access** to data.

What is data governance?

- Data governance is a collection of practices and processes that help to ensure the formal management of data assets within an organization.

- It encompasses the complete life cycle of IT investment, from strategic planning to the day-to-day operations of the IT function.



DAMA-DMBOK2 Data Management Framework

What is data governance?

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

Why is data governance important?

- Regulatory compliance – with increased regulation comes compliance that needs to be implemented and followed
- Reduce risk – effective data governance enhances data security and privacy
- Improve processes – when everyone follows the same standards, projects and management become more efficient

Data governance principles

A data governance program should be:

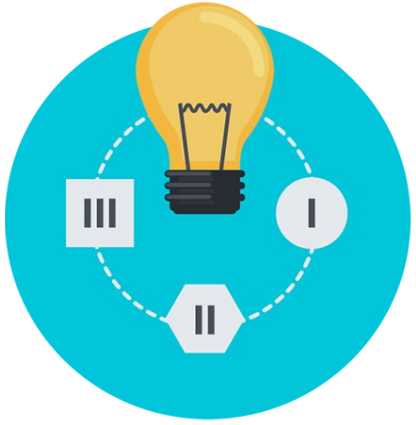
- Sustainable –it survives beyond the initial implementation
- Embedded – data governance should be present in all processes related to data
- Measured – there should be some defined metrics to help demonstrate value to the organization

Data governance strategy

A data governance program might be documented using:

Charters

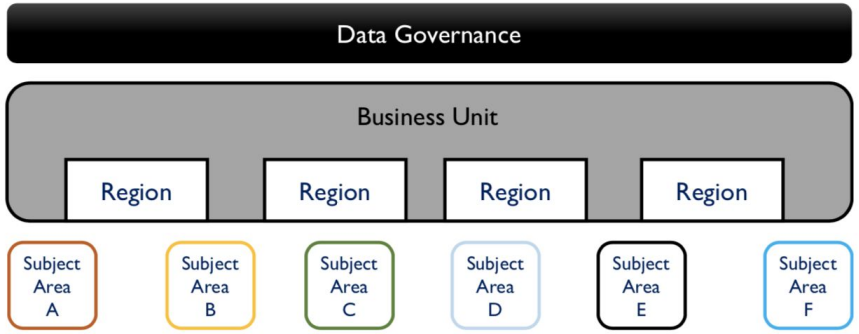
Implementation roadmaps



Operating frameworks / accountabilities

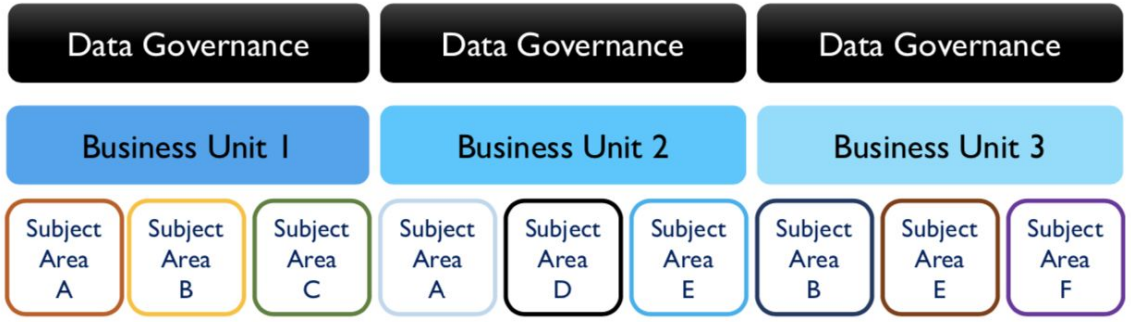
Plans for operational success

Data governance models



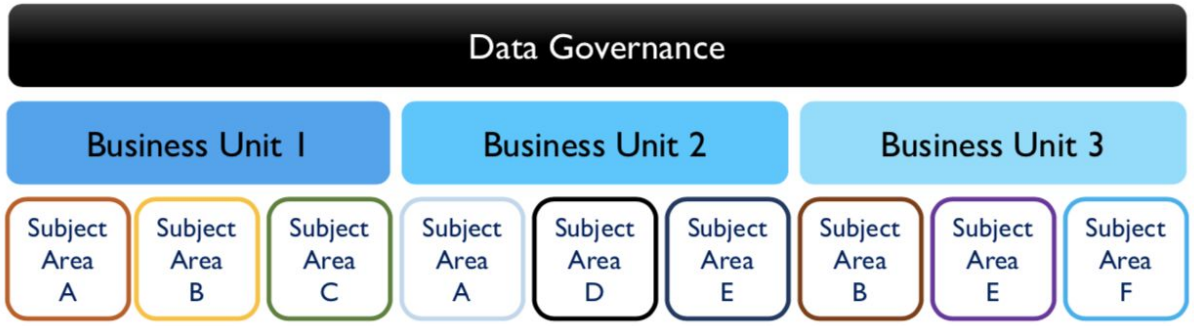
Centralized

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.

Poll question: data governance

Which governance model do you think is suitable for your organization?

- Centralized
- Replicated
- Federated
- None of the above

Poll question: data governance

After purchasing three companies, an organization is interested in ensuring high quality data across the enterprise, which analytics governance strategy will probably best support that goal?

- Centralized
- Replicated
- Federated
- None of the above

Data governance: process maturity

Level	Description
0 Non-existent	Complete lack of any recognizable processes; have not even recognized that there is an issue to be addressed
1 Initial / ad hoc	Enterprise has recognized that the issues exist and need to be addressed but there are no standardized processes (only ad hoc approaches); overall approach to management is disorganized
2 Repeatable but intuitive	Processes have developed to the stage where similar procedures are followed by different people undertaking the same task; no formal training or communication; high degree of reliance on the knowledge of individuals and, therefore, errors are likely
3 Defined	Procedures have been standardized, documented, and communicated; it's mandated that these processes should be followed; however, it is unlikely that deviations will be detected
4 Managed and measurable	Management monitors and measures compliance with procedures and takes action where processes appear not to be working effectively; processes are under constant improvement and provide good practice; automation and tools are used in a limited or fragmented way
5 Optimized	Processes have been refined to a level of best practice, based on the results of continuous improvement and maturity modelling with other enterprises; IT is used in an integrated way to automate the workflow, providing tools to improve quality and effectiveness, making the enterprise quick to adapt

What do leaders need to know?

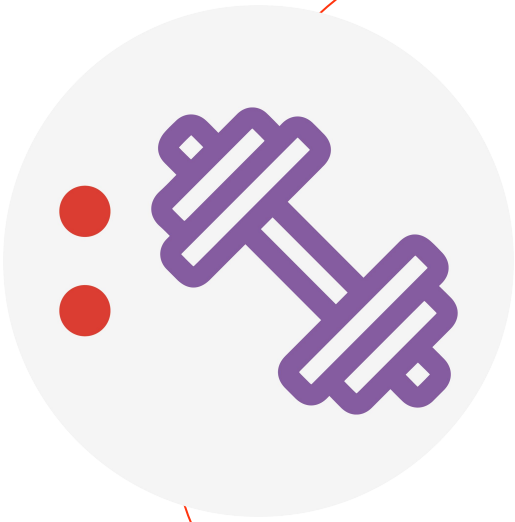
- Target maturity levels would be expected to vary for individual IT processes, IT infrastructure, and industry characteristics

- Differences in maturity come from factors such as the risks facing the enterprise and the contribution of processes to value generation and service delivery

- It does not make sense to be at level 5 for every IT process because the benefits could not justify the costs of achieving and maintaining that level

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.



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.

Agenda

Day 1

- Data and its uses
- Data analytics overview
- Data governance
- Data ethics

- What are data ethics?
- What does a Federal manager need to know about data ethics?

What is data ethics?

Data ethics is a newer branch of ethics that studies and evaluates moral problems related to:

- Data (including generation, recording, curation, processing, dissemination, sharing, and use)
- Algorithms (including artificial intelligence, artificial agents, machine learning, and robots)
- Corresponding practices (including responsible innovation, programming, hacking, and professional codes)

Source: University of Oxford

Why data ethics?

- Data science has huge opportunities, but those opportunities are accompanied by complex data ethical challenges.
    - To formulate and support morally good solutions (e.g., right conducts or right values)
    - To maximize the value of data science for our societies, for all of us and for our environments

The best single thing you can do to further data ethics is to talk about data ethics!

Source: University of Oxford

FDS: Data Ethics Framework

- In December 2020, GSA published a Data Ethics Framework.

- The Framework's purpose is to guide federal leaders and data users as they make ethical decisions when acquiring, managing, and using data to support their agency's mission.

Federal Data Strategy

Data Ethics Framework

https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

Federal Data Ethics Tenets

Existing frameworks

- O'Reilly's 5 Cs: consent, clarity, consistency, control, consequences

- UK Government Data Ethics Framework
  1. Start with clear user need and public benefit.
  2. Be aware of relevant legislation and codes of practice.
  3. Use data that is proportionate to the user need.
  4. Understand the limitations of the data.
  5. Ensure robust practices and work within your skillset.
  6. Make your work transparent and be accountable.
  7. Embed data use responsibly.

- GDPR regulations developed in Europe to help individuals control their data

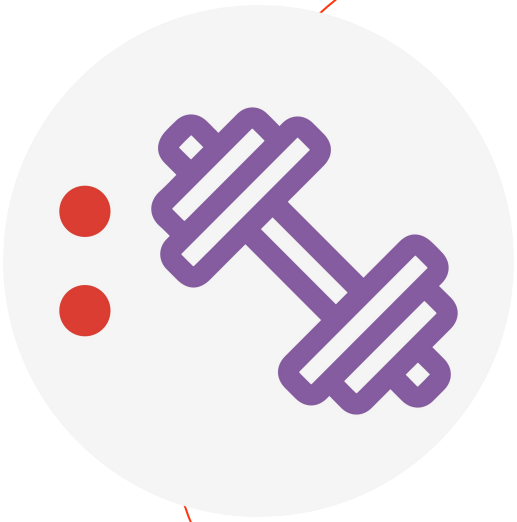
Data Society guidelines

1. Ownership: Who owns the data? Do you have the right to collect the data?
2. History: How long can you store the data?
3. Privacy: Who controls access to the data?
4. Uses: What kinds of inferences can you make?
5. Math: How do you prevent machine learning algorithms from learning the biases of the past? Understanding how the math works is imperative for ethical data science!

Activity: data ethics

- Turn to **page 9** of your participant guide to the Data ethics activity.

- Read the scenario excerpted from the Data Ethics Framework and answer the questions that follow.



End of Day 1

Data and its uses
Data analytics overview
Data governance
Data ethics

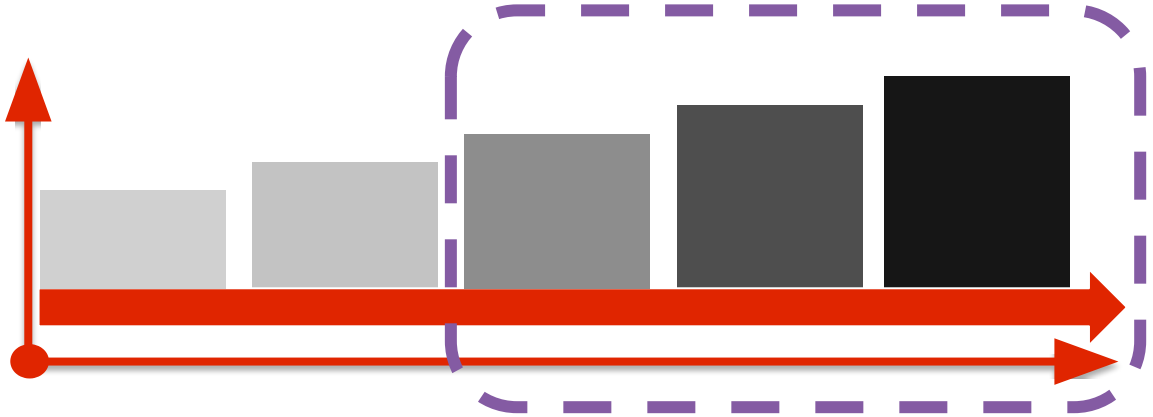
DATASOCIETY:

DATA LITERACY FOR MANAGERS

Day 2

Recap

- To reach the realm of data science organizations require:
    - quality data
    - an innovative environment
    - resources, with the requisite knowledge and technical skillsets to use them



Agenda

Day 2

- Building a data-driven culture
- Data tools
- Data teams
- The data science process
- Putting together a project

- What is a data-driven culture?
- How can a Federal manager encourage data-driven practices?

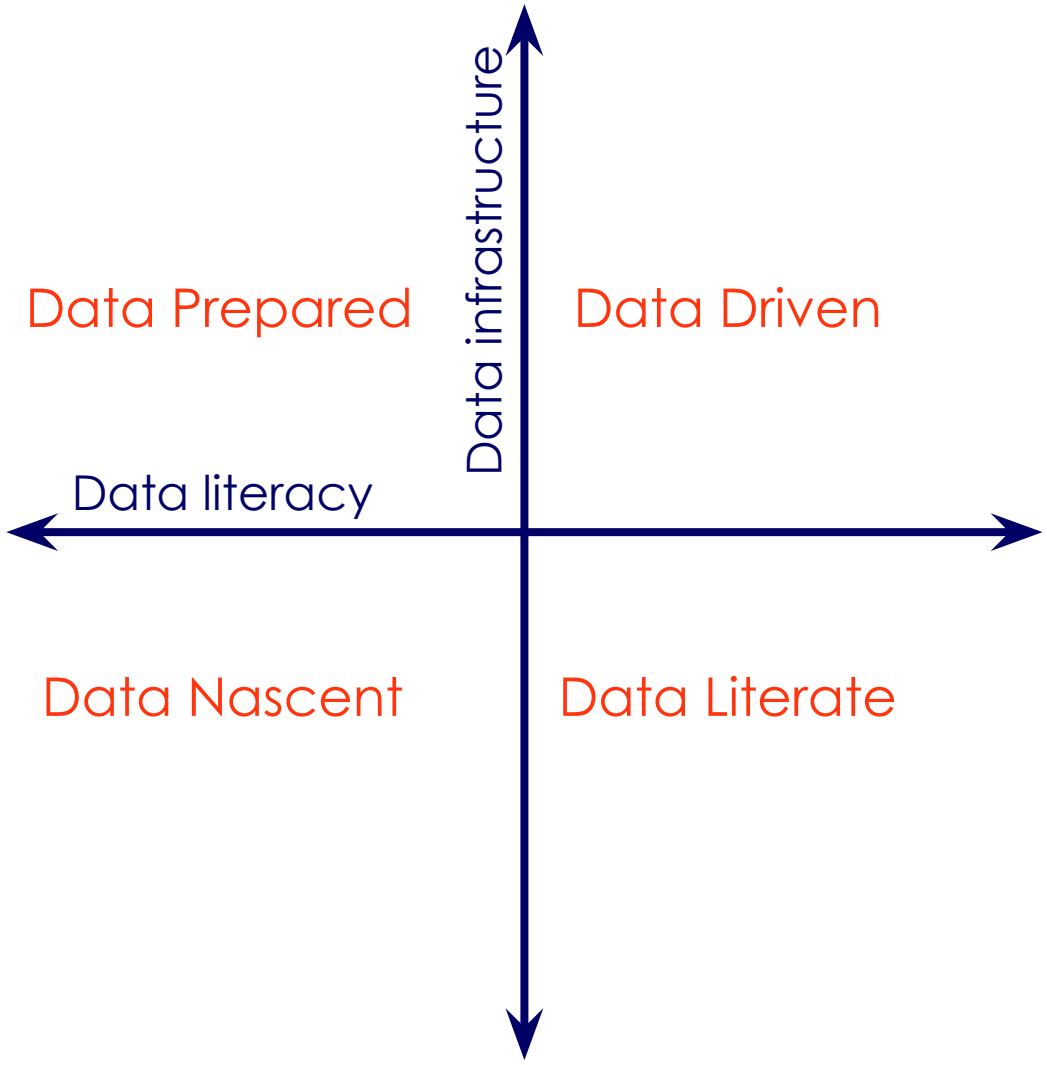
Chat question

What does it mean to be data driven?

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

DATA ACCESS

Can staff access data easily and in a timely manner?

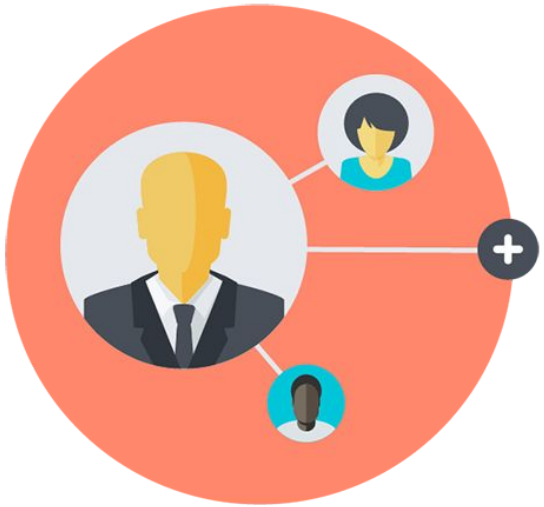
DATA STORAGE

Is the data stored securely with a backup?

DATA COLLECTION

Is data collected in a timely and clean way?

Data literacy



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.

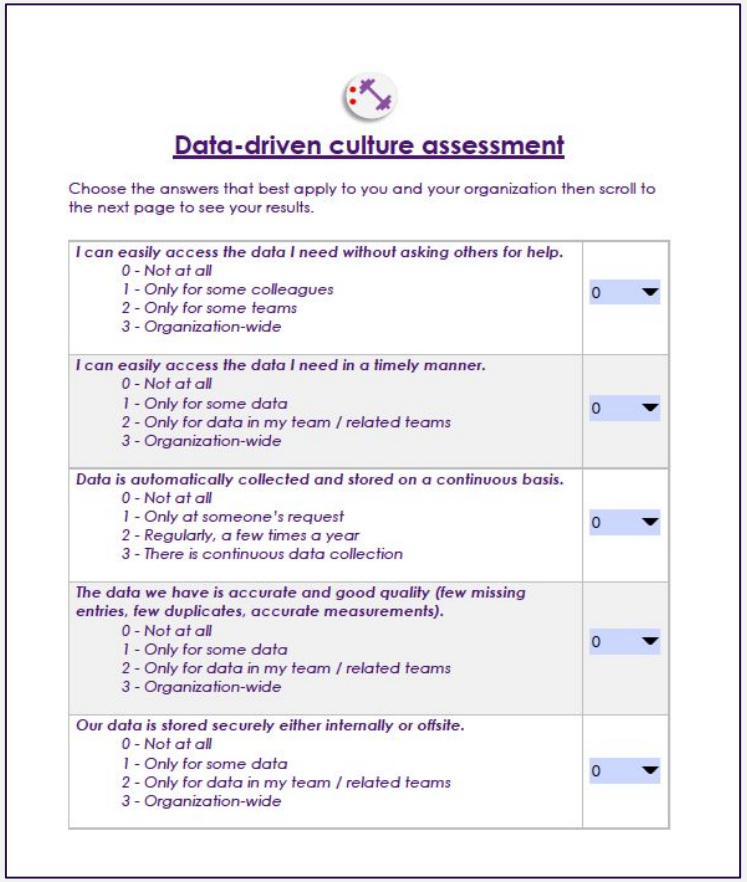
Walmart Corporate, via Flickr

Example: IRM

- Milan needed to replace its slow computers.
- By pulling and analyzing data on computer read/write speeds and hard drive usage, IRM was able to change the purchase order specs.
- Over $50,000 was saved by eliminating unnecessary requirements.

Activity: Are you data driven?

Turn to **page 10** of your participant guide to the **Data-driven culture assessment** to evaluate your team.

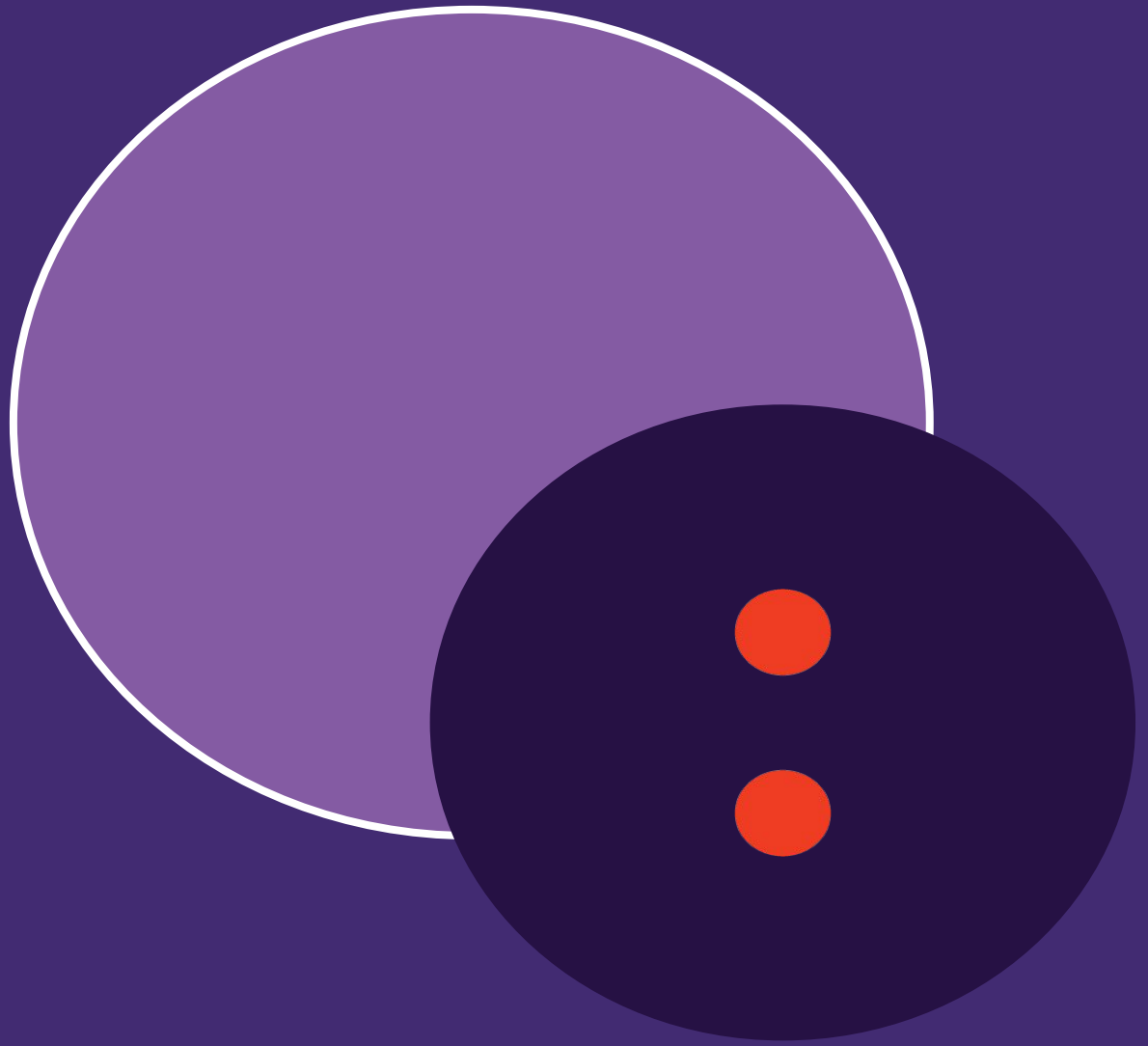


Data-driven culture assessment

Choose the answers that best apply to you and your organization then scroll to the next page to see your results.

I can easily access the data I need without asking others for help. *0 - Not at all* *1 - Only for some colleagues* *2 - Only for some teams* *3 - Organization-wide*	0
I can easily access the data I need in a timely manner. *0 - Not at all* *1 - Only for some data* *2 - Only for data in my team / related teams* *3 - Organization-wide*	0
Data is automatically collected and stored on a continuous basis. *0 - Not at all* *1 - Only at someone's request* *2 - Regularly, a few times a year* *3 - There is continuous data collection*	0
The data we have is accurate and good quality (few missing entries, few duplicates, accurate measurements). *0 - Not at all* *1 - Only for some data* *2 - Only for data in my team / related teams* *3 - Organization-wide*	0
Our data is stored securely either internally or offsite. *0 - Not at all* *1 - Only for some data* *2 - Only for data in my team / related teams* *3 - Organization-wide*	0

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.

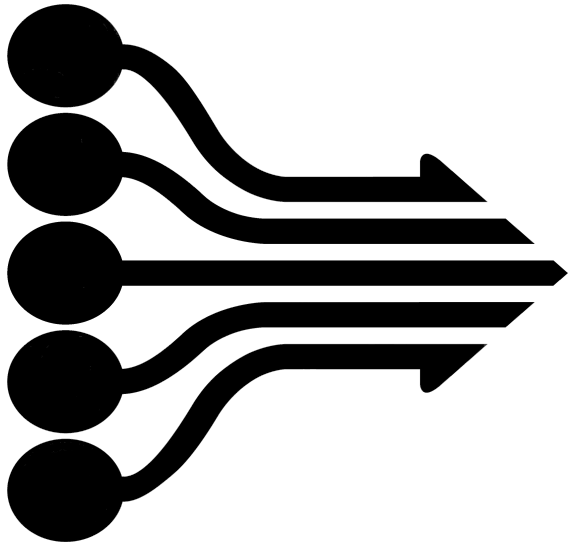
Step 2: Invest in data infrastructure and strategy

- 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

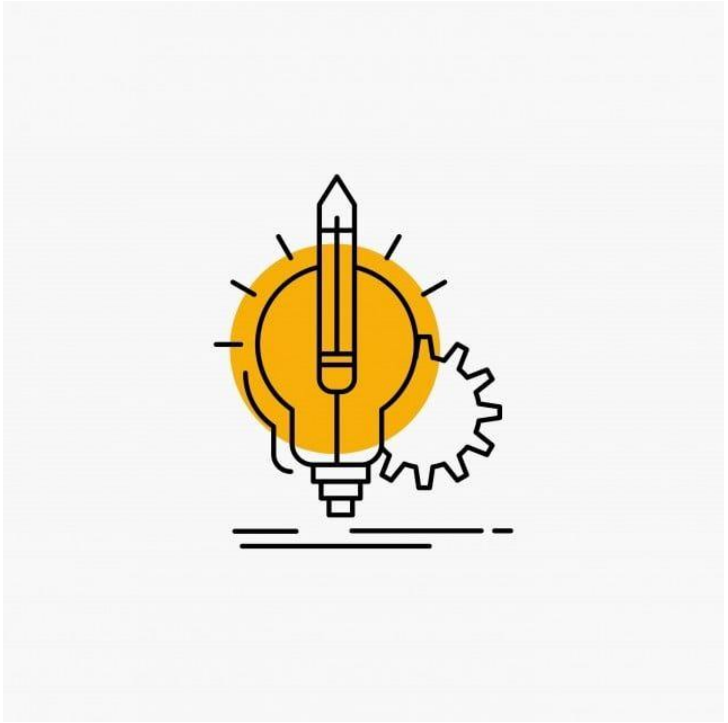
- Data quality is just as crucial as data quantity.
- If you do not have new data in your repository, you might be looking at outdated data and fake reality.
- Keep collecting more relevant, new data.
- Use data mining and software tools to clean and maintain the quality of the data automatically.

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

Chat 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?

Data solutions

Break

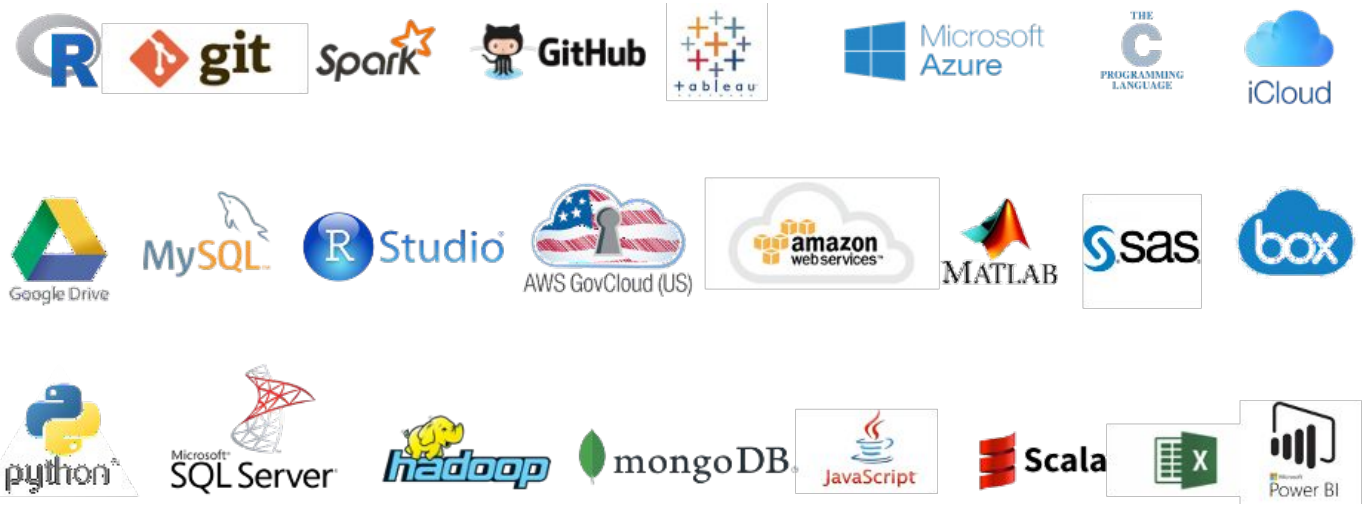
Agenda

Day 2

- Building a data-driven culture
- Data tools
- Data teams
- The data science process
- Putting together a project

- What types of tools do data scientists use to do their work?

Data tools



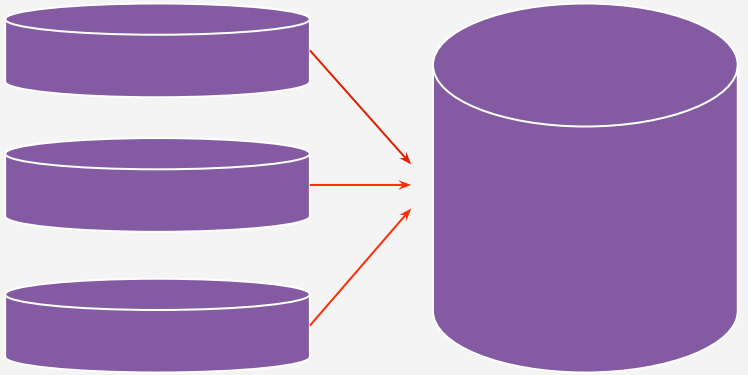
- There's no shortage of tools in the data analytics space.
- There are different tools for different functions, but most overlap in their offerings.

Storage tools

- Databases
  - Relational
  - Non-relational (NoSQL)

- Data warehouses & data lakes

y1	x1	x2	x3
A	F	X	P
B	G	Y	Q
C	H	Z	W



Cleaning tools

- Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

- Example tools: Drake, OpenRefine, DataWrangler, Data Cleaner, Winpure Data Cleaning Tool

Analysis tools

- Analysis tools make it easier to sort through data in order to identify patterns, trends, relationships, correlations, and anomalies that would otherwise be difficult to detect.

- Tools known to be in use at State:
  - Excel
  - R
  - Python
  - SAS
  - WordSmith

Visualization tools

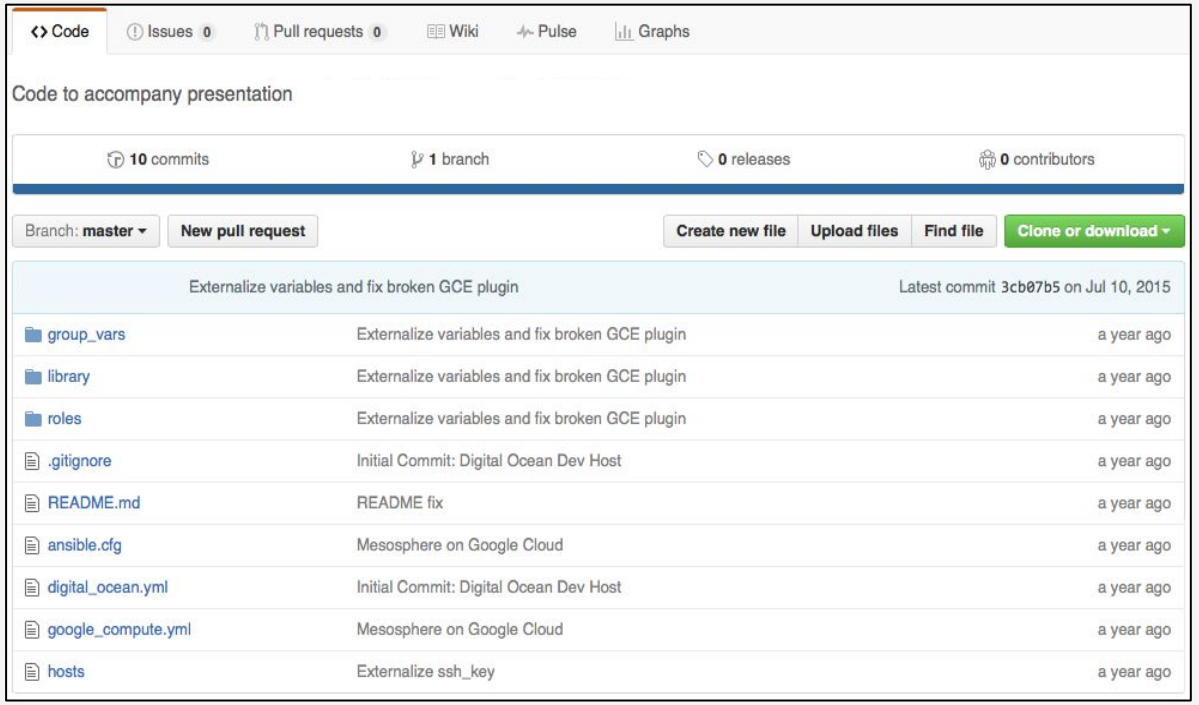
- Visualization gives a visual or graphical representation of data/concepts.

- Tools known to be in use at State:
  - Excel
  - Power BI
  - Tableau
  - R and RStudio
  - Python
  - Power BI
  - MicroStrategy

Collaboration tools

- Collaboration tools offer version control, workflow, bug tracking, task management, etc.

- Example tools: Git, GitHub



Other technologies

- Several other technologies enable and support data analytics, including:
    - Application programming interfaces (APIs). An API is a computing interface that allows two applications to talk to each other. Using them speeds up data acquisition.
    - Graphical processing units (GPUs). A GPU is an electronic circuit specially designed to process graphics such as images and video. Using them can help speed up computation.
    - Cloud computing. Cloud computing offers fast and flexible servers, storage, databases, networking, software, analytics, and intelligence over the Internet.

Questions to guide tool selection

1. What types of technologies are needed for working with data at various stages of the data pipeline?
2. How do the different tools and technologies compare in their functionality, strengths, and weaknesses?
3. Do you have staff who can be trained or know how to use particular tools?
4. Do you have budget constraints you need to be mindful of?
5. Is it on the approved software list?

Chat question

Let's imagine you want to create data visualizations for an upcoming report.

What tool will you use? Why?

Agenda

Day 2

- Building a data-driven culture
- Data tools
- Data teams
- The data science process
- Putting together a project

- Who is on the data team?
- How do data teams fit within an organization?

Data Analyst

- Ensures that collected data is relevant and exhaustive while also interpreting the analytics results
- Main role and responsibilities include:
  - Wrangling the data
  - Managing the data
  - Creating basic analyses and visualizations
- Core skills to include: SQL, R / Python, Tableau / Power BI

Data Scientist

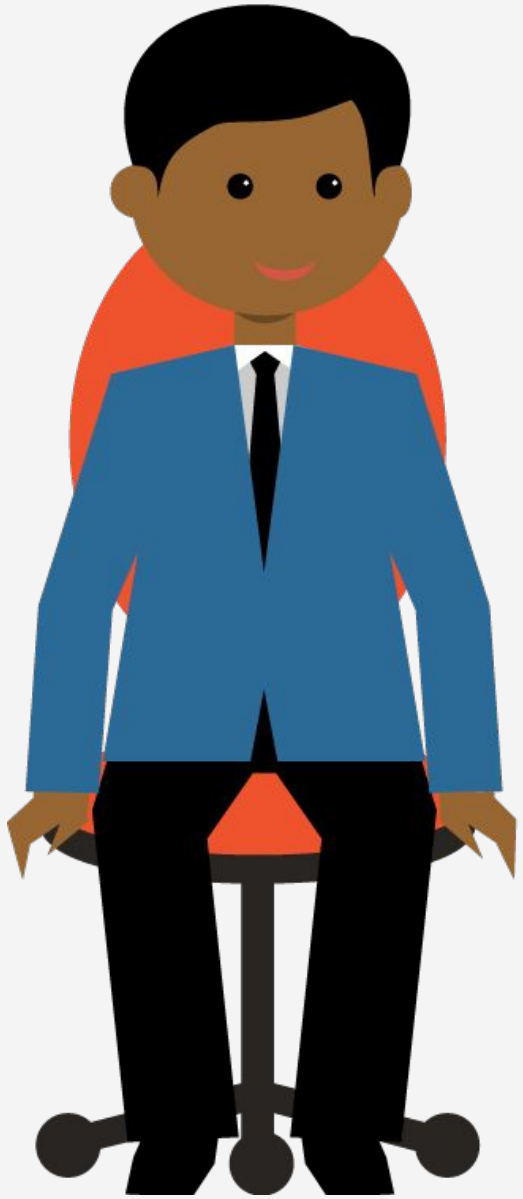
- Builds upon the analysts' data work to develop predictive models and complex algorithms
- Main role and responsibilities include:
  - Asking the right questions from the data
  - Building more complex predictive models
  - Interpreting the results critically and communicating them well
- Core skills to include: R, Python, Spark, Hadoop

Data Engineer

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

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

- Core skills to include: AWS, MongoDB, MySQL, Hadoop, C++, Azure

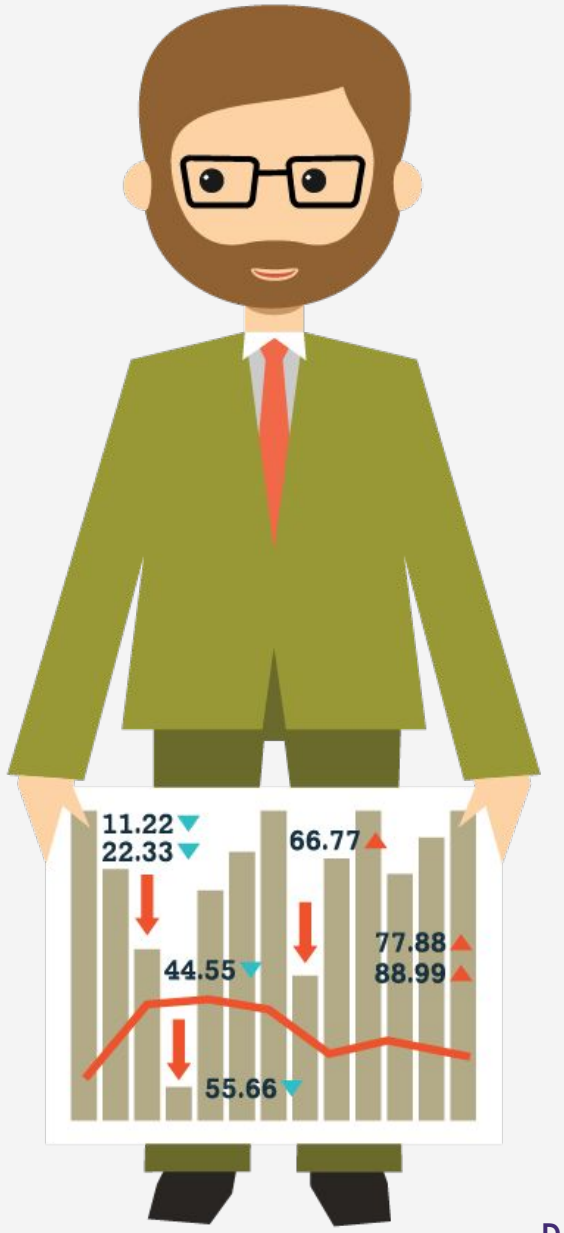


MLOps Engineer

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

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

- Core skills to include: distributed computing principles, networking, database architecture



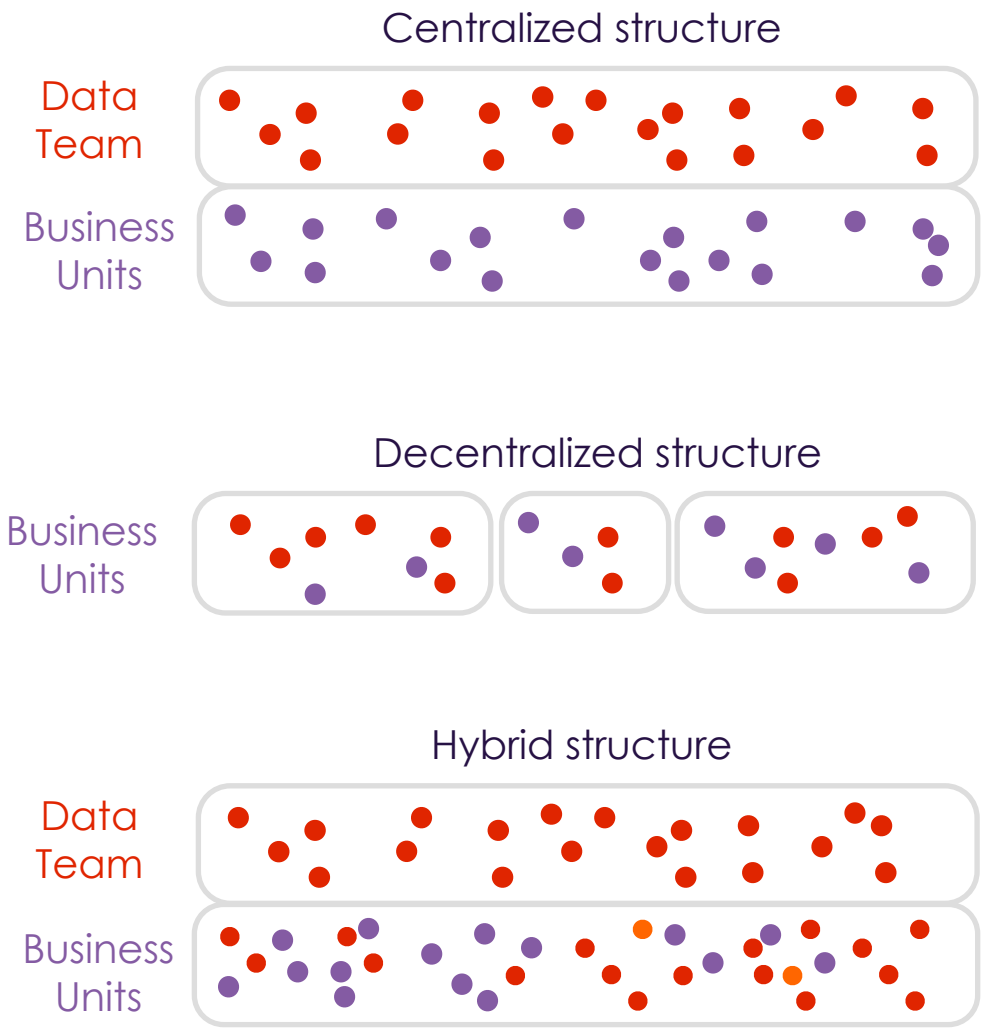
Data Science Manager

- Oversees and directs data science teams and projects and is the bridge between data and non-data people

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

- Core skills to include: management experience, programming skills (R / Python / SQL), strong communication

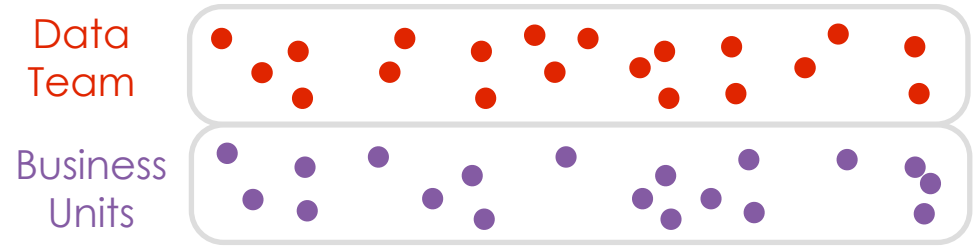
Team structures



- Data analysts
- Business analysts

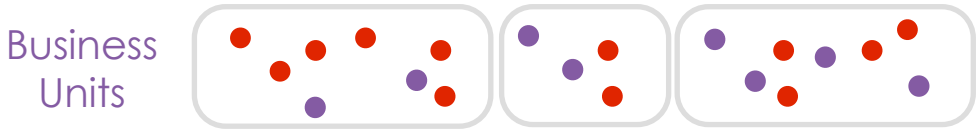
Pros and cons

Centralized structure



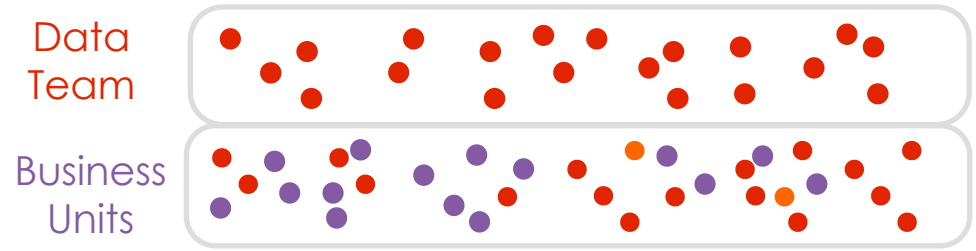
+ easier to standardize team processes
- harder to coordinate projects to meet strategic goals

Decentralized structure



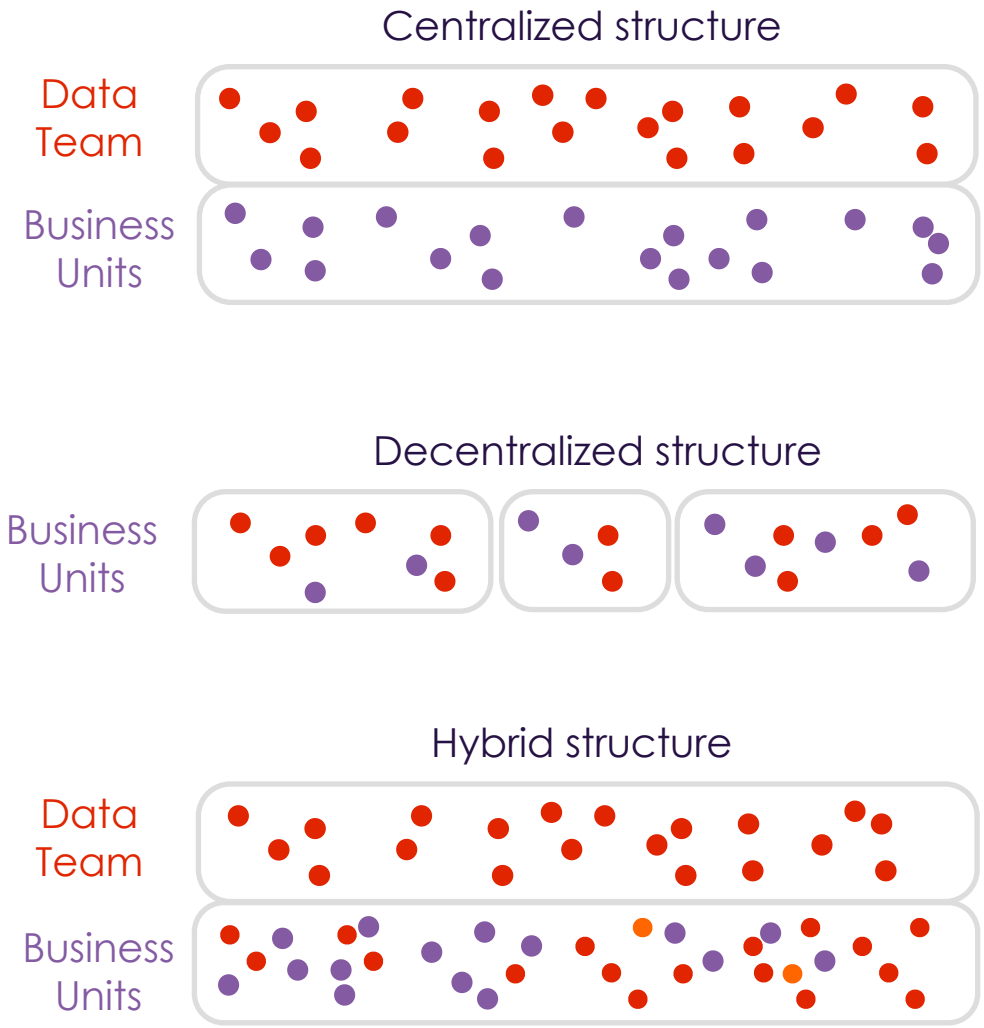
+ easier to coordinate projects to meet strategic goals
- leads to inconsistent & redundant data usage across organization

Hybrid structure



+ easier to standardize team processes
+ easier to coordinate projects to meet strategic goals

Polling question



Which best describes the structure of the data teams in your organization?

- Centralized
- Decentralized
- Hybrid

Another option…

Contracting a team

Strengths

- Flexible cost structure can adapt to changing budgets
- Easy to change staff if people don't work out
- Quickly add staff with new skills

Weaknesses

- Internal know-how is not built up
- Data science does not become an endemic capability
- The organization becomes dependent on forces outside of its control

Hiring a team

Strengths

- Data science becomes an endemic capability—better decision making becomes part of the DNA
- Internal know-how is developed and sustained—the analytics capability has a strong foundation

Weaknesses

- State-of-the-art capabilities may still need to be brought in from the outside ("rented")
- Organizational challenge: data science must remain impartial to internal dynamics

Polling question

What would be the best option for your organization?

- Contracting a team
- Hiring a team

What are the key factor(s) in making that decision?

- Recruitment/ training time
- Cost
- Internal know how
- Flexibility
- Not depending on outside forces

Break

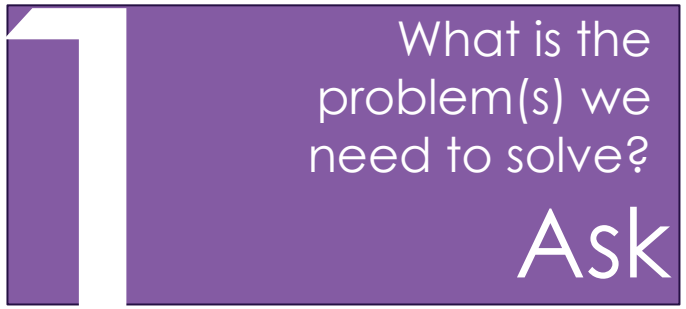
Agenda

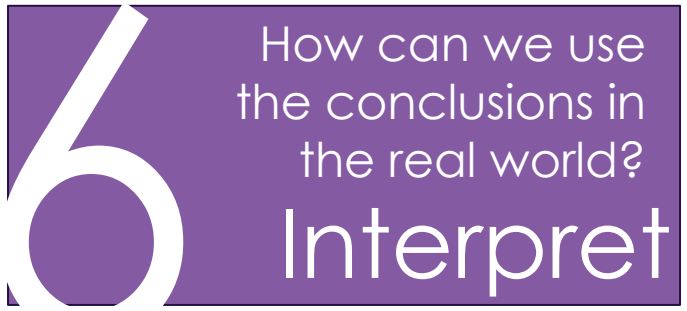
Day 2

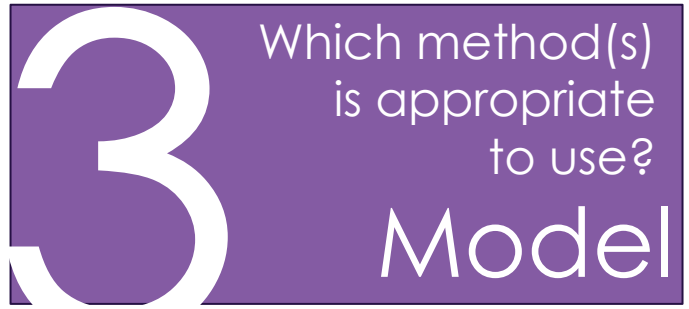
- Building a data-driven culture
- Data tools
- Data teams
- The data science process
- Putting together a project

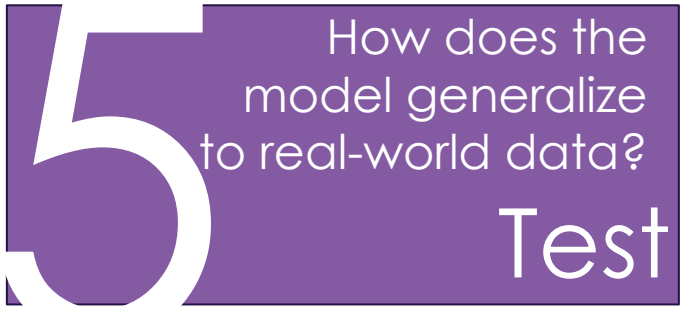
- What are the six stages of the typical data science process?

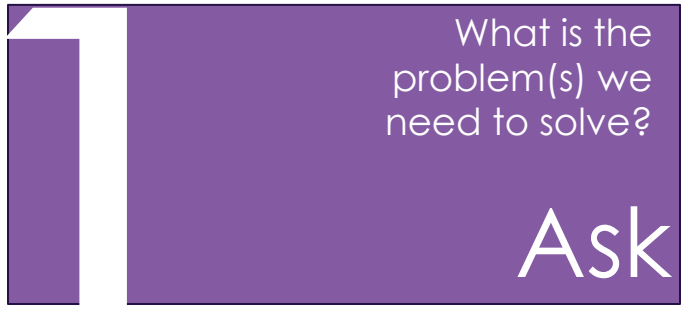
Typical data science process











- 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.

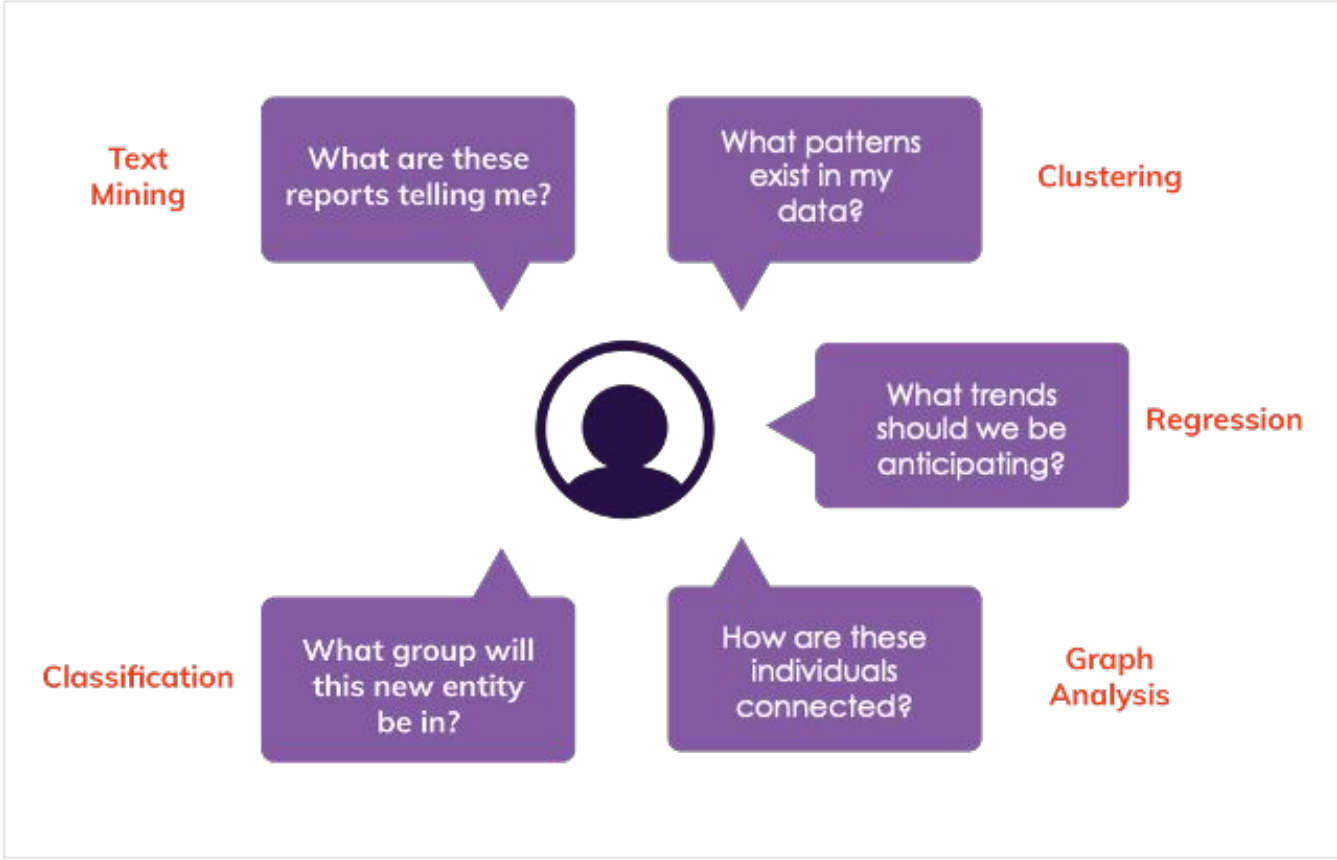
I'm sure the data is good enough as is. → Where can I read about how the data was collected and how the metrics are defined?

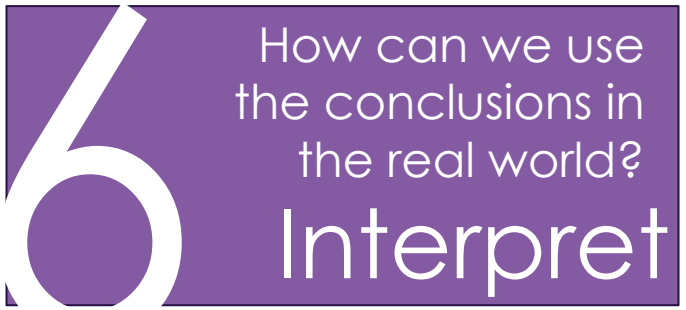
3 Which method(s) is appropriate to use? **Model**

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.

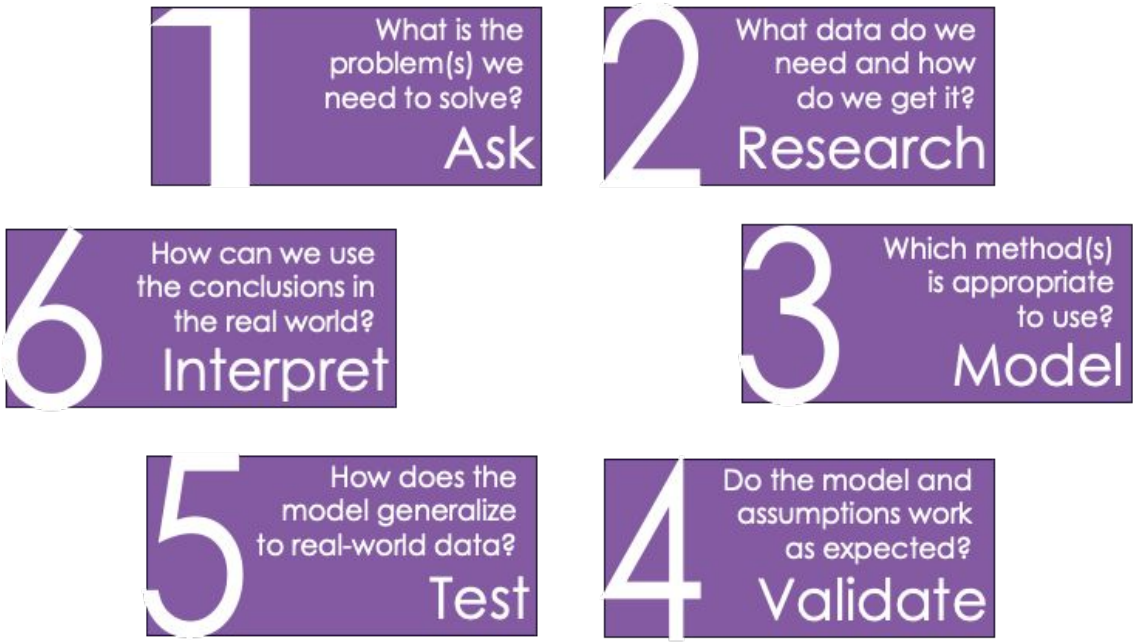
- They present the data and make recommendations based on the data, their domain knowledge, and stakeholder needs.

Example

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

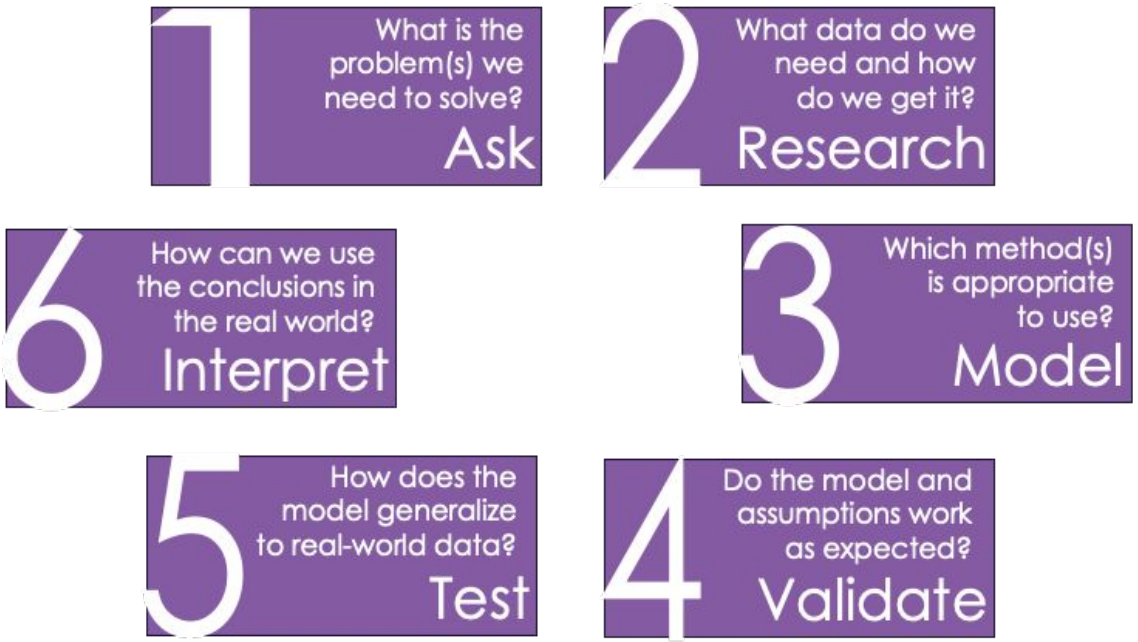
→ How can I best convey the results that matter most to my end users?

Chat question



Which part of the data science process do you think data teams spend the most time on? Why?

Chat question



As a manager, which part of the data science process do you think you should spend the most time on? Why?

Agenda

Day 2

- Building a data-driven culture
- Data tools
- Data teams
- The data science process
- Putting together a project

- How do I identify feasible and impactful data projects?

Planning a data project

- A successful and comprehensive data project is way beyond just programming.
- It involves sophisticated planning and a large amount of communication.
- In this section, we will practice planning an impactful and feasible project.

Planning a data project

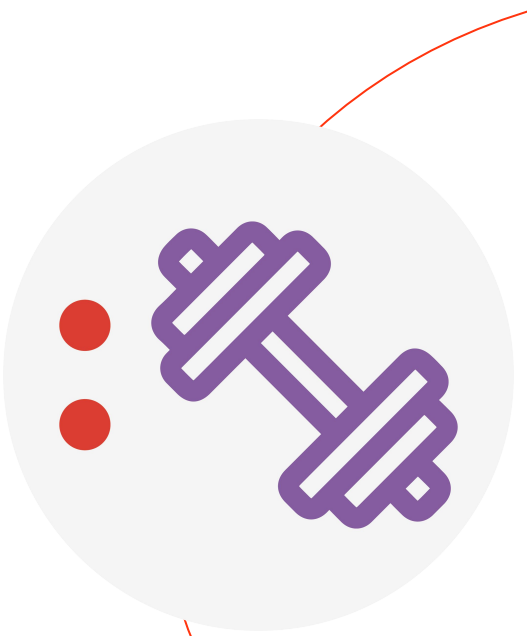
In previous classes, participants have tacked projects such as:

- Using data to prioritize the distribution of COVID vaccines to Department employees, family members, and members of the diplomatic community.
- Improving equity in the post bidding process using data.
- Determining the impact of bilateral engagement (e.g., meetings and trips) on a country's child abduction indicators, according to the data.
- Proving, with data, that counternarcotics funding in Colombia has reduced cocaine consumption.

Activity: brainstorm ideas

- Turn to **page 13** 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.



End of Day 2

Building a data-driven culture
Data tools
Data teams
The data science process
Putting together a project

DATASOCIETY:

DATA LITERACY FOR MANAGERS

Day 3

World's Smartest Home

Agenda

Day 3

- Foundational data science methods
- Advanced data science methods

- What are the basics of machine learning?
- What is clustering and how is it used?
- What is classification and how is it used?
- What is regression and how is it used?

Why learn about these methods?

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

What's an algorithm?

What is machine learning?

- Machine learning 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
  - social-media feeds like Facebook and Twitter
  - voice assistants like Siri and Alexa

- Most is categorized as either supervised or unsupervised.

Supervised learning

- You have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.
- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
- Requires labeled data (i.e., data tagged with one or more labels identifying certain properties, characteristics, or classifications)
- Example: emails are classified as spam/not spam based on how their features compare to the features of emails that a human "Marked as Spam."

Unsupervised learning

- You only have input data (x) and no corresponding output variables. The goal is to model the underlying structure or distribution in the data in order to learn more about the data.

- In other words, the machine looks for whatever patterns it can find.

- Example: for marketing purposes, finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying record

Polling question

The goal of this type of machine learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Do you think this statement describes supervised machine learning or unsupervised machine learning?

Polling question

The Stanford Dogs Dataset contains 20,580 images. Each image is categorized into 1 of 120 different dog breed categories.

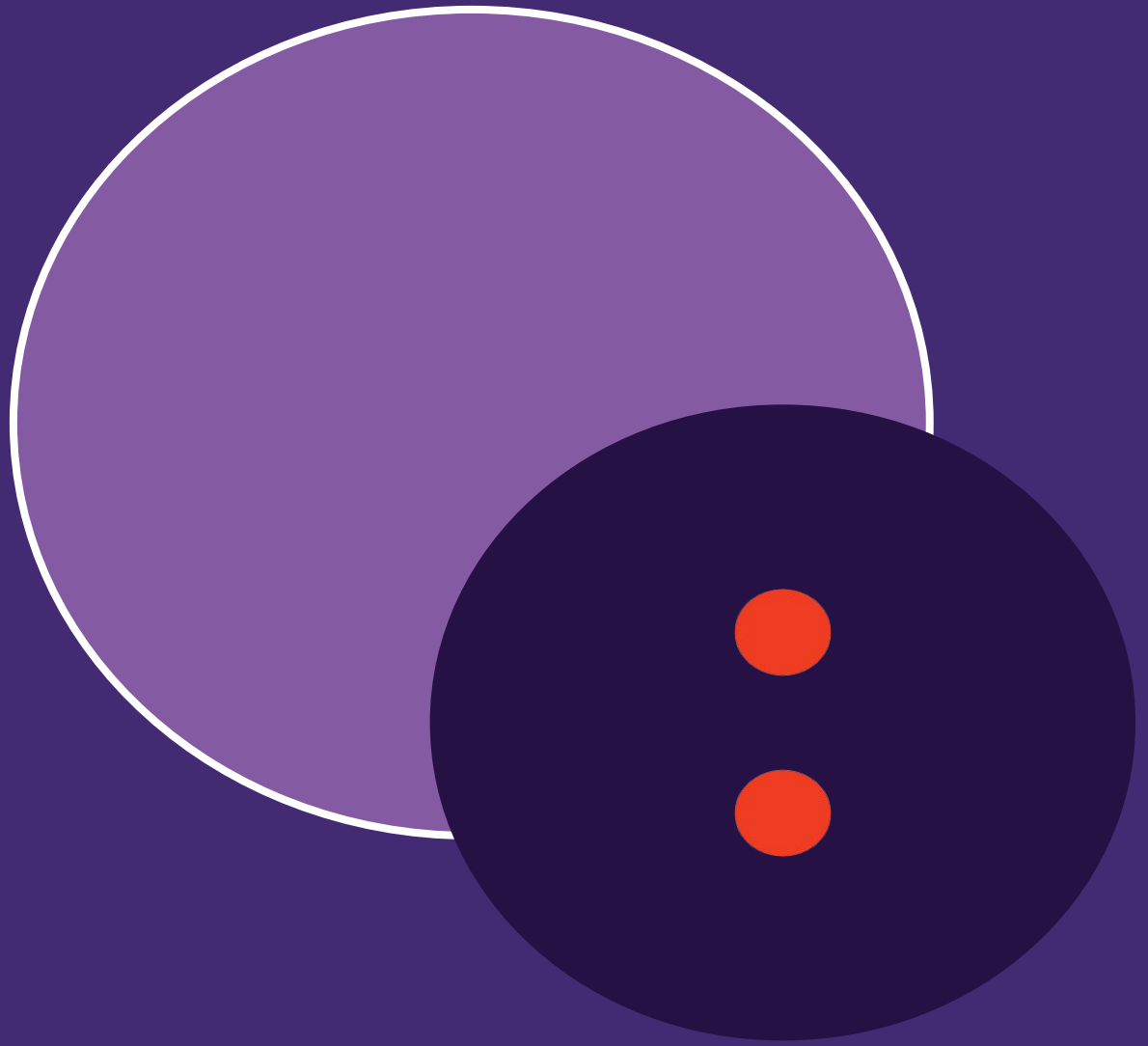
Based on the information provided, is this dataset suitable for use with supervised machine learning techniques?

Before we go further…

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

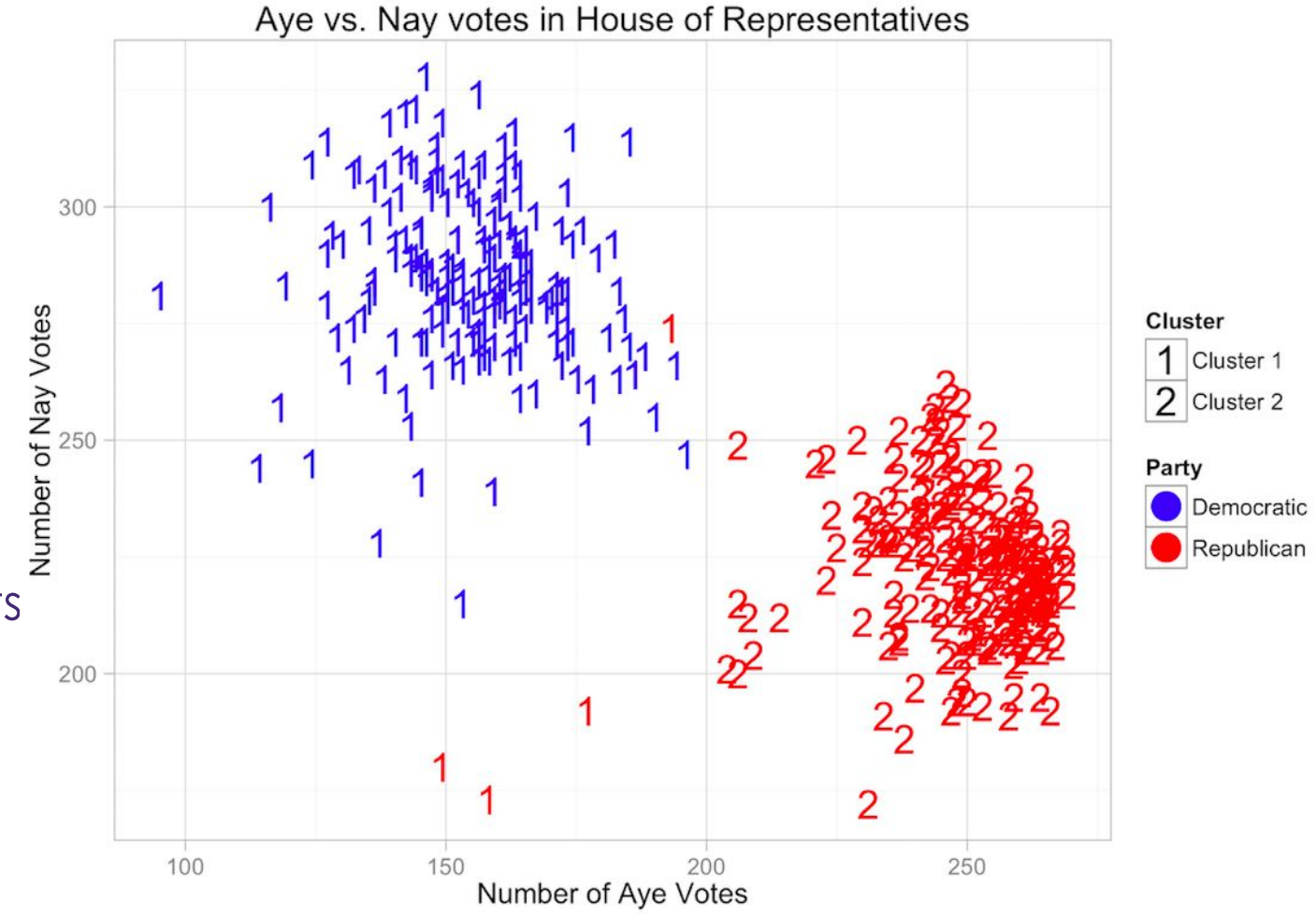
- The two big components that drive the decision for which method to use are: the question you're asking, and the data you have.

Clustering



Clustering

- Clustering is a type of unsupervised machine learning.
- You find similarities between data points and create groups (clusters) based on those similarities.
- It tries to find whether there is a relationship between the data points when the classes are unknown.



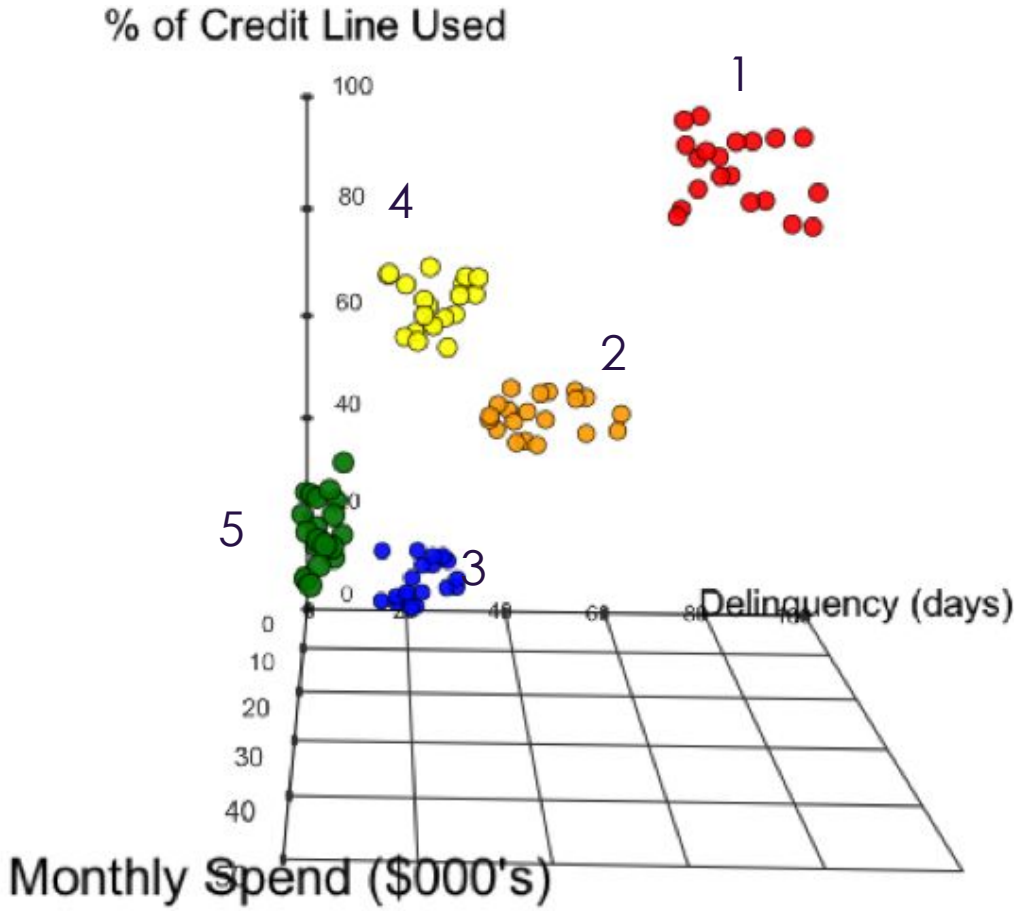
How can you use clustering?

- 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.

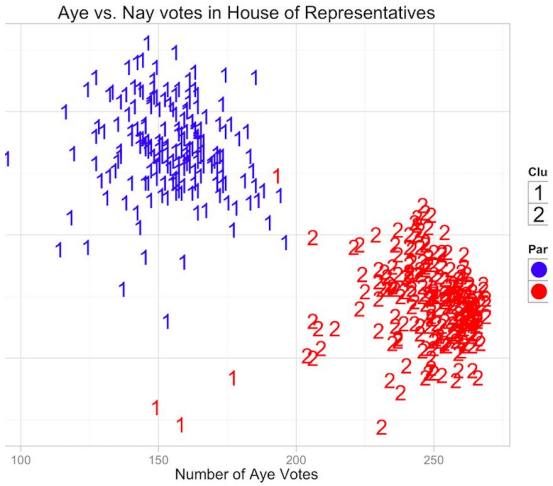
Example: credit line optimization

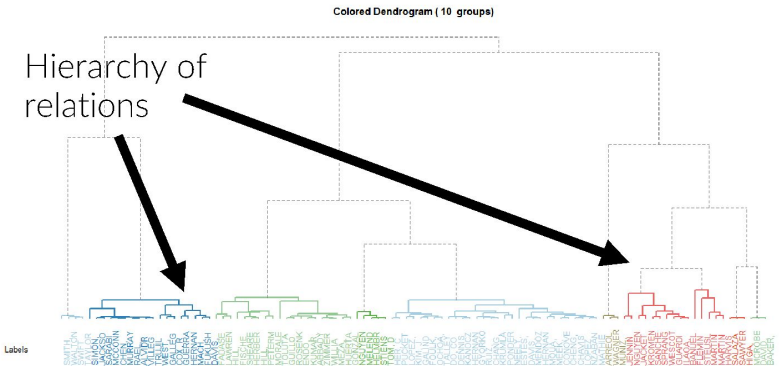
- GE Capital created a model to predict customer behavior and offer tailored products.
- The clusters were defined using existing GE Capital data—based on days delinquent, monthly spend, and percent of credit line used.
- Led to more targeted marketing and specific offers to those groups.

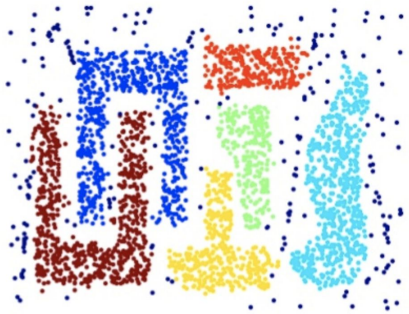


Types of clustering

- Centroid - iterative clustering algorithms where the proximity of data points is translated into similarity
- Hierarchical - assumes that the closer the data points are to each other, the more similar they are
- Density-based - searches for areas of varied density of data points in the dataset and clusters based on the density



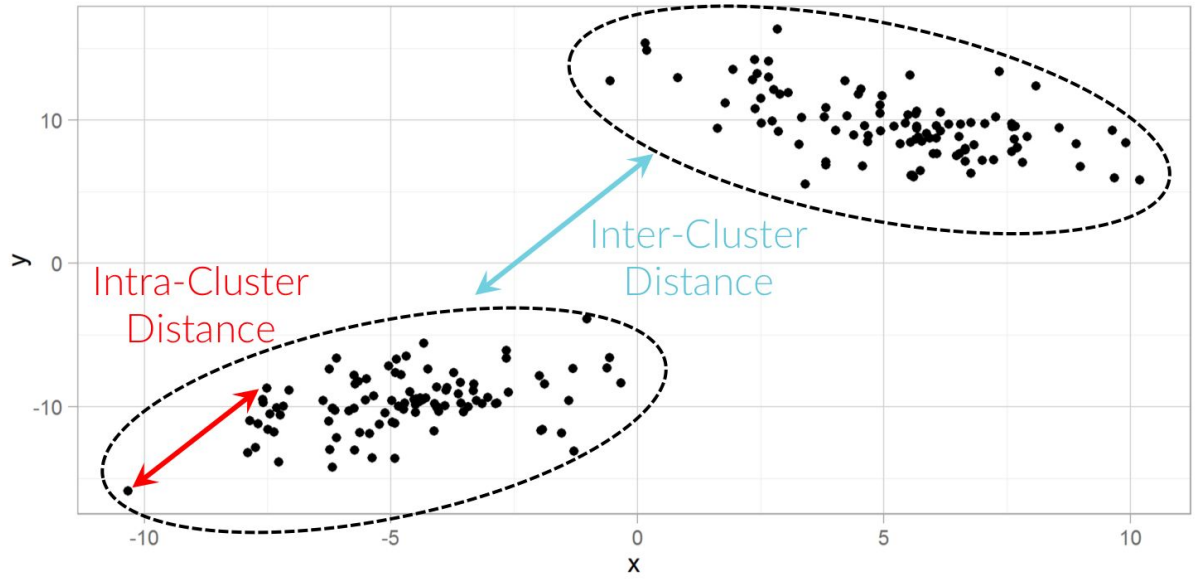




Evaluating the accuracy of the model

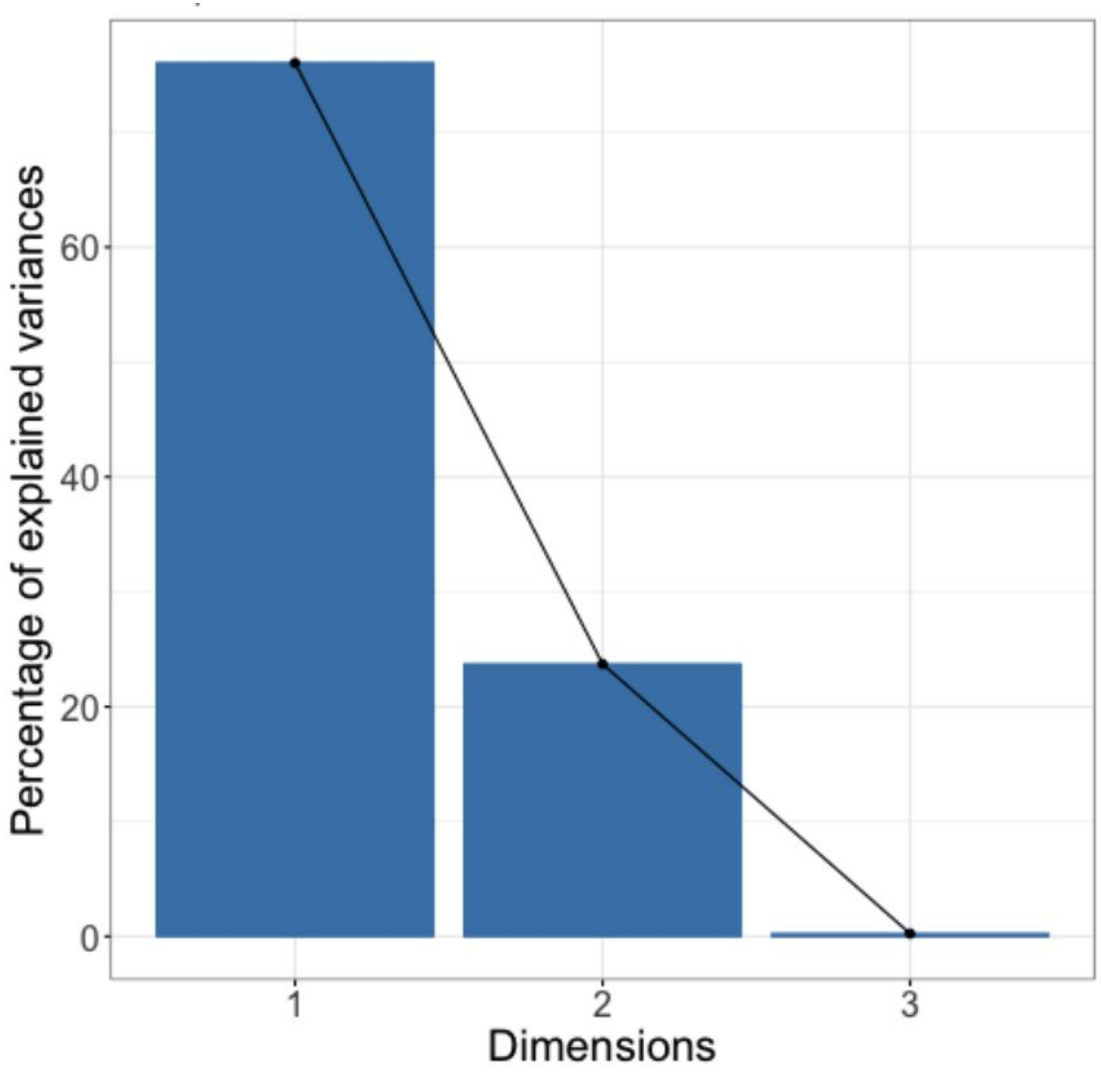
- Goal of clustering is to maximize the separation between clusters and minimize the distance within clusters
- The ratio of inter-cluster variance to total variance can help you assess the performance of algorithms, although this is dependent on the model you use

$$\frac{\text{Variation explained by clusters}}{\text{inter-cluster variance}}{\text{total variance}}$$



Evaluating the accuracy of the model

- A screeplot identifies the contribution of each variable on the explained variance of the model.
- Good for identifying important components of a model



Questions managers should ask

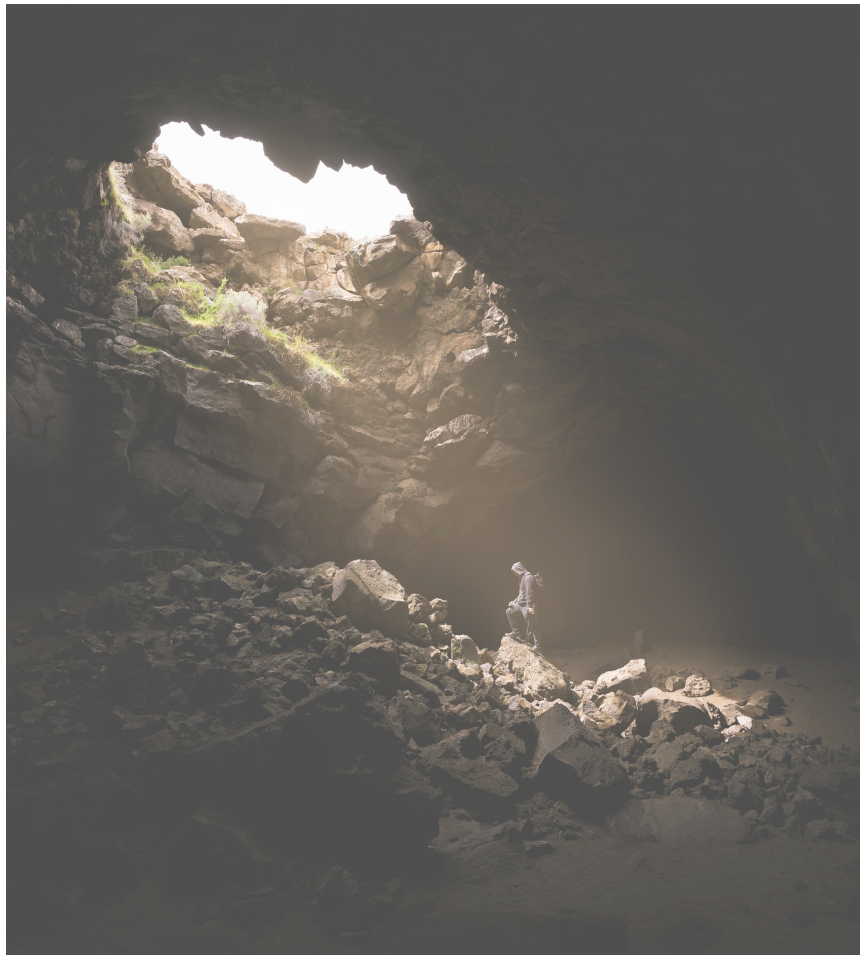
1. How was the distance measure identified?
2. Did you scale the data appropriately?
3. How many clusters do you expect or want? Why?
4. Does your algorithm scale to the size of the data?
5. What can we learn from the groups that the algorithm identified?

Common pitfalls with clustering

- Clustering algorithms don't scale well to large datasets
  - "Curse of dimensionality" – as the dimensions increase, the data points become sparse and increases distance and similarity between points

- 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!

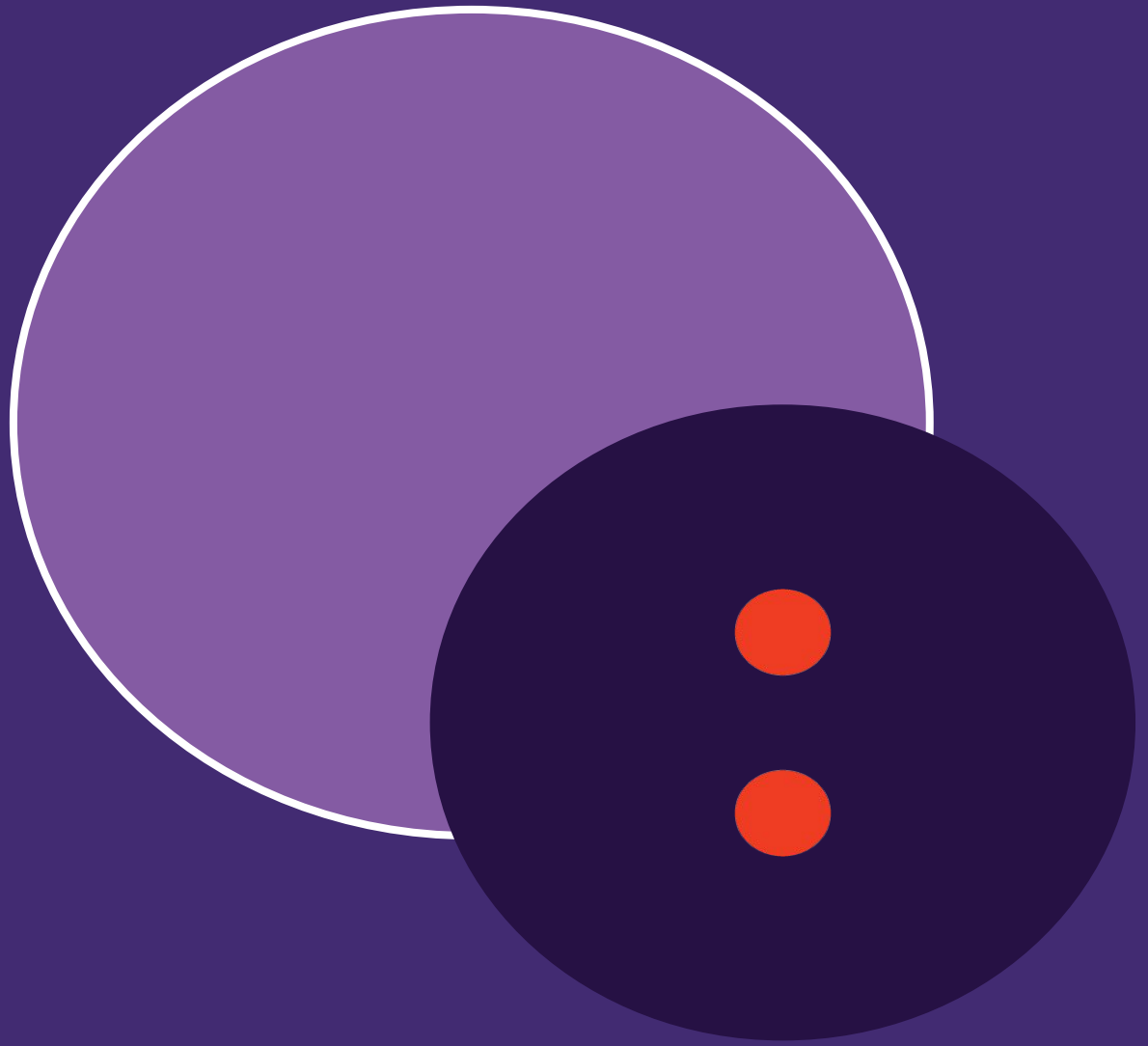


Recap: when should you use clustering?

- Use clustering when:

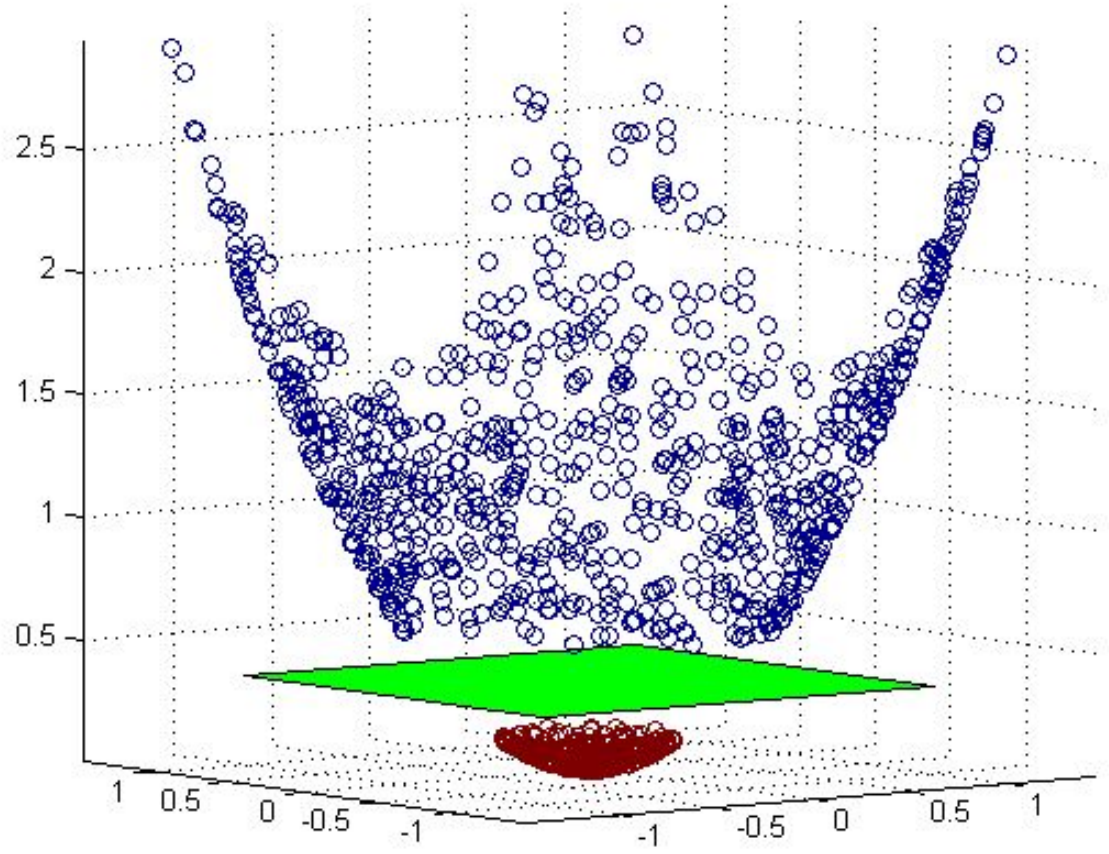
1. You have an unlabeled dataset
2. The dataset has multiple attributes
3. You need to identify patterns in your data
4. You need to find groups in your data

Classification



Classification

- Classification is a type of supervised machine learning.
- It is the process of assigning new data points to known classes.
- The assignment is done based on the similarity of new data points to existing data points with known class assignment (category or behavior pattern).



How can you use classification?

Classification answers the questions:

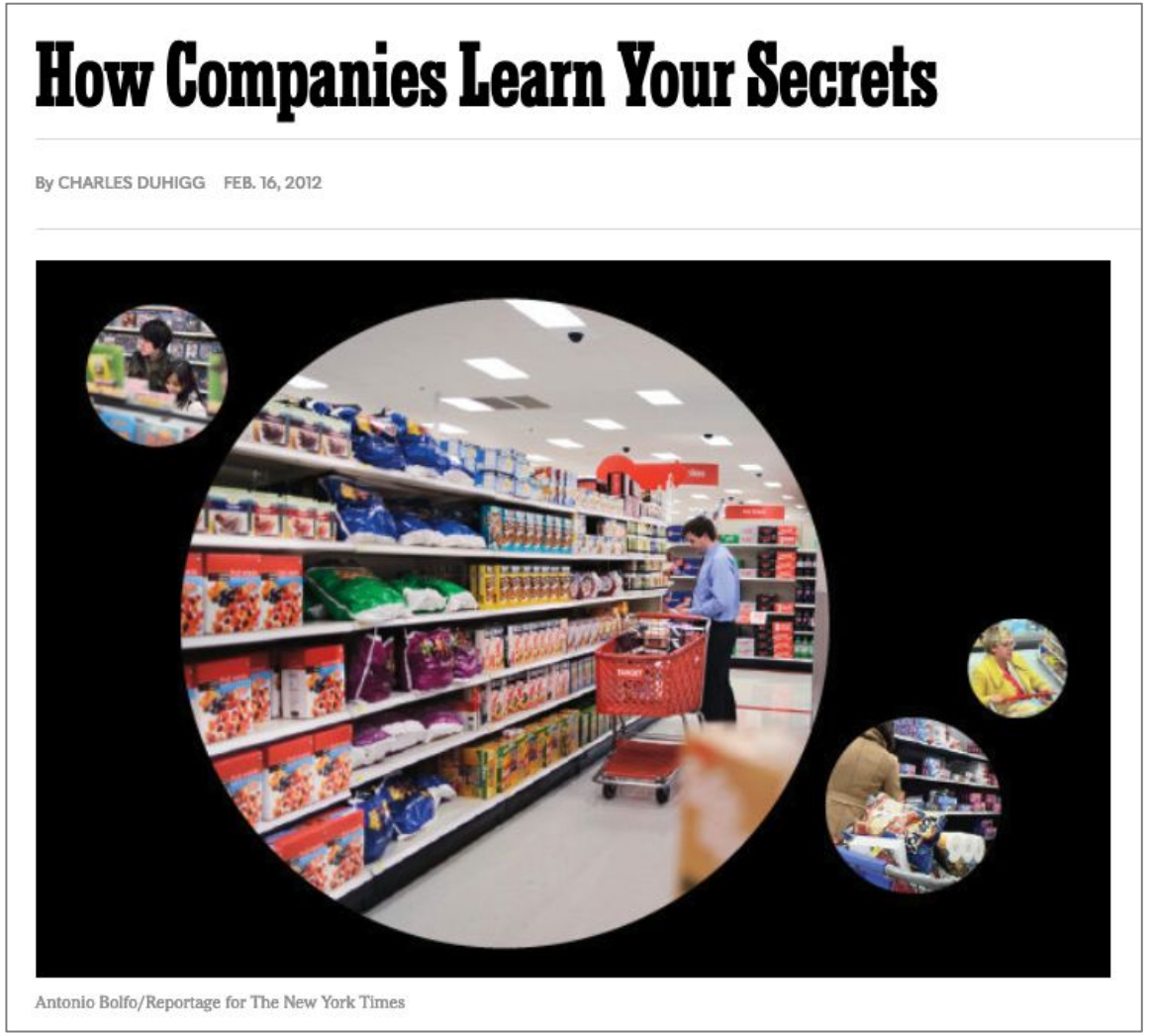
1. Which is the probability of an object / person being in a particular group?
2. What category is this person / object in?
3. What is this person / object most similar to?

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

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

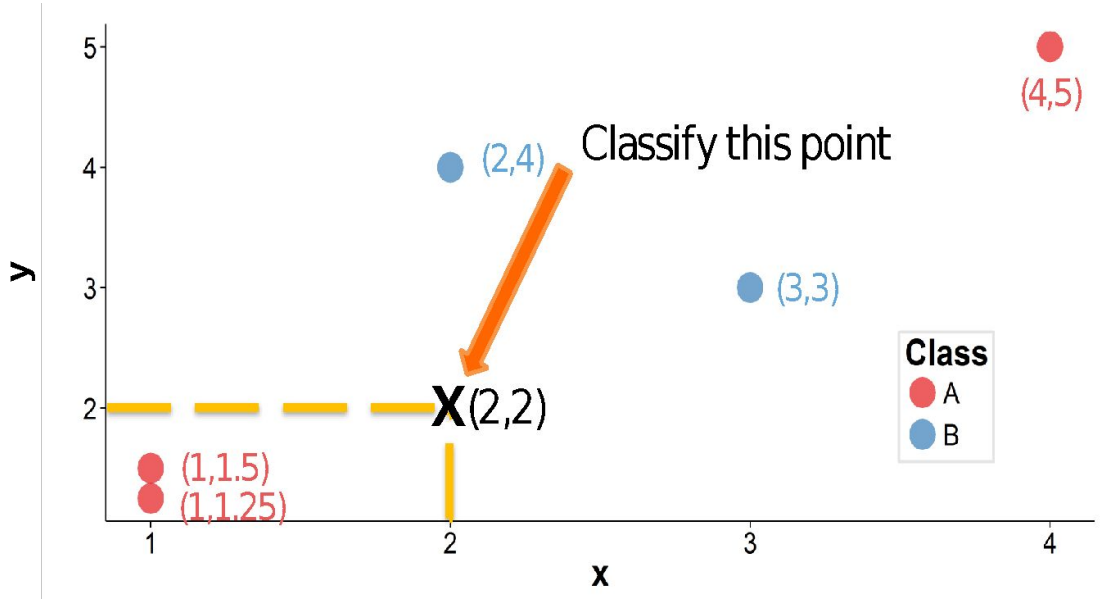
http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0

Chat question

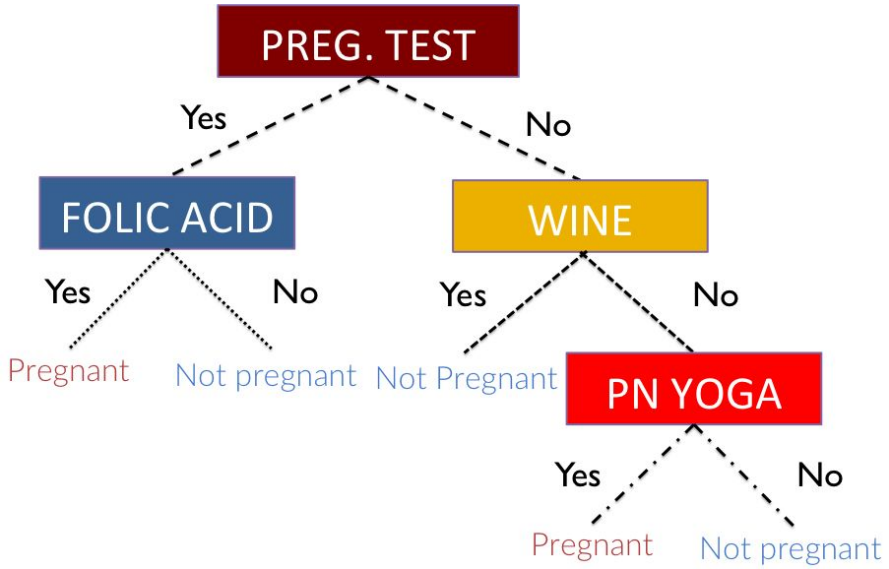
Time out! What ethical implications might Target's pregnancy predications have raised?

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

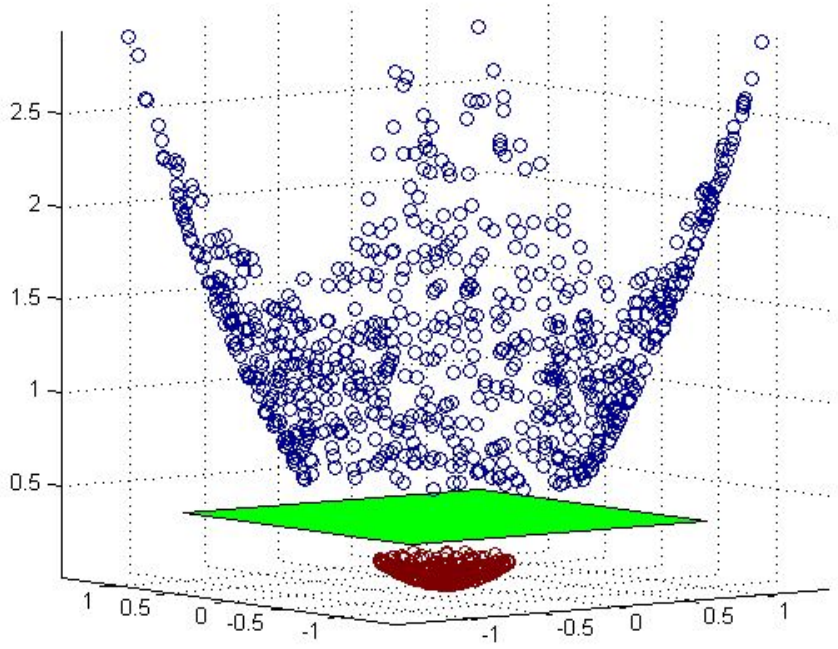


- Decision trees – uses a tree-like graph or model of decisions and their possible consequences to classify data



Common classifiers

- Support vector machines – separates data points by class using an optimal hyperplane



- Logistic regression – determines the probability of a data point to be part of a certain class or not

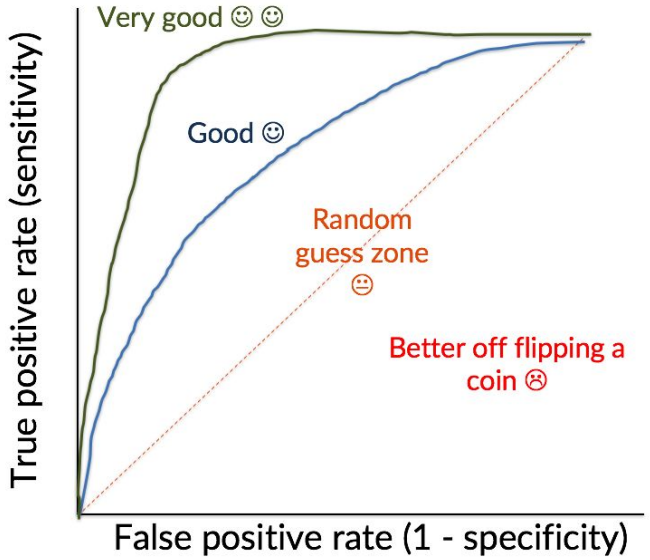
Evaluating accuracy of a model

- In order to determine the accuracy of the model, you need to split your data into a training set and a test set.
- Then, compare the outcomes that the model produced to the actual outcomes to determine how accurate your model is, and how well it generalizes to new data.
- This is called a confusion matrix.

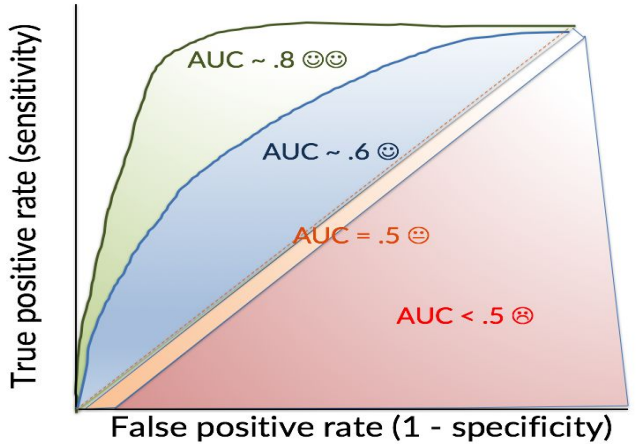
	Y1	Y2	Predicted totals
Predicted Y1	True positive (TP)	False positive (FP)	Total predicted positive
Predicted Y2	False negative (FN)	True negative (TN)	Total predicted negative
Actual totals	Total positives	Total negatives	Total

Accuracy, cont'd.

- Next, you can plot the ROC (receiver operator characteristic), which is the true positive rate against the false positive rate at different thresholds.



- Another metric to plot is called the AUC (area under curve), which compares classification models to measure predictive accuracy.
  The AUC should be above .5 to say the model is better than a random guess.



Questions managers should ask

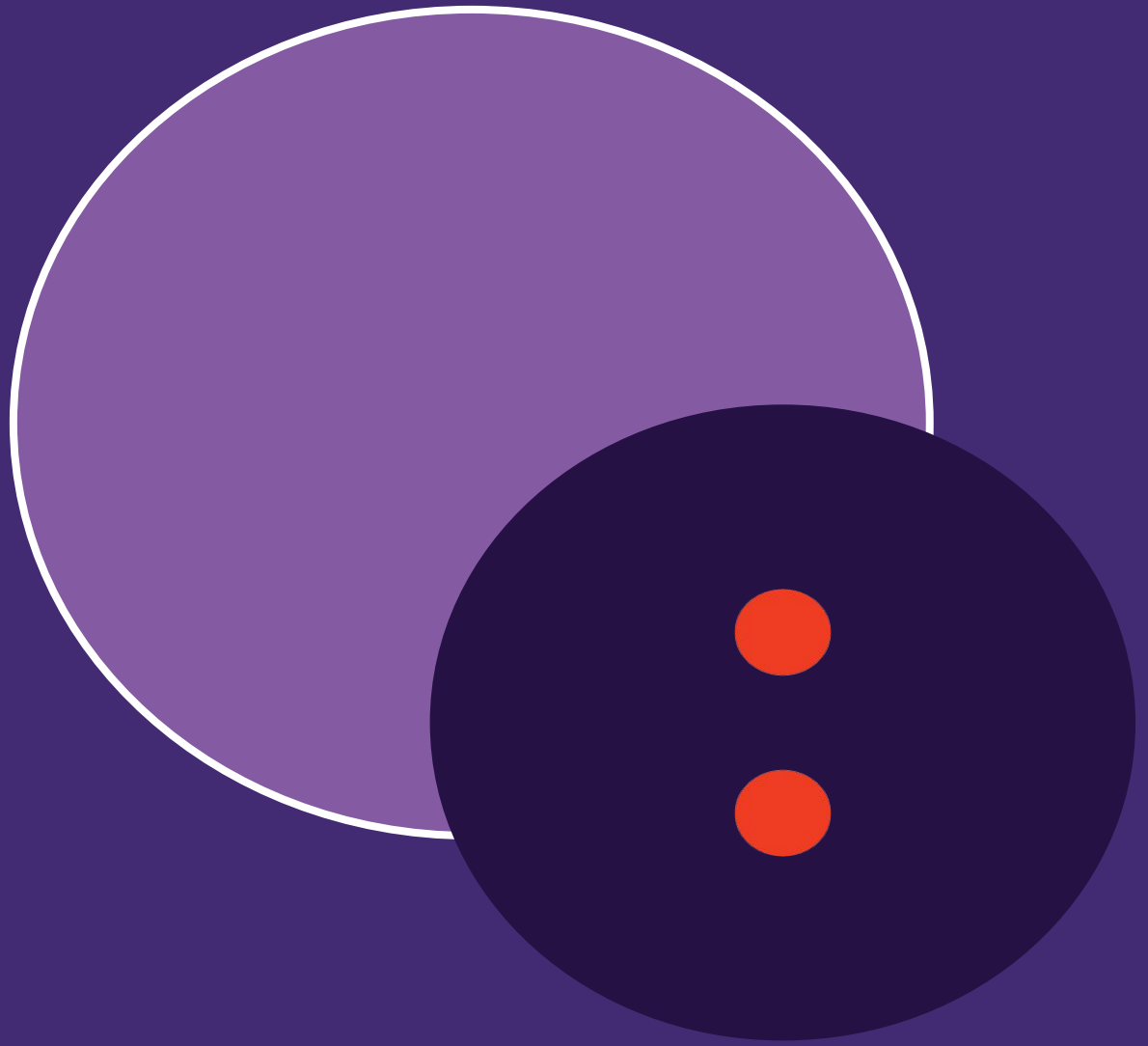
1. How was the distance measure identified?
2. Did you scale the data appropriately?
3. How did you split the test and training data?
4. What thresholds did you use for AUC and ROC?

Recap: when should you use classification?

- Use classification when:

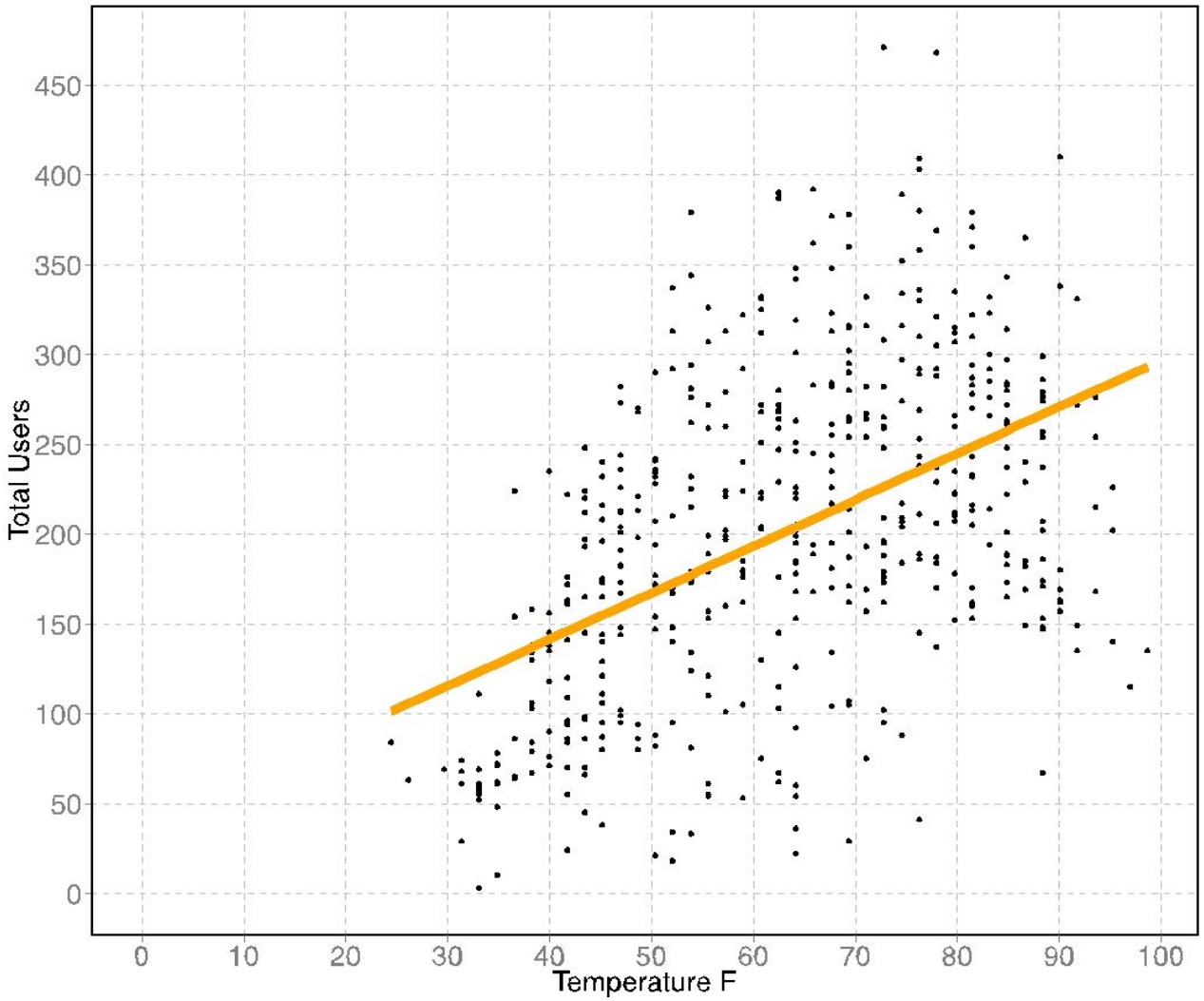
1. You have a labeled dataset 
2. You want to predict group assignments 
3. You want to predict behaviors / events 
4. You want to identify important attributes 

Regression



Regression

- Regression is a type of supervised machine learning.
- It predicts the value of a variable based on the value of another variable or several variables.
- It's used to examine and calculate the relationship between a variable of interest (dependent variable) and one or more explanatory variables (predictors or independent variables).

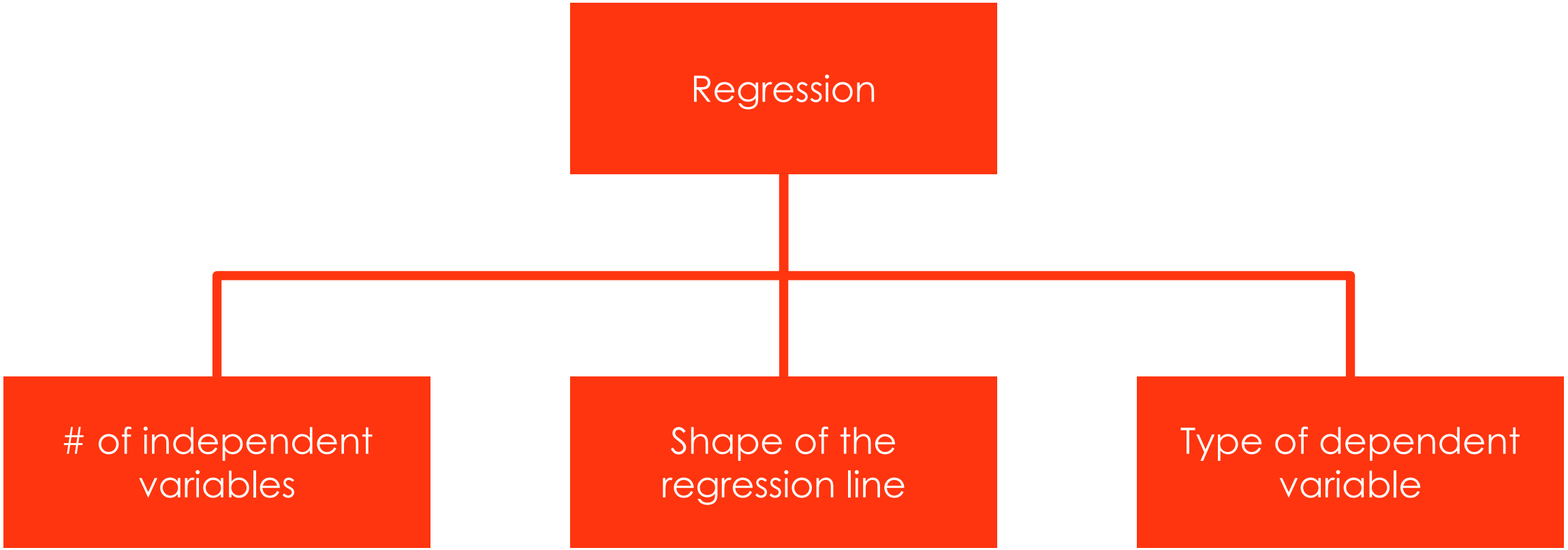


How can you use regression?

- Regression answers the 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?

- Domain knowledge is key!
  - We can predict political instability in countries
  - We can predict how tourism season affects a country's economy

Types of regression techniques



Use case: predicting city movements

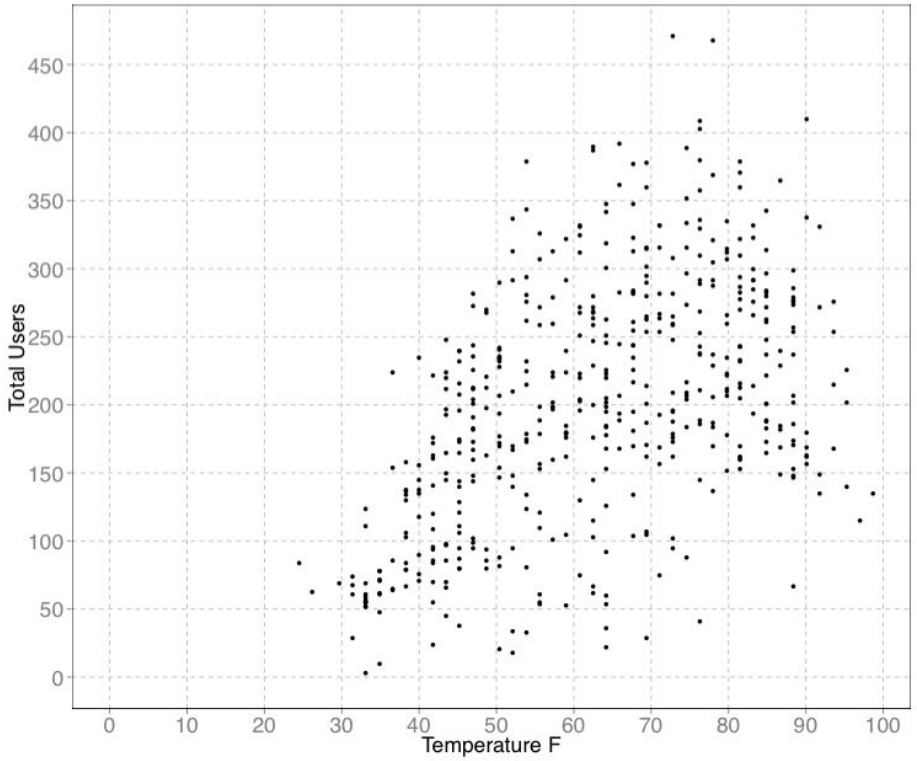
- There are over 500 bike-sharing programs around the world with over 500,000 bikes.
- Automated systems track numerous data points providing a treasure trove of data about the mobility of residents.
- Data can be used to forecast the number of bikes required and adjust pricing based on demand.

Chat question

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

Simple linear regression

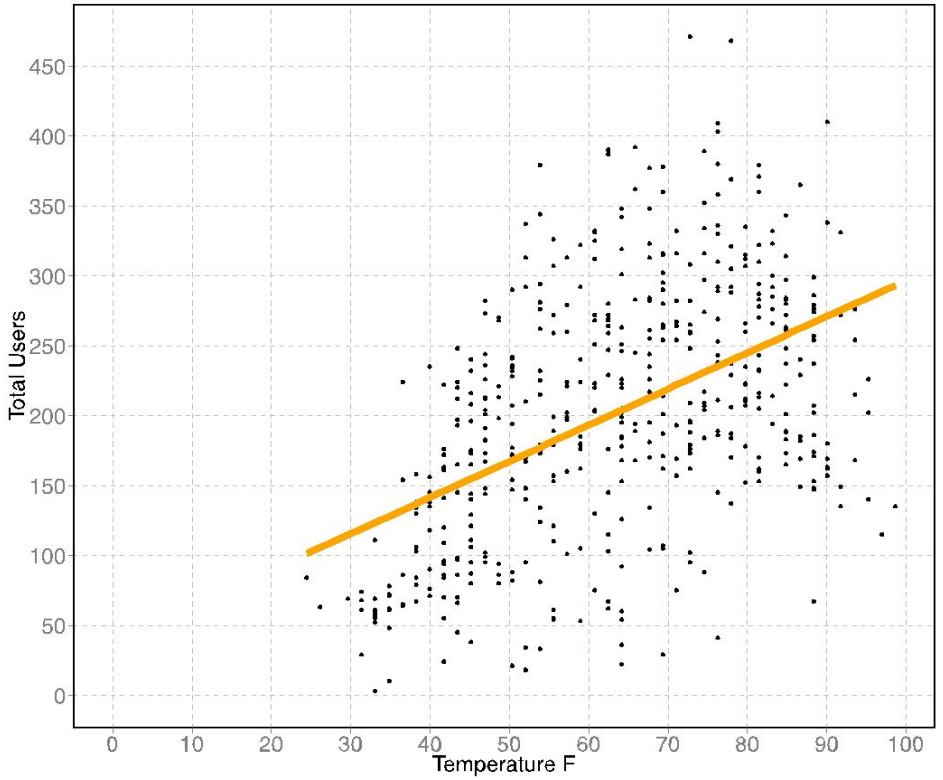
1. Gather data on variables in question
2. Plot the data
3. Draw the line to best fit the data



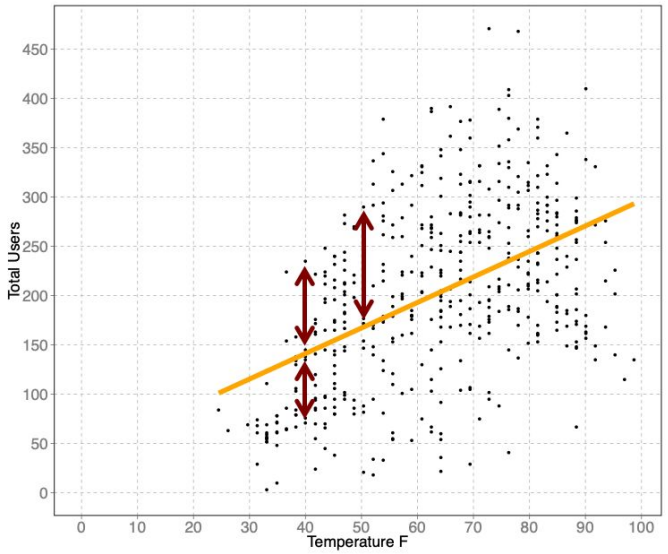
$y = mx + b$

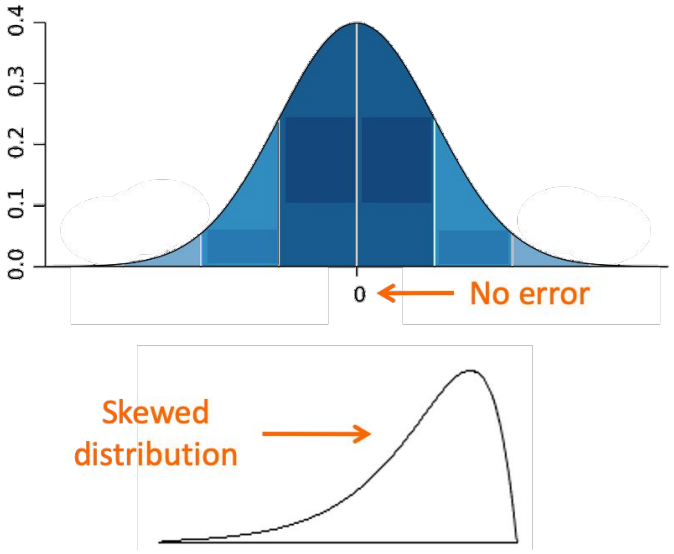
Number of bike users
=
2.6 * (Temperature)
+ 37.6

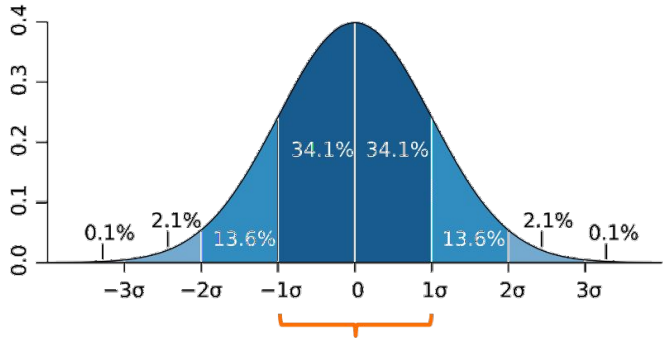
4. Evaluate model performance
   - Measure error
   - Deal with outliers
   -



Measure error







- 68.2% of errors are within 1σ away from the average or best fit line
- 95.4% of errors are within 2σ
- 99.6% of errors are within 3σ

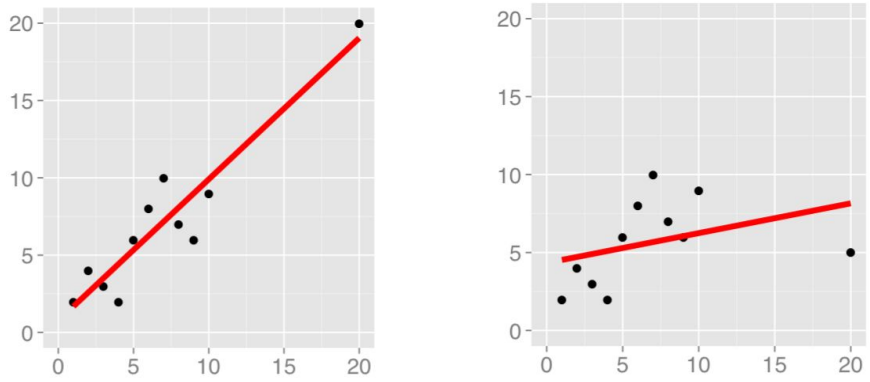
Variance. How widely dispersed is actual data from the expected data?

Randomness. Are the errors random or is there bias in the model?

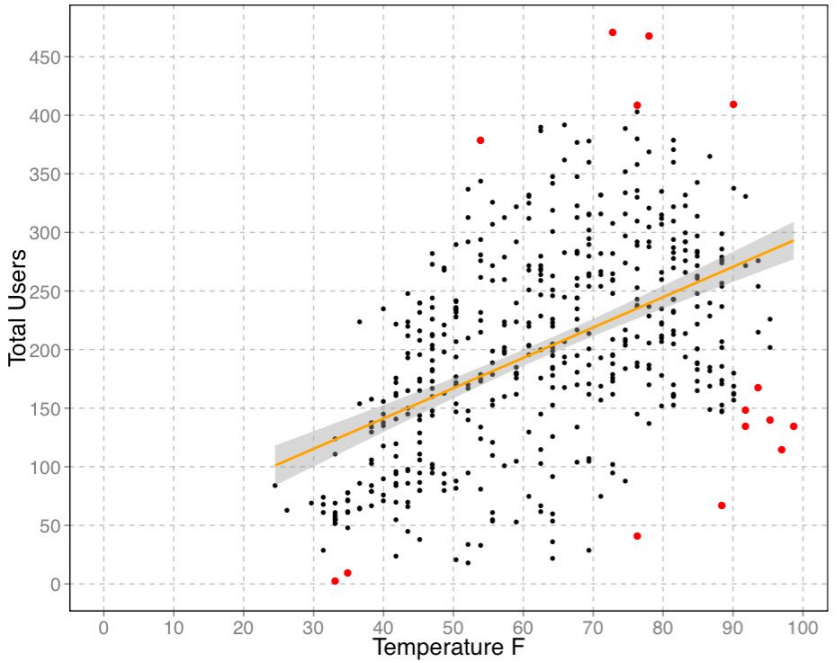
Standard deviation/Certainty. What proportion of data points fall within a given range? How likely is a value to be in that range?

Deal with outliers

- Just one outlier can have a very negative impact on a linear regression if it is not identified and handled properly.



- Methods such as scatterplots, box-and-whisker plots, and Cook's distance can be used to identify outliers.



Determine accuracy

- Look at:
  - Covariance: measures how changes in one variable effects another variable
  - Correlation: identifies the strength of the relationship between the variables
  - p-values: probability that pattern exists through random chance, and not a relationship between the variables

- R^2 determines the accuracy of a regression model. It's the proportion of variance in the outcome variable that's accounted for by regression
  - e.g., "about 40% of the variance in the number of bike users is explained by the temperature"

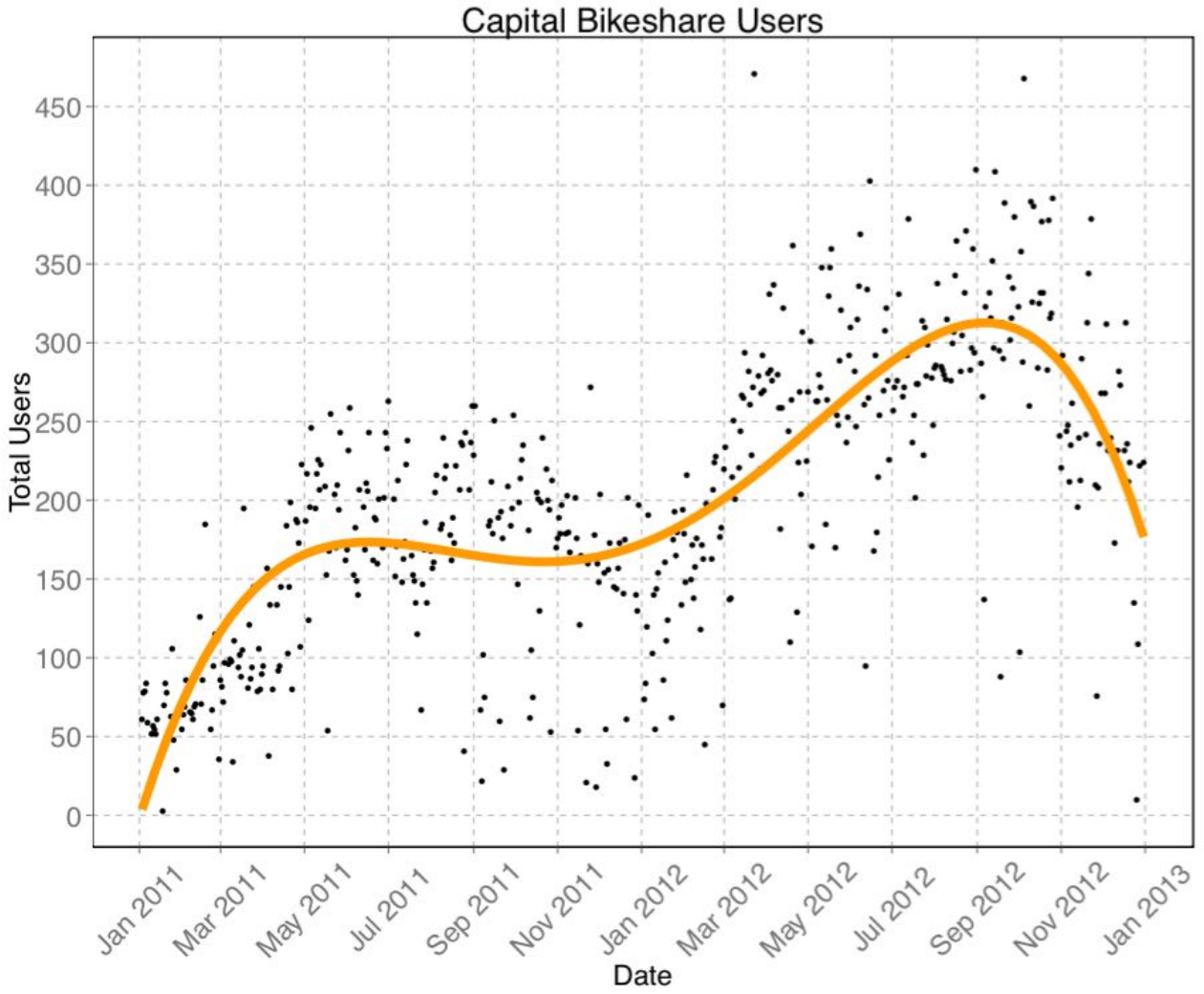
Multiple linear regression

- Has more than one independent variable
  - e.g., How do several variables (temperature, humidity, day of the week, time of day) affect demand for bikes?

- Added concerns:
  - Multicollinearity: when 2 or more independent variables are strongly correlated to one another you may be effectively double counting an effect
  - Autocorrelation: when the correlation between the values of the same variables is based on related objects
  - Heteroskedasticity: when the variability of a variable is unequal across the range of values of a second variable that predicts it

Other types of regression

- Nonlinear Regression
- Binary Logistic Regression
- Ordinal Logistic Regression
- Nominal Logistic Regression
- Ridge Regression
- Lasso Regression
- Partial Least Squares Regression
- Polynomial Regression
- Logistic Regression
- Quantile Regression
- Elastic Net Regression
- Principal Components Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Ecologic Regression
- Bayesian Regression
- Jackknife Regression



Questions managers should ask

1. How well do we understand the underlying data distribution?
2. Did you identify any outliers? Were they significant? Did you remove them?
3. Did you test the variables for multicollinearity so as not to double-count their effects?
4. What was the R^2 metric?

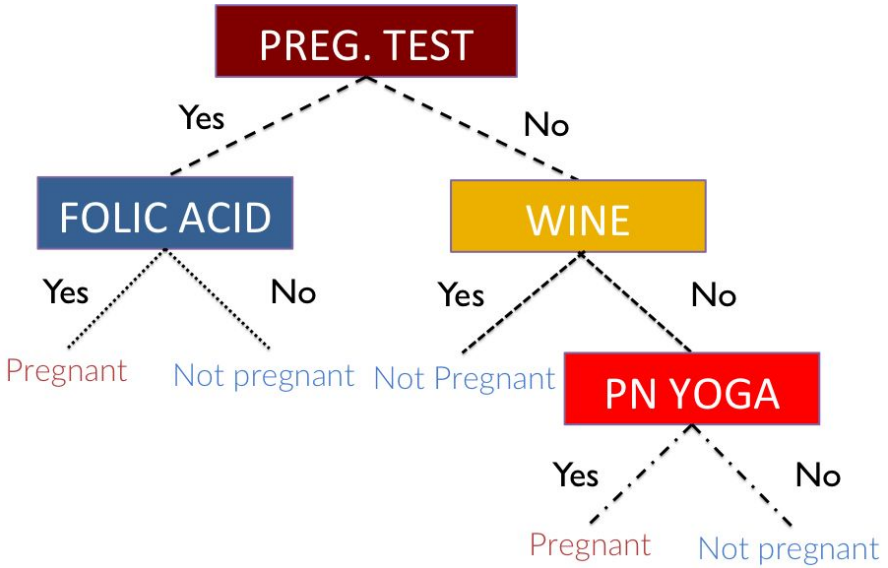
Recap: when should you use regression?

- Use regression when:

1. You have a labeled dataset 
2. You want to predict trends 
3. You want to anticipate needs or shortages 

Polling question

Do you think the decision tree shown below depicts a classification method?



Polling question

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

How Machines Learn

Break

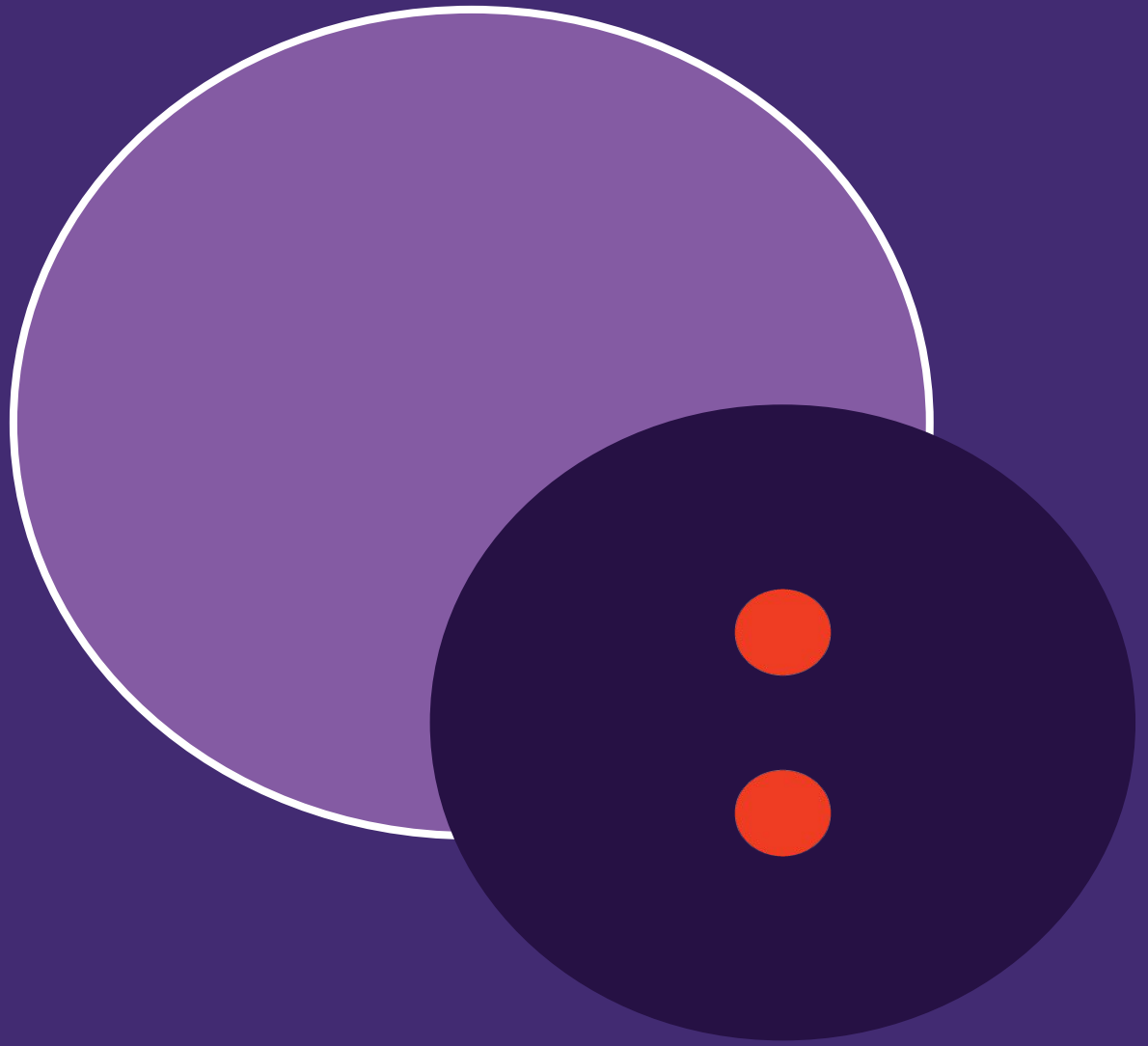
Agenda

Day 3

- Foundational data science methods
- Advanced data science methods

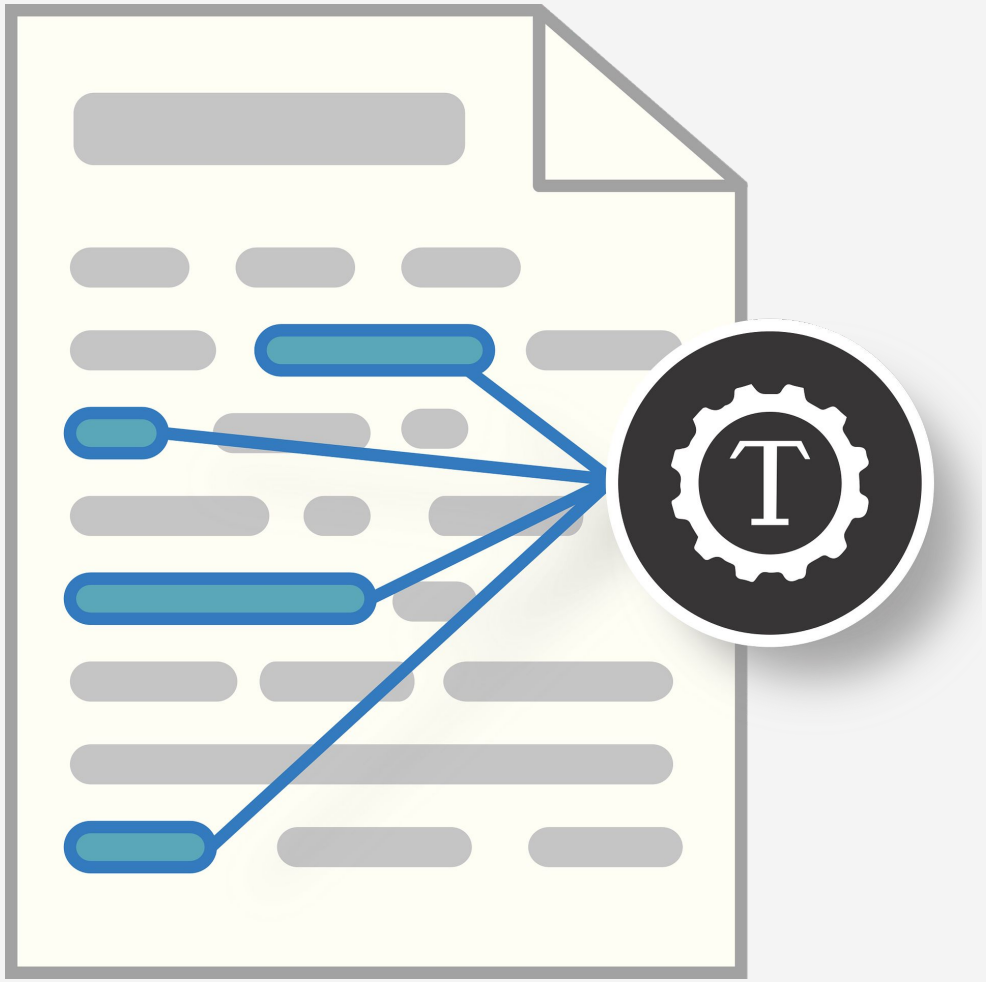
- What is text mining and how is it used?
- What is graph analysis and how is it used?
- What are neural networks and how are they used?

Text Mining



Text mining

- Text mining employs methods from various fields including mathematics, statistics, computational linguistics, and programming.
- It's the process of getting insightful and valuable information out of text data.
- Includes entity extraction, document classification, and sentiment analysis.

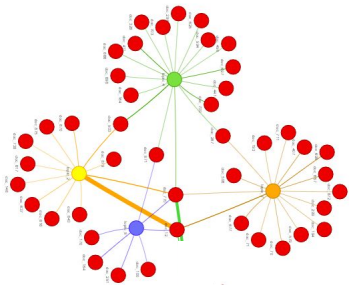


Text mining branches

Text mining

Entity extraction

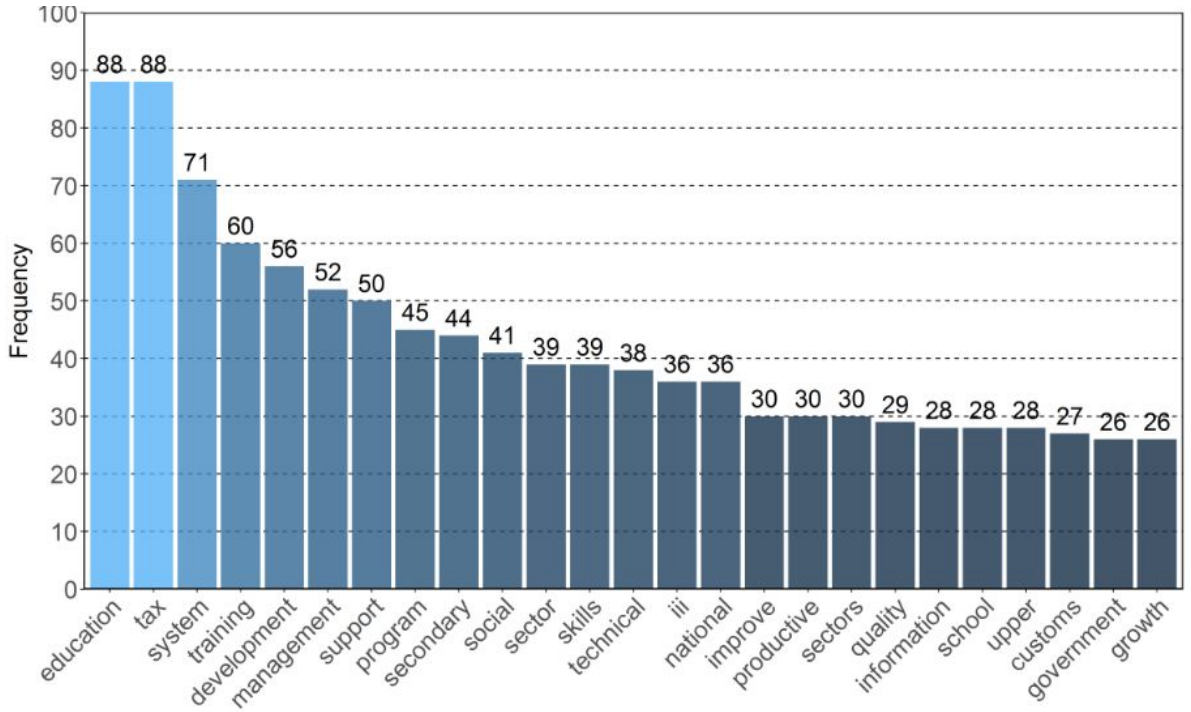
Document classification

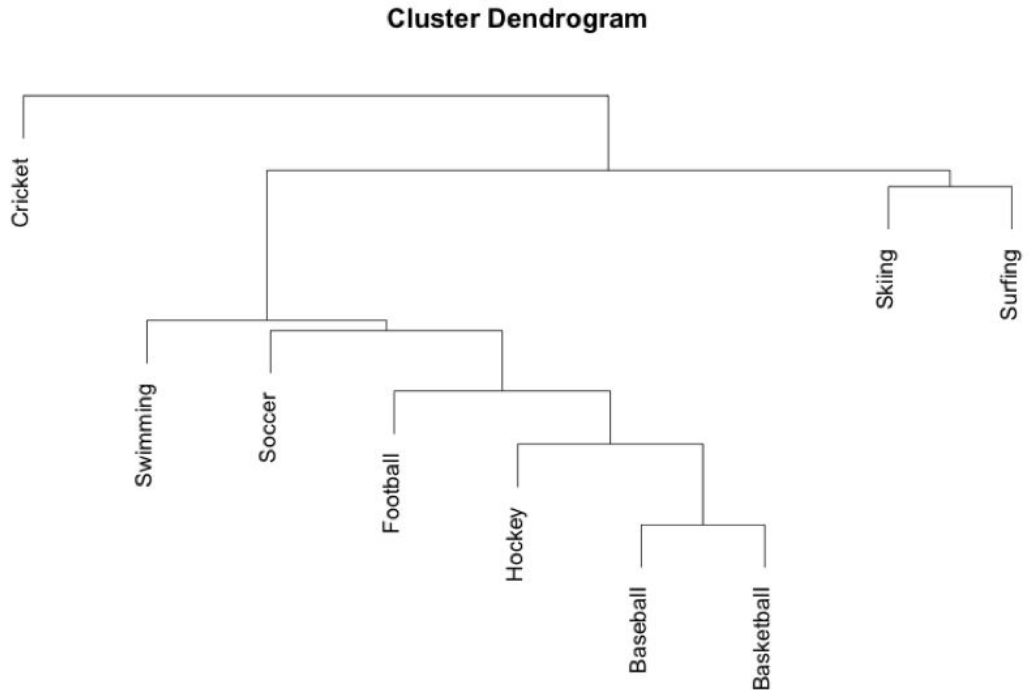


Sentiment analysis

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.

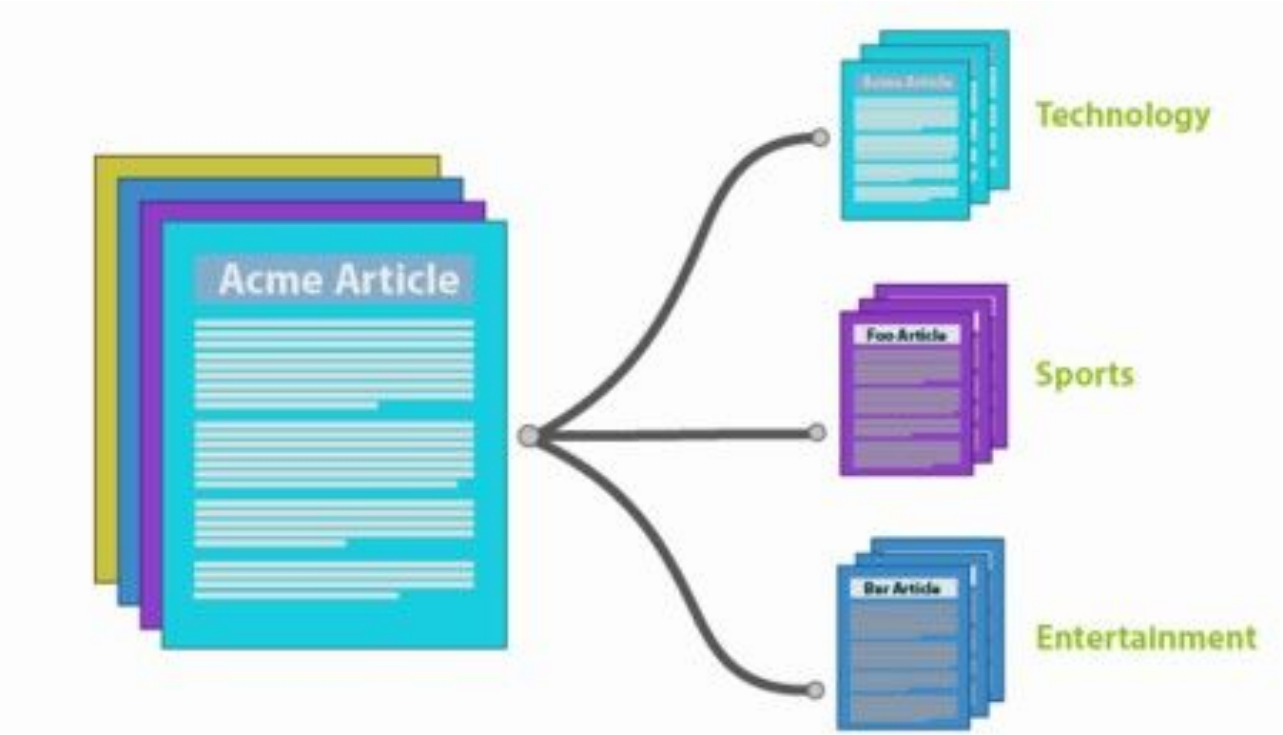




Document classification

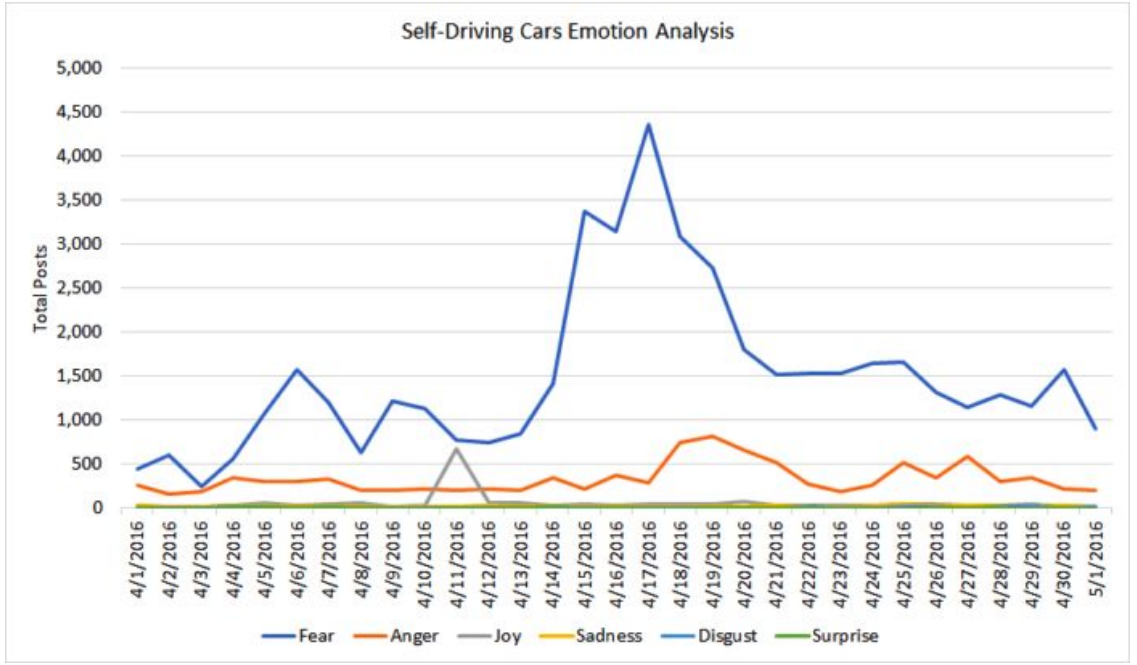
- Use document classification when you want to sort through documents and identify groups of similar articles.

- 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

Evaluating accuracy of our model

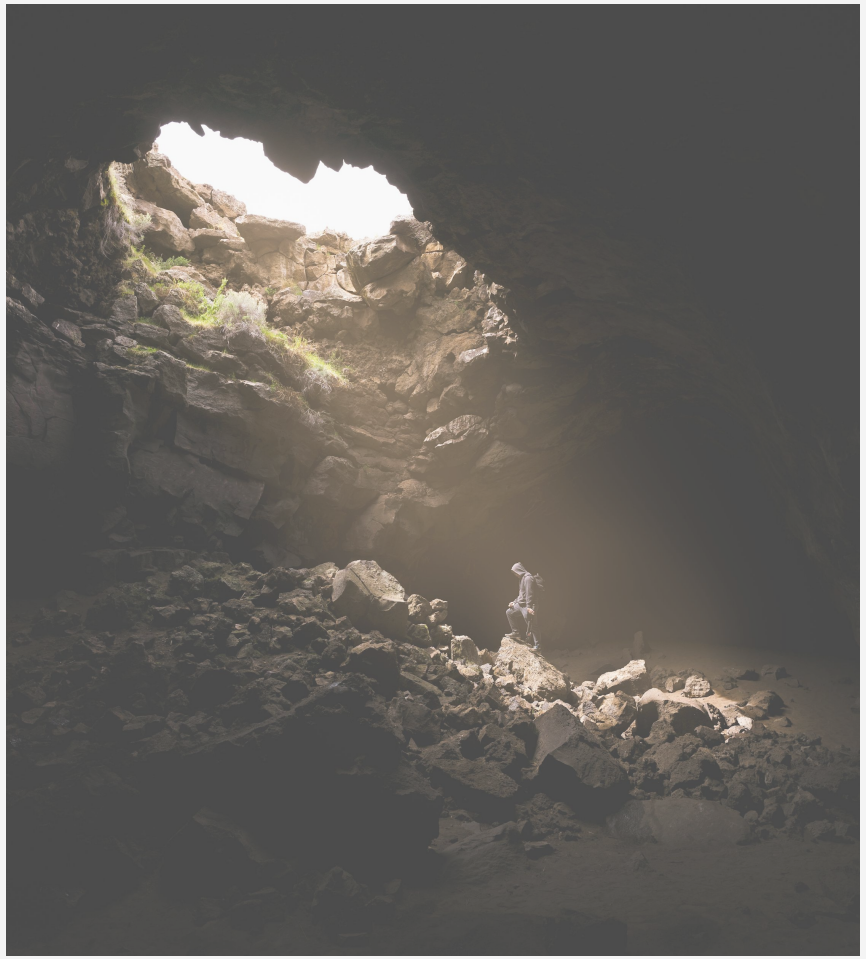
- This is a tricky subject!

- Text analysis and text mining rely on other methods that we've introduced in this class, such as clustering and classification. You'll need to use the evaluation methods for those particular models.

- In terms of sanity-checking the text mining process, look for unhelpful stop words (frequent words that don't provide additional information) and see if the topics generally make sense.

Common pitfalls with text mining

- Cleaning text is extremely messy and time consuming – this is a key problem in text mining projects.
- Existing dictionaries are not a panacea for catching the nuances of language – typically, there need to be manual additions of other words.
- Using the right methods and metrics to classify and cluster documents correctly.



Questions managers should ask

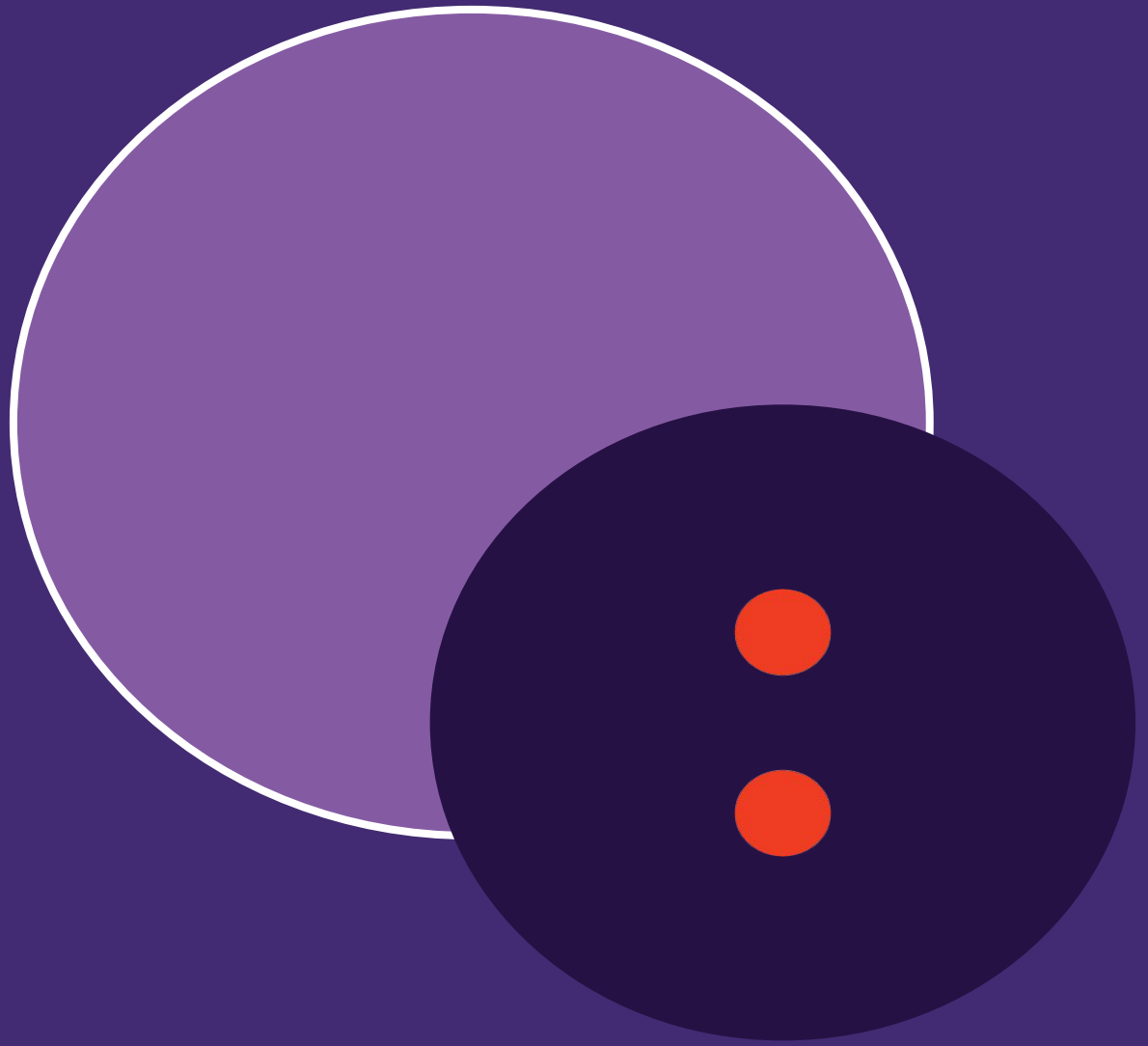
1. How does the model take sarcasm / irony / colloquialisms into account?

2. Is there an existing library of reference words that can assist you in text mining?

3. Does that reference library include misspellings, alternate versions of words, symbols, different parts of speech or compound terms?

4. How do the topics change over time?

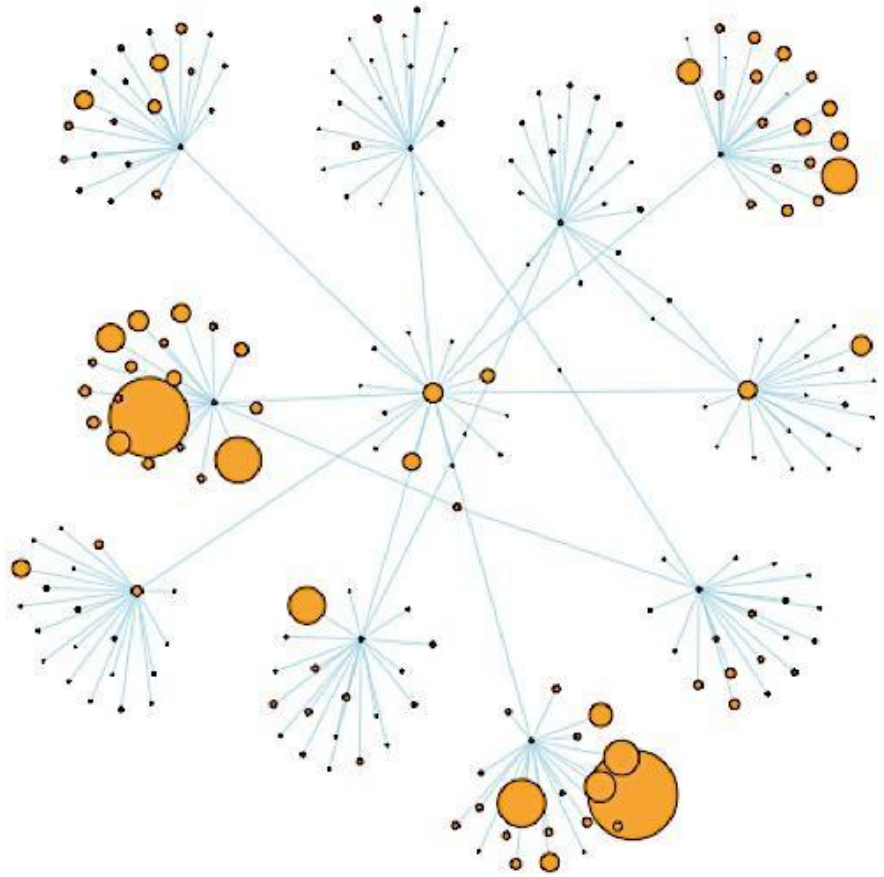
Graph analysis



Graph analysis

- Graph analysis (also known as network analysis) seeks to find patterns within a network, a set of points connected by lines that represent connections.

- Networks can represent organizational relationships; communications patterns; economic relationships; environmental relationships; connections based on interests, preferences and similarities; as well as geographic relationships.



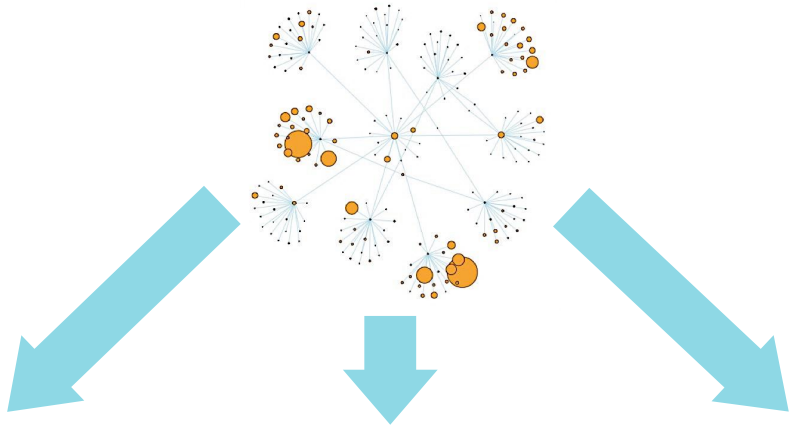
Example: IBM & a volcano

- In April 2010 a volcano in Iceland halted flights throughout Europe.
- IBM's internal analytics software alerted the team that IBM's supply chain link most relevant to the eruption was in Hong Kong – not Europe!
- The software showed that when flights resumed after the eruption was over, IBM would need to quickly move a backlog of components from Asian manufacturers to European customers. A bottleneck in Hong Kong would result.
- IBM booked additional space on commercial flights to help transport the backlog.

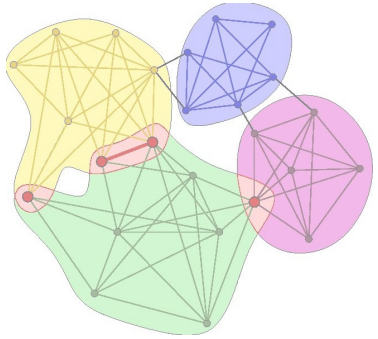
Source: Big Data Driven Supply Chain Management by Nada R. Sanders

Types of graph analysis

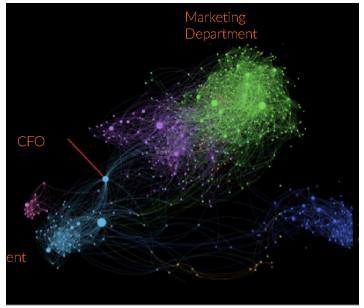
Graph analysis



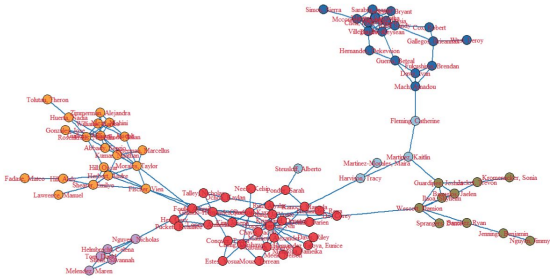
Community detection



Centrality metrics



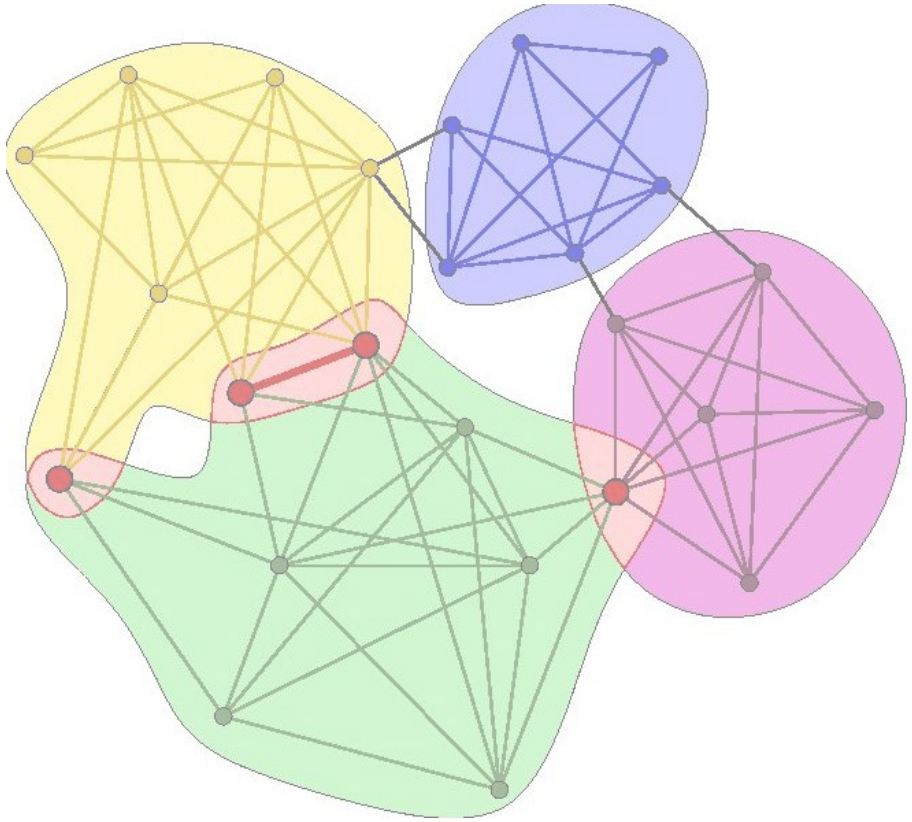
Social Media



Community detection

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

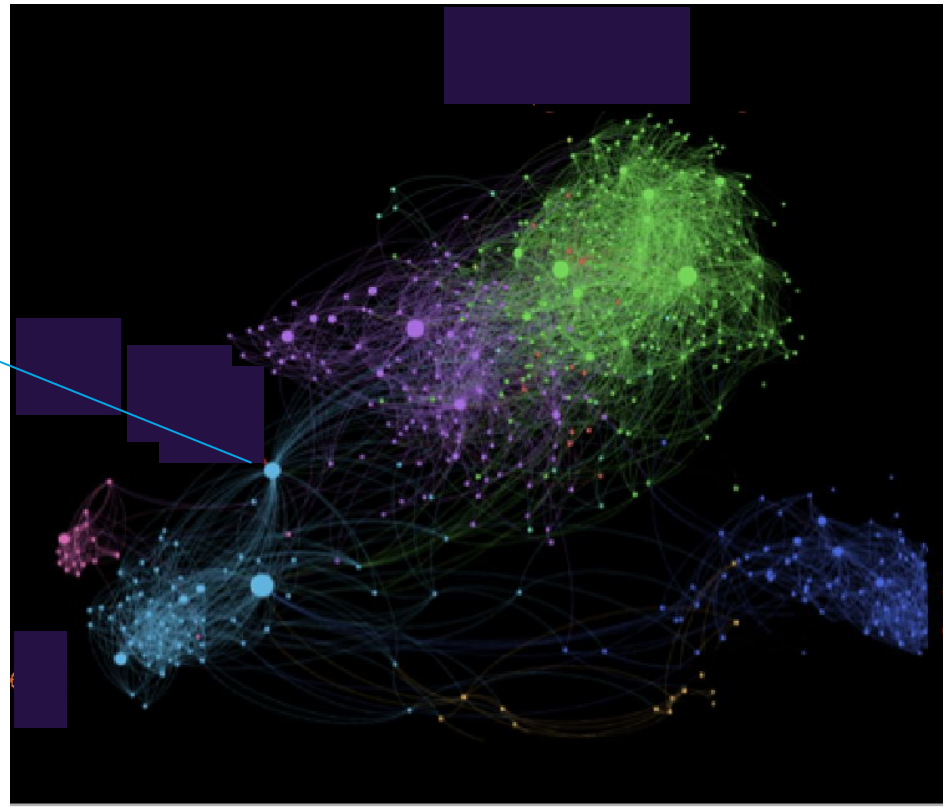
- Identifies groups of individuals / nodes that belong together; can detect latent connections and communities.



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.

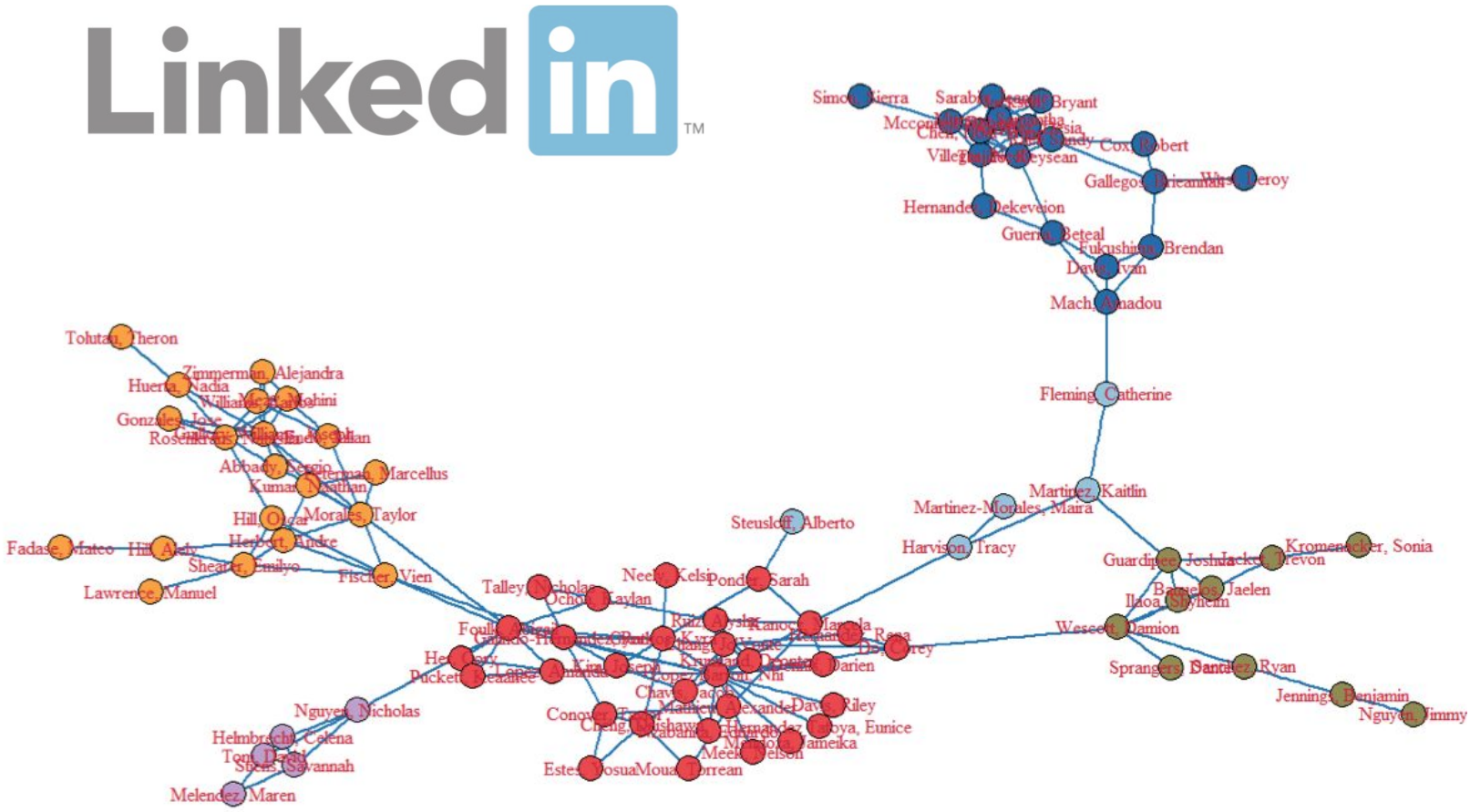
This email network shows how a company communicates.



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.



Ways to measure networks

Metric	Purpose
# of nodes	How many participants are included in the network?
# of edges	How many connections exist in a network?
Distance	How long does it take for information to travel through a network?
Degree (in-, out-)	Direction of connections, is someone a follower or an opinion leader?
Degree centrality	How many other people/objects can someone/something reach?
Closeness centrality	On average, how quickly can someone/something reach every other point in the network?
Betweenness centrality	How important is someone/something as a connector to the structure of the network?
Eigenvector centrality	How important is someone/something based on who/what else they are connected to?
Tie strength	How strong or significant is a connection between two people/objects?
Density	How sparse and fragile or inter-connected and resilient is a network?
Jaccard Index	How similar or redundant are 2 people/elements of a network?

Evaluating accuracy of our model

- This is a tricky subject!

- Graph analysis relies on other methods that we've introduced in this class, such as clustering and classification. You'll need to use the evaluation methods for those particular models.

- In terms of sanity-checking the process, look at how the nodes are accounted for in each community and determine what threshold makes the most sense for your analysis.

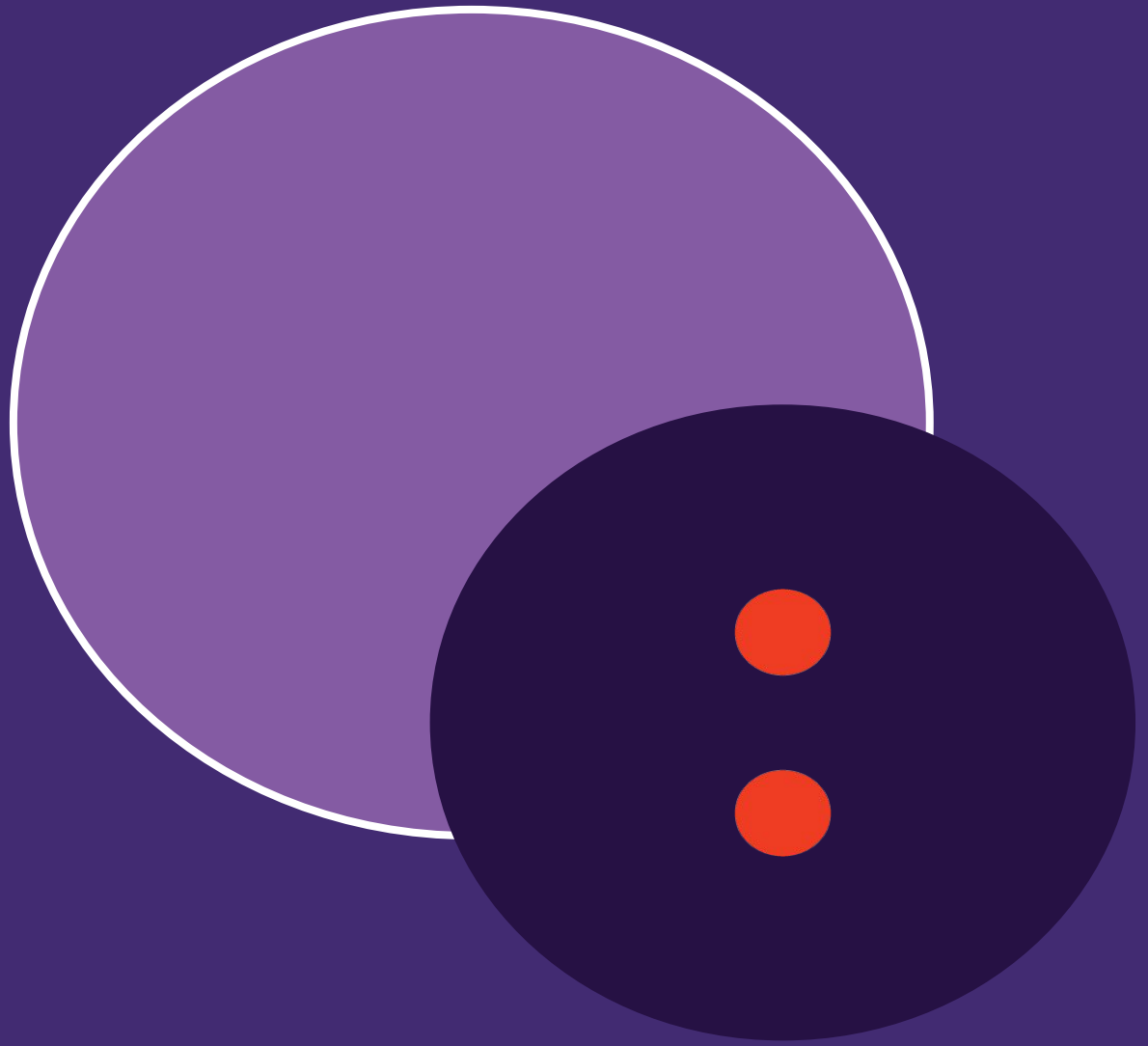
Questions managers should ask

- 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?

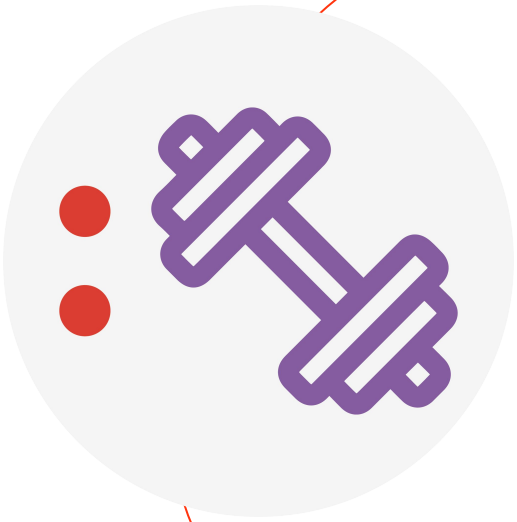
- What metrics did you use to evaluate the proximity between nodes / communities?

Neural networks



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.

Neural networks

- A neural network is born ignorant and builds on itself to get smarter and smarter.

- It starts out with a guess, and then tries to make better guesses as it learns from its mistakes.

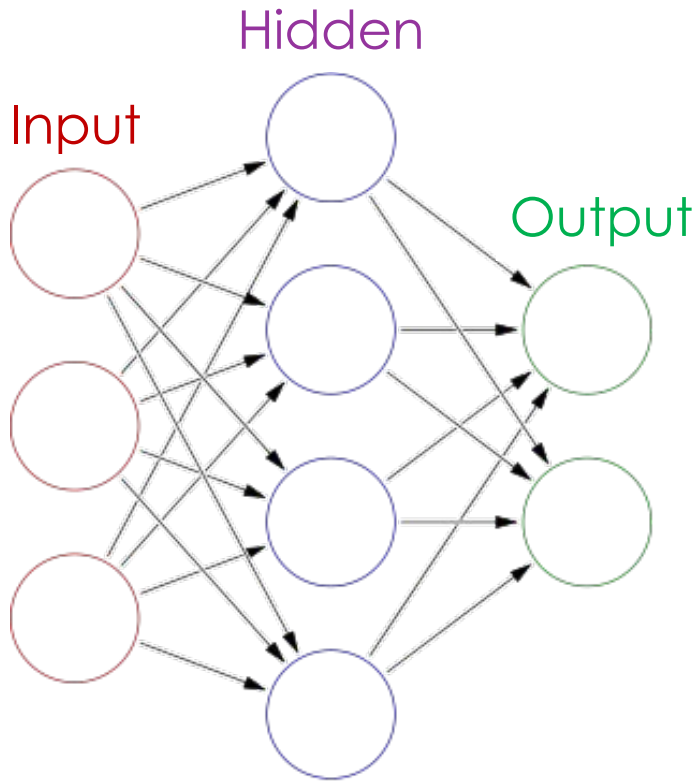
- Neural networks cover the same topics that we've reviewed previously. In theory, you can apply them to almost any method!

Intuition: neural networks

- Neural networks are made up of perceptrons.

- A simple perceptron has 3 layers:
  - Input: observations that enter the model
  - Hidden layer: composed of an activation function that derives the output based on inputs and other factors
  - Output: target variable you want to predict

- Once the output is produced, the model measures the error, then walks it back over the model to adjust its performance and reduce errors.



A perceptron acts like a neuron.

What data do you need?

1. Relevant: data must resemble the real-world data you hope to process as much as possible

2. Properly classified: in order for a deep-learning solution to correctly classify, a labeled dataset is needed. If a labeled dataset is not available, someone needs to actually apply the labels to the raw data

3. Formatted: all data needs to be vectorized, and the vectors should be the same length when they enter the neural net

4. Minimum data requirement: this may vary with the complexity of the problem, but 100,000 instances in total across all categories is a good place to start

Neural networks: pros and cons

- Pros
  - Neural networks are highly versatile.
  - They are fairly insensitive to noise in your data.
  - They are well-equipped to handle fuzzy and convoluted relationships.

- Cons
  - It's a black box – those hidden layers are difficult to explain and evaluate.
  - They are in danger of overfitting the training data, so it might not generalize as well to new information.
  - An experienced data scientist should develop the parameters of hidden layers and nodes.

Chat question

We started our discussion on neural networks with a drawing activity...

How many of your drawings did the neural network guess correctly?

Does that mean you are a good (or bad) artist?

Key points

- Don't accept an analysis at face value – you need to ask the right questions!

- Most data analyses incorporate multiple methods in order to determine which one is the most accurate.

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

End of Day 3

Foundational data science methods
Advanced data science methods

DATASOCIETY:

DATA LITERACY FOR MANAGERS

DAY 4

Agenda

Day 4

- Data visualization
- Misleading statistics & visual distortions
- Data storytelling

- What is data visualization?
- How to I pick and design visuals?

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.

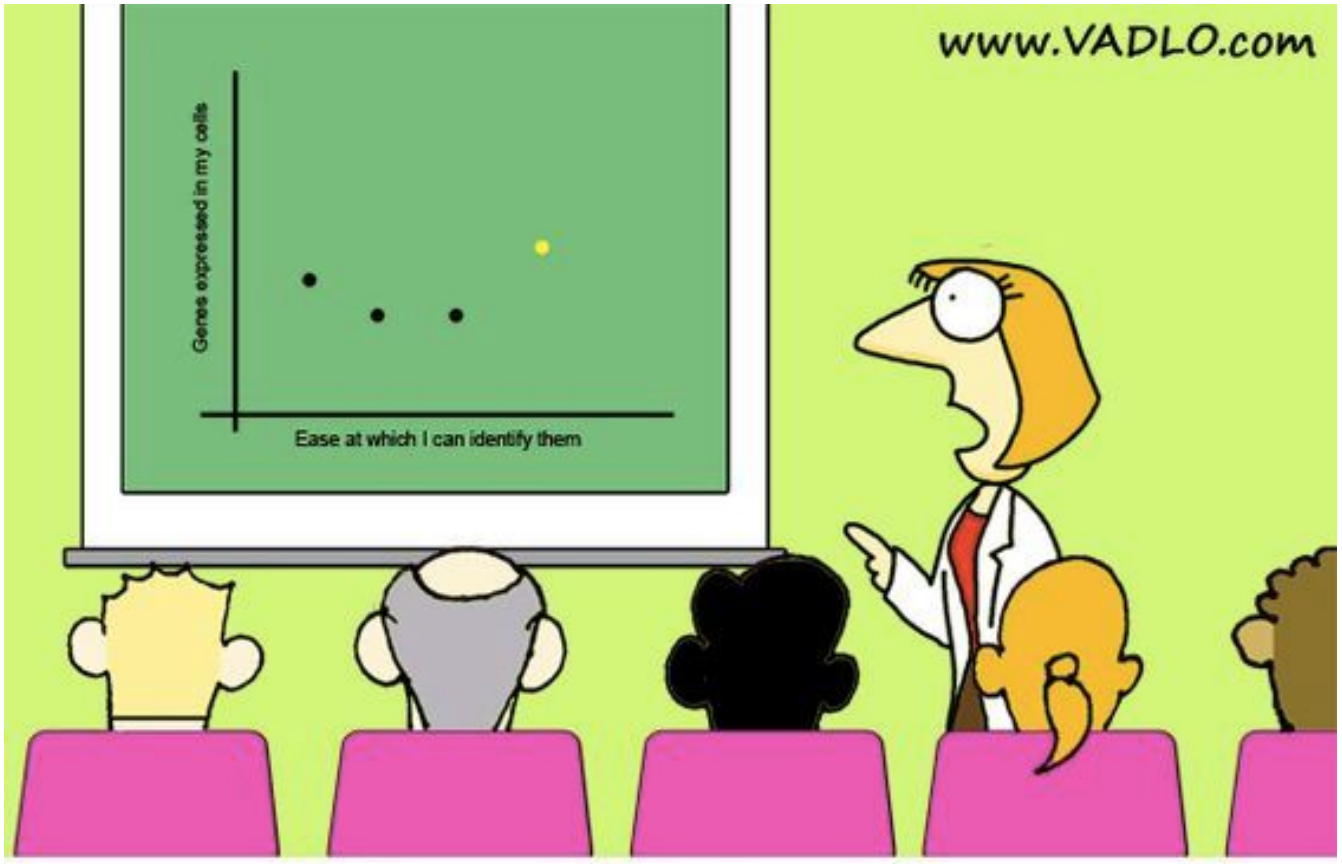
Chat questions

- What types of data visualization does your organization produce?

- What improvements would you like to see in the visualizations created or used by your organization? Why?

Getting it right

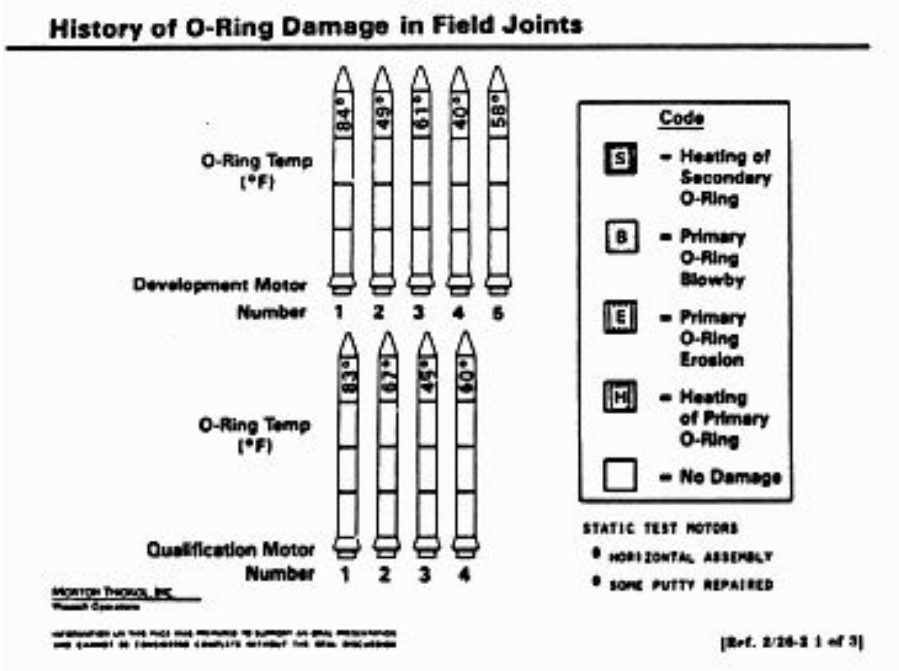
Using visualizations incorrectly can cause you to lose your audience, lose the value in your data, and ultimately lead to poor decision making.

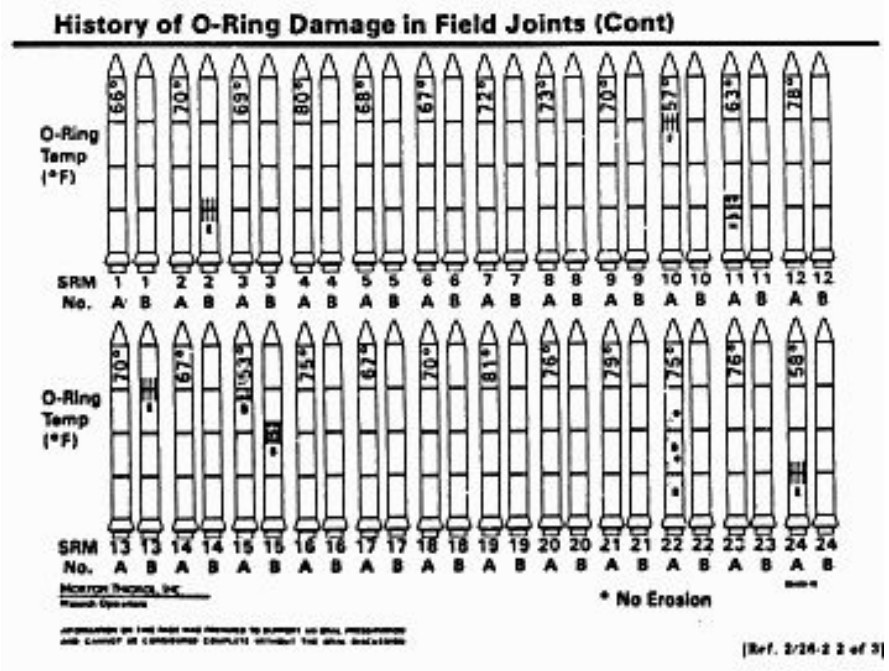


"Same graph as last year,
but now I have an additional dot."

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. 1593-1594

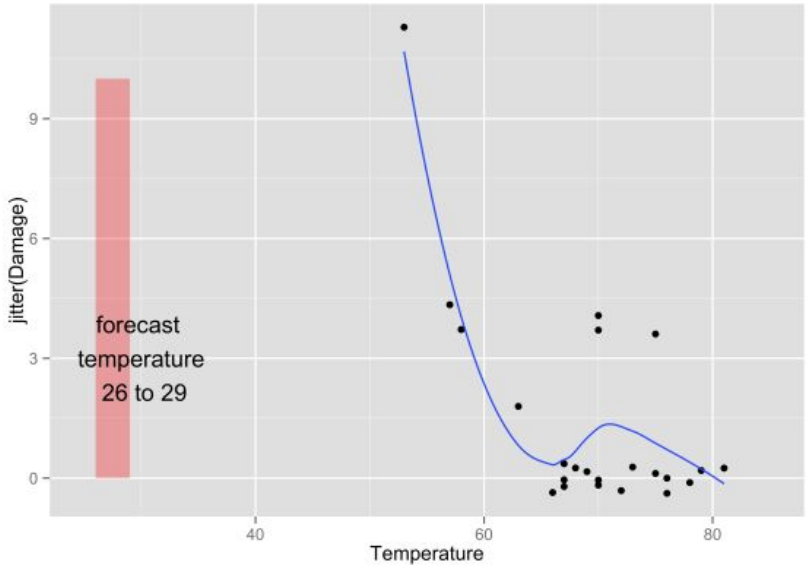
Example: The Challenger

- 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/

Using appropriate visuals

- We'll start by talking about how to pick and design the right visual for your purpose.
- Then, we'll discuss common mistakes.
- Later, we'll talk about how to avoid being misled by visualizations.

To get started with data viz

1. Know your audience and understand how it processes visual information. (Who)
2. Determine what you're trying to visualize and what kind of information you want to communicate. (What)
3. Choose a type of visual that conveys the information in the best and simplest form for your audience. (How)

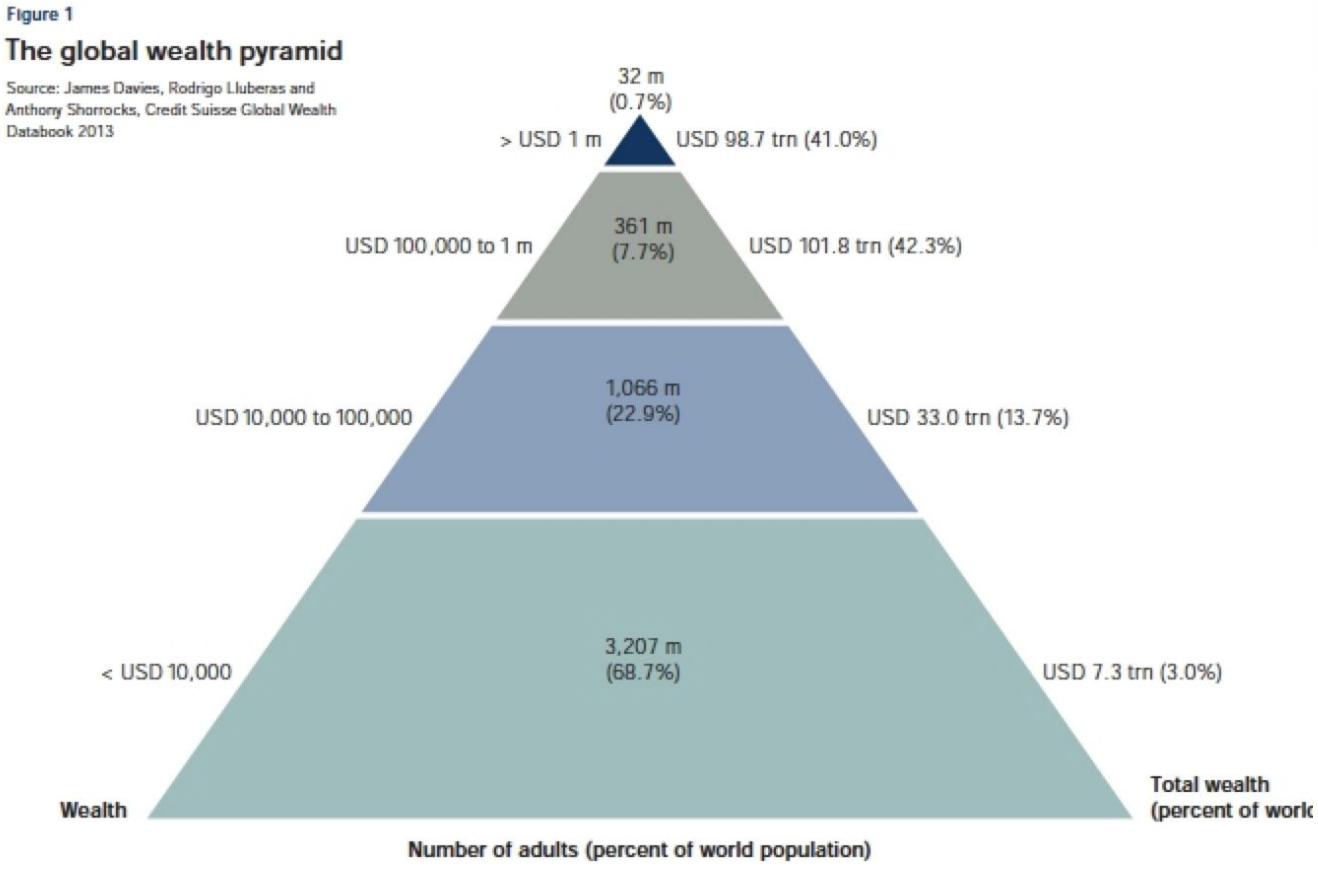
Who

- Know your audience and understand how it processes visual information.

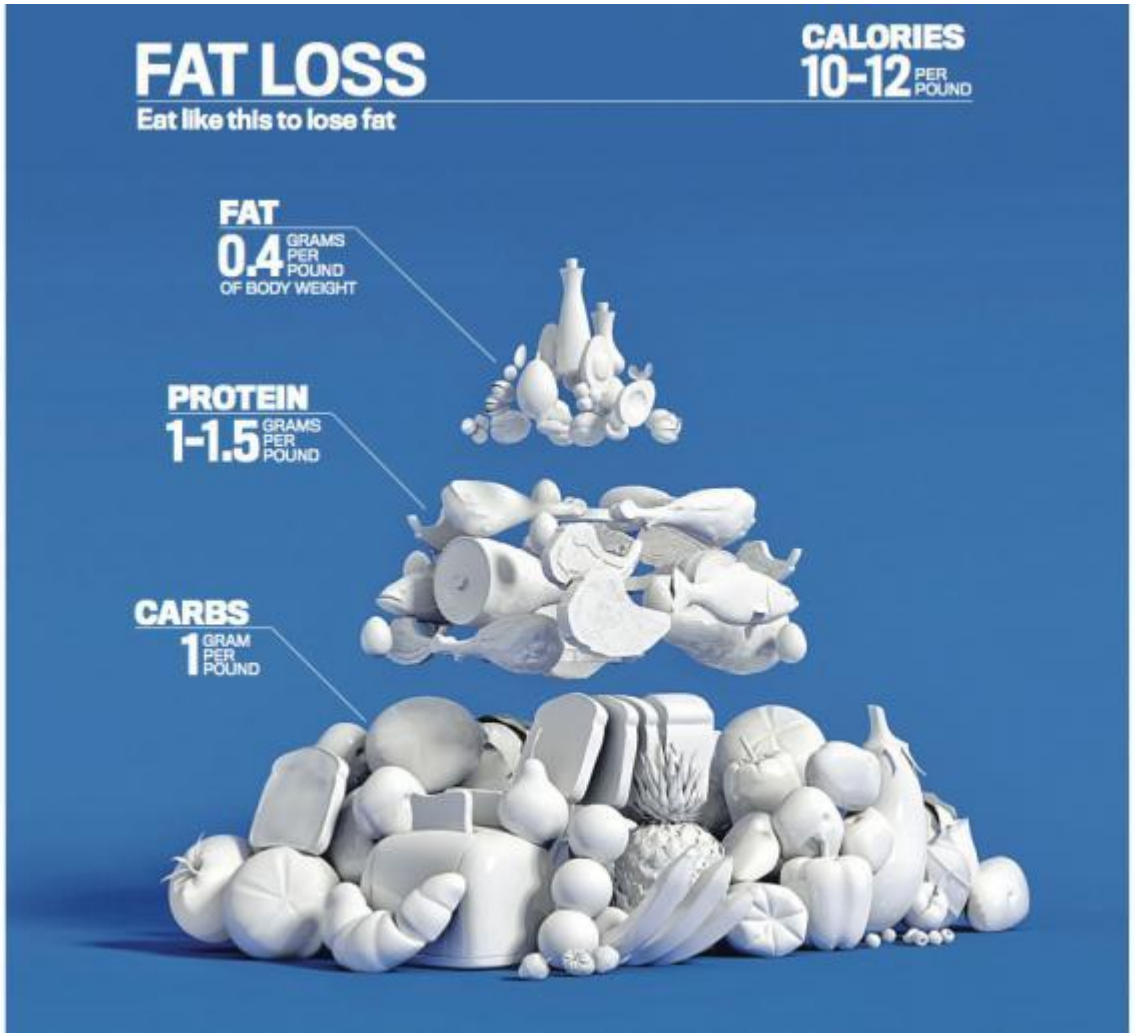
- Consider audience familiarity. For example:
  - High-level executives are generally well-versed in visual data, so use a variety of methods to stand out
  - Less-experienced audiences will want it kept simple (e.g., pie charts, bar graphs, and word maps)

- Consider how the visualization will be used by the audience:
  - Is it for executives to use to make decisions?
  - Is it to inform the public?

Chat: who is the audience?



https://www.oreilly.com/library/view/learning-highcharts-4/9781783287451/ch08s04.html



http://www.mensfitness.com/nutrition/what-to-eat/mens-fitness-food-pyramid

What

- Determine what you're trying to visualize and what kind of information you want to communicate.

- Remember, the audience only knows as much as you tell them:
  - Do you want them to explore the data on their own? (exploratory analysis)
  - Do you want to tell a specific story about the data? (explanatory purposes)

- If the message is explanatory, consider:
  - What type of data you have on which to base the analysis?
  - What are the audience's topmost concerns or requirements?
  - What decisions can be made based on the results you provide?

How

- Choose a type of visual that conveys the information in the best and simplest form for your audience.

- The type of visual you use depends primarily on two things:
  1. the data you want to communicate
  2. what you want to convey about that data

- Then, choose the visual that will be easiest for your audience to read.
  - Aim for them to "get it" in 30 seconds or less.

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.

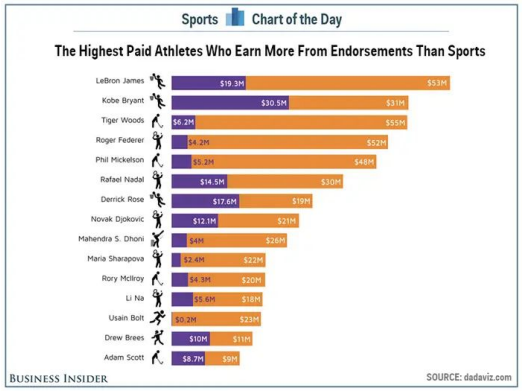
Name	Total Hours	Billable Hours
Aiello, Francisco P.	1,880	1,504
Eidson, Virginia D.	2,300	1,280
McVay, Dorothy	1,905	1,086
Ramos, Emilio Pabón	2,037	1,426

Product	Weight	Price
Toaster (UK)	1.05 kg	£17,49
Toaster (US)	3.13 lbs.	$29.99
Toaster (South Africa)	1.07 kg	R239,00

Common messages

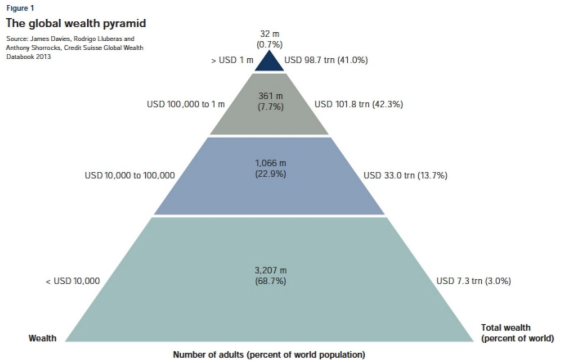
Comparison

Evaluate and compare values between two or more data points



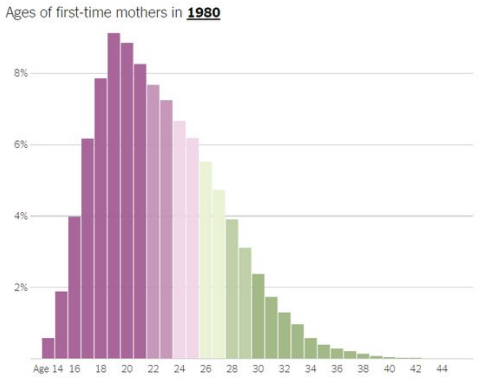
Composition

Understand how individual parts make up a whole



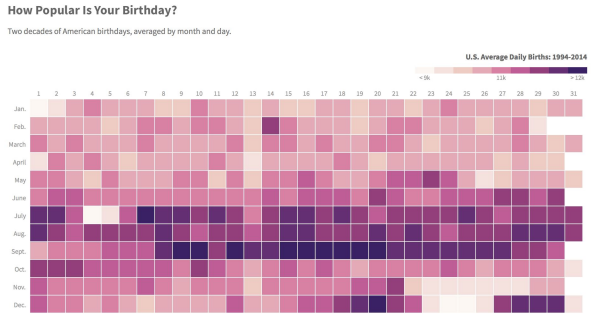
Distribution

Combine comparison and composition



Relationships

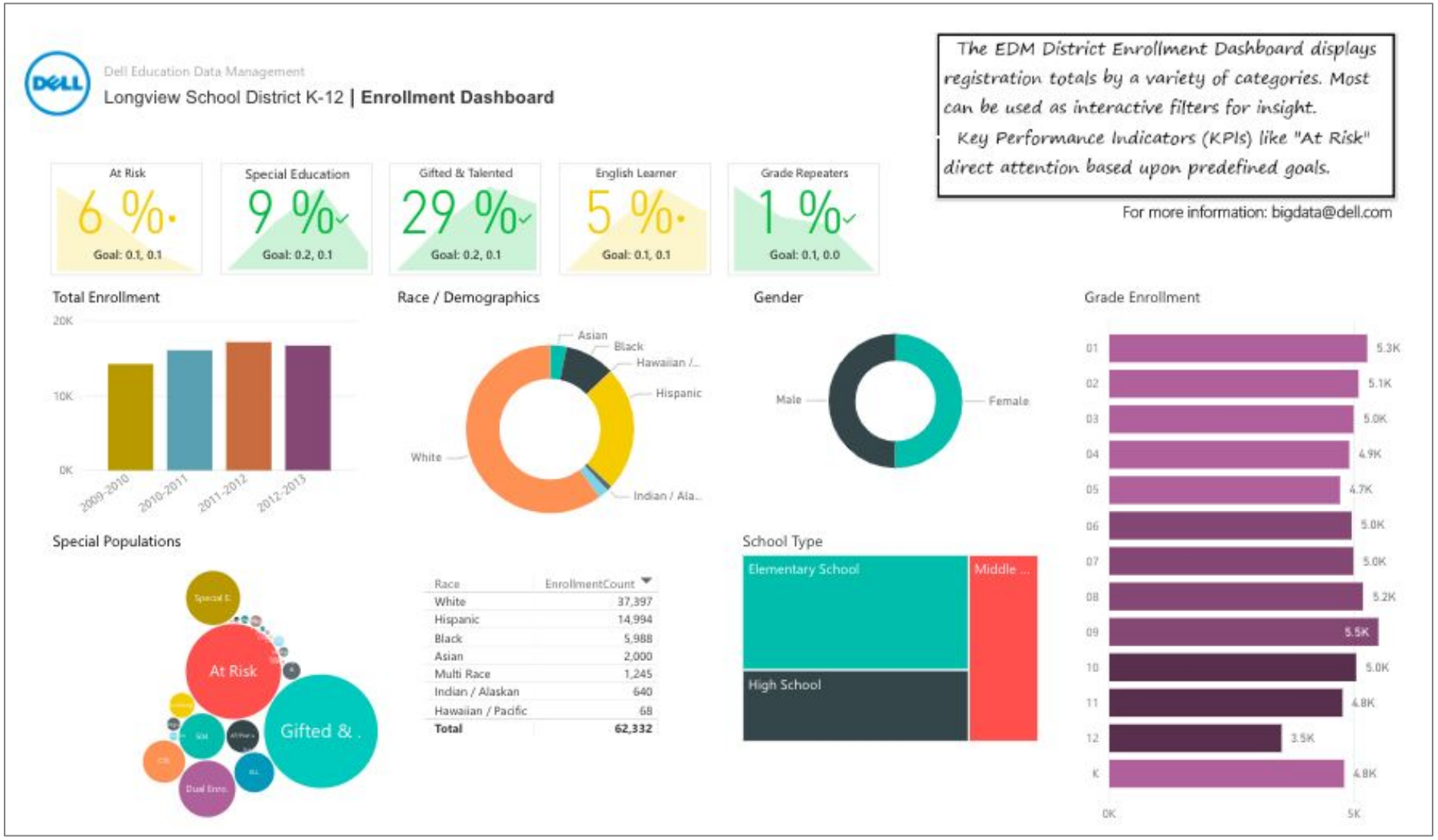
Represent the correlation or connection between 2+ variables



What if it's more complicated?

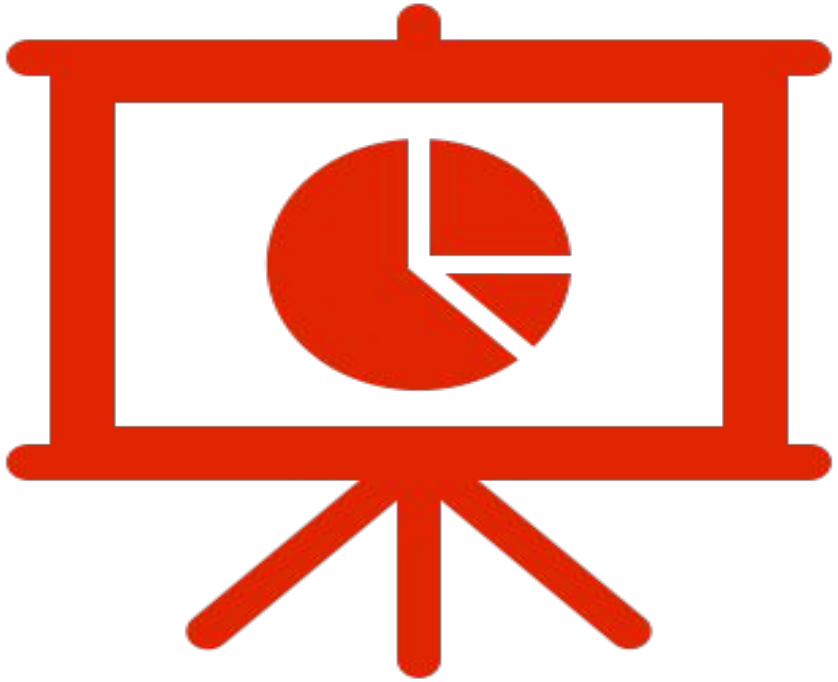
- A dashboard is a collection of visual reports that display important metrics and KPIs, usually in real-time.

Data visualization	Data dashboard
• a visual representation of your data, such as a chart • can be static or dynamic • typically shows data for a single metric, such as electricity usage	• a collection of data visualizations assembled into a single, unified view • might display data visualizations for electricity usage, energy costs, CO2 emissions, and peak/off peak use



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.

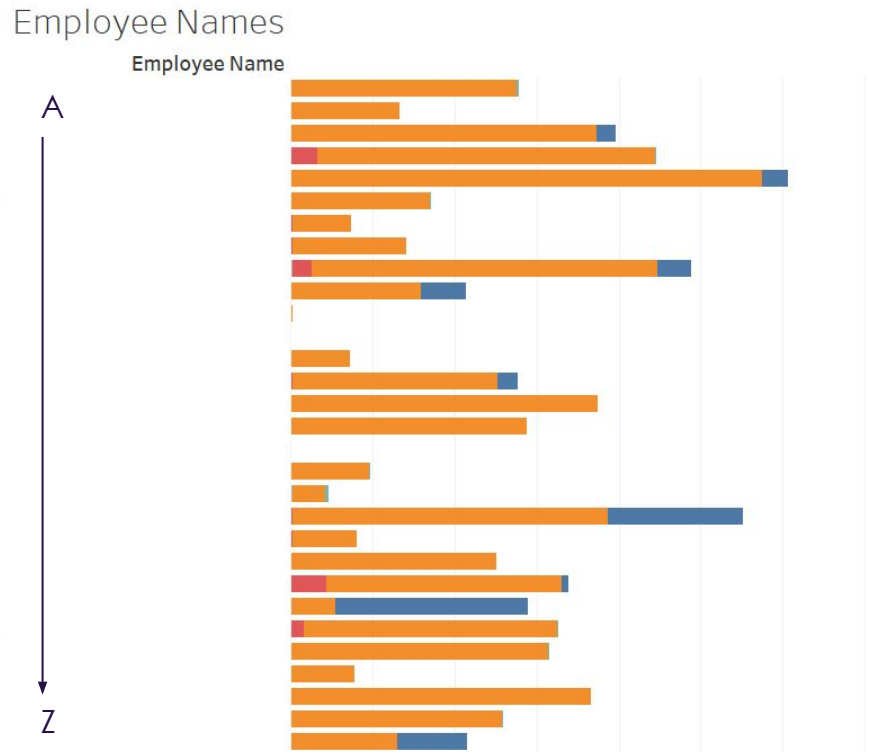
Theory

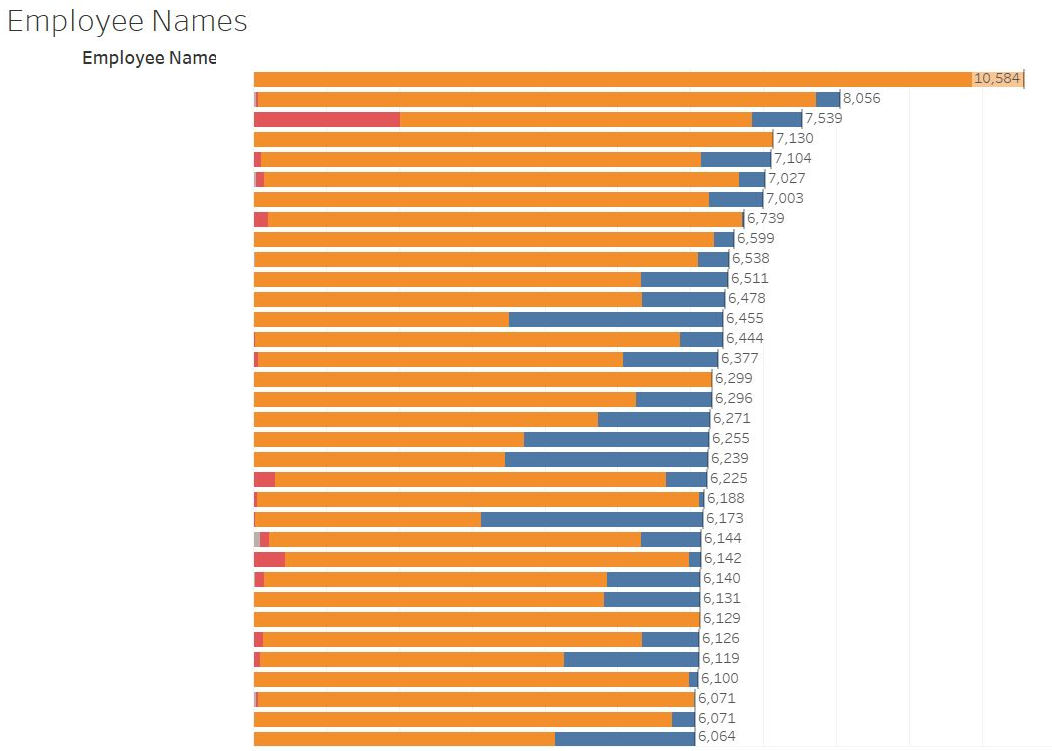
- The visual design tips we'll review today draw on theory such as:
  - the building blocks of visual design described by the Interaction Design Foundation
  - the four categories of preattentive visual attributes described in Colin Ware's book, Information Visualization: Perception for Design
  - the 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.





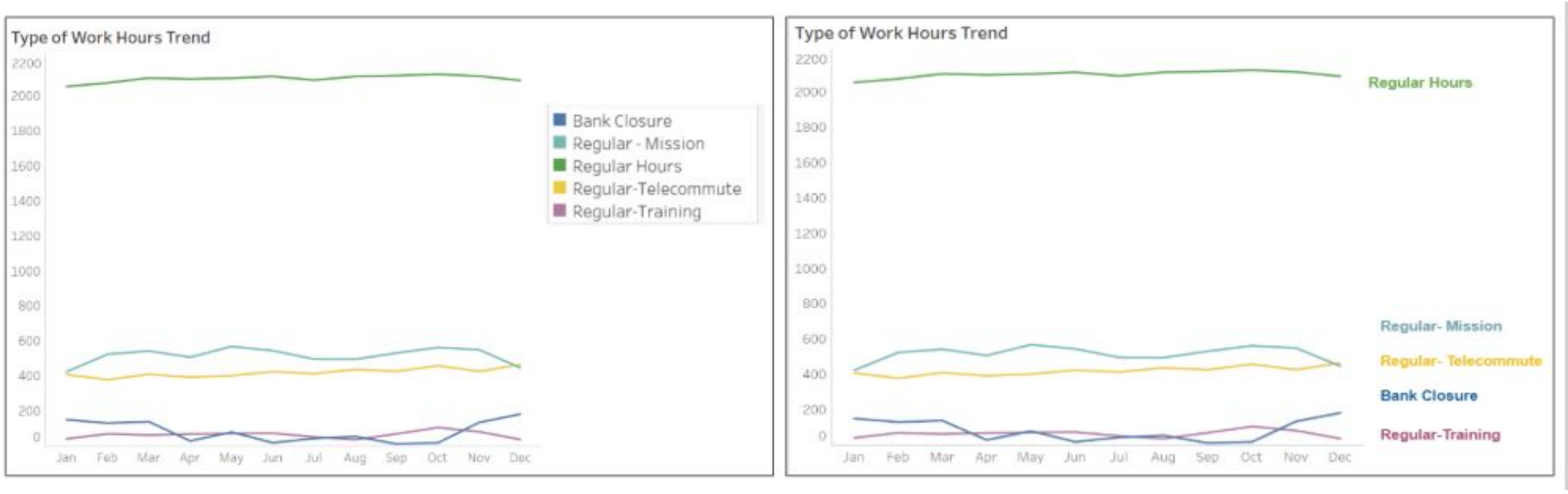
The left chart is sorted alphabetically; the right by value.

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).

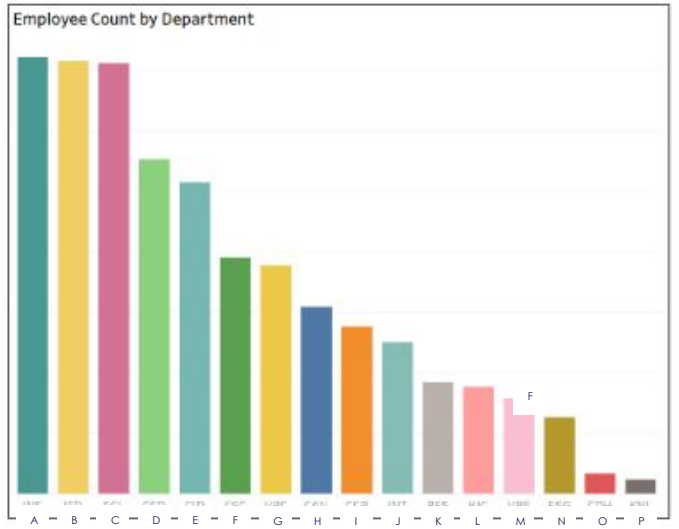


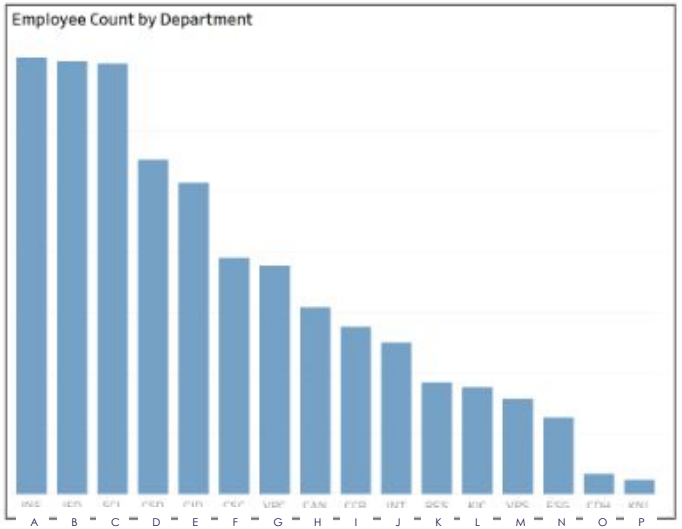
Tip 3

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

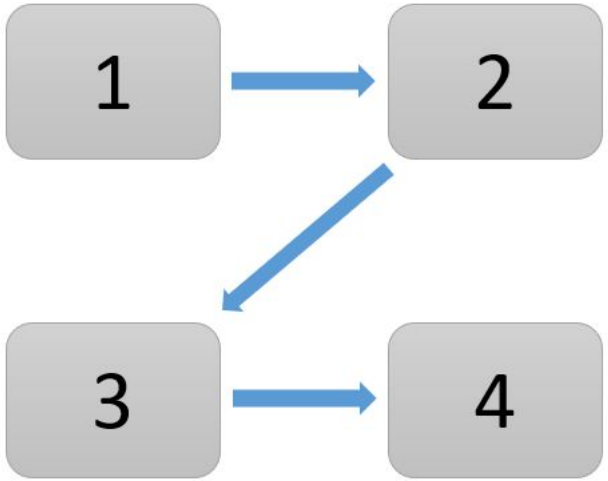




Tip 4

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.

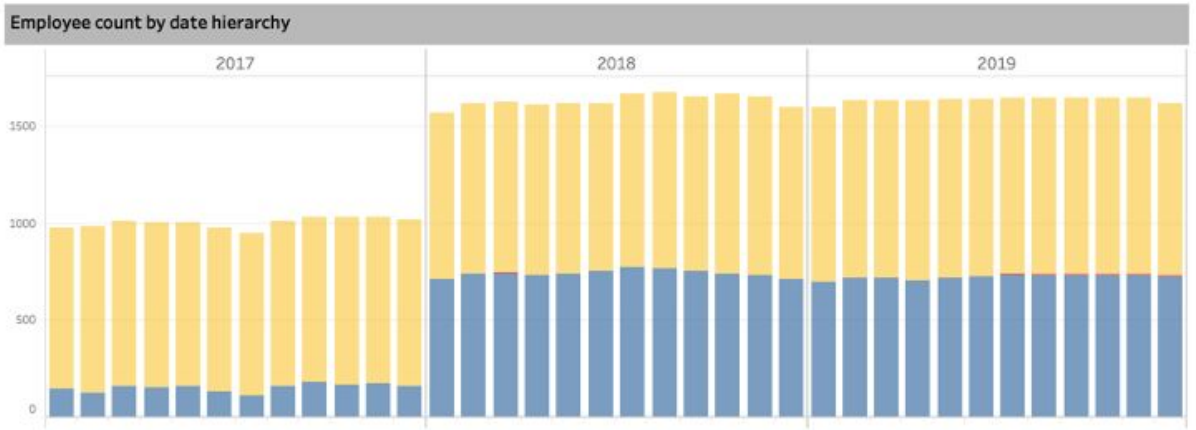
- Also, remember that the top of the page is the most precious.

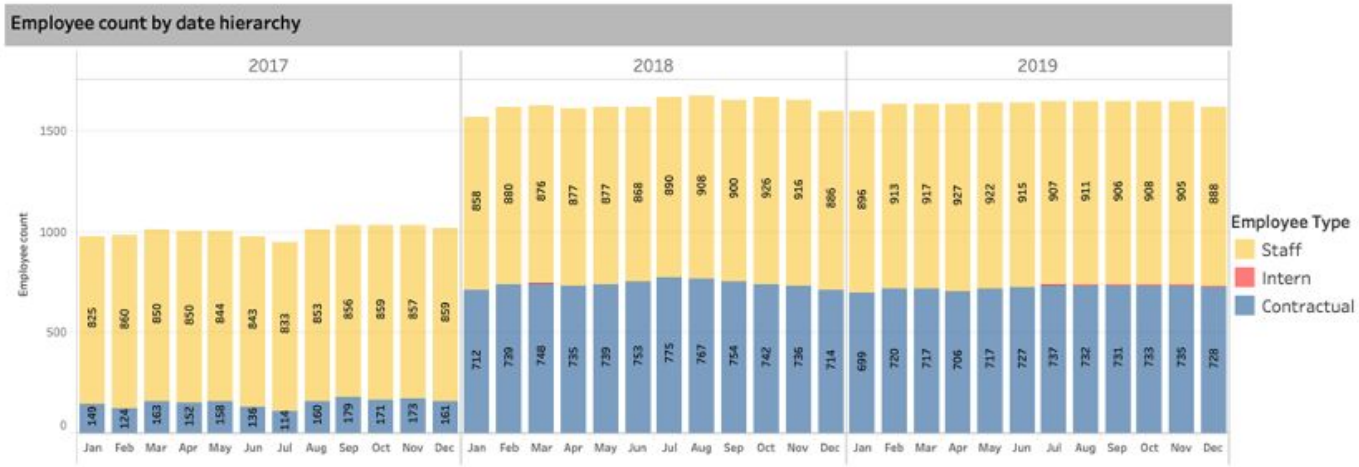
Tip 5

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.



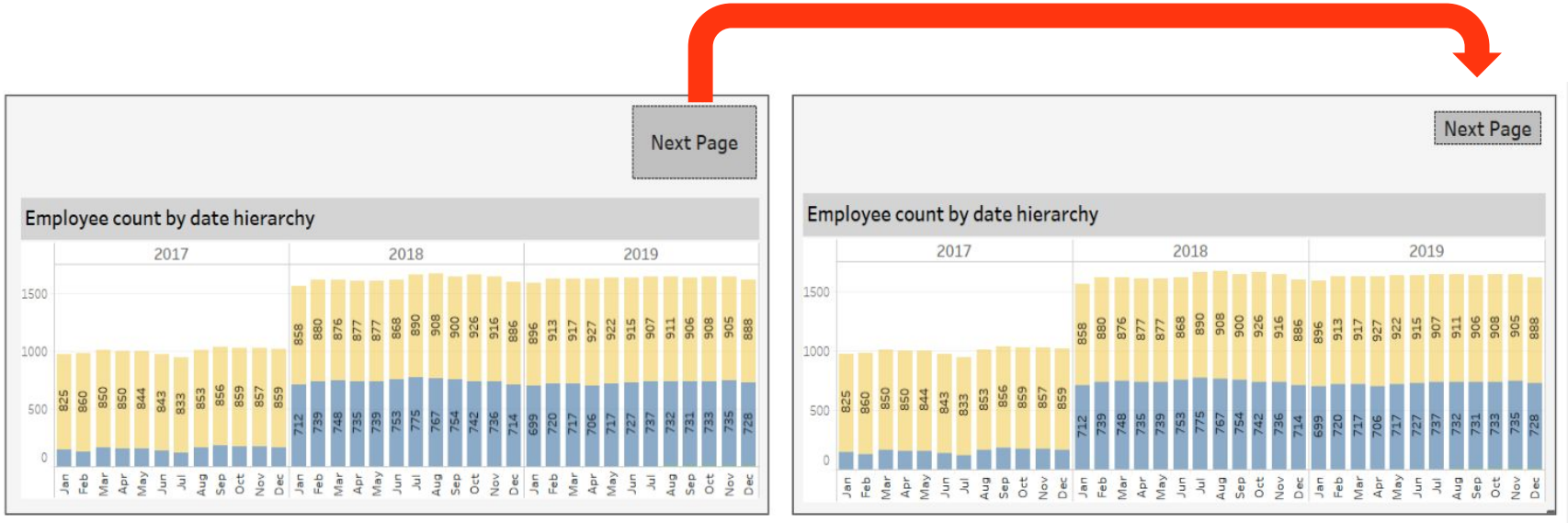


The labels and legends used in the bottom chart makes it easier to understand.

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.



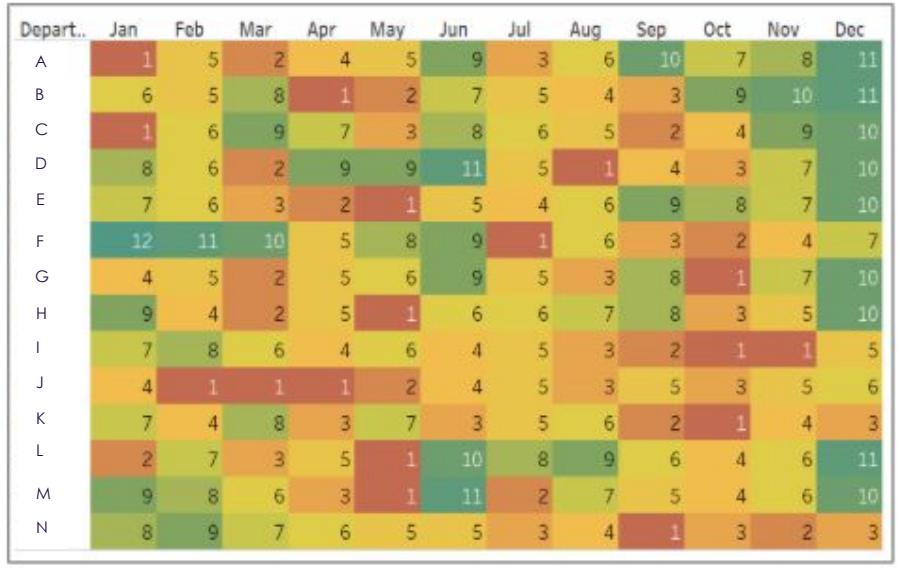
Resizing the "Next Page" button deemphasizes its importance.

Tip 7

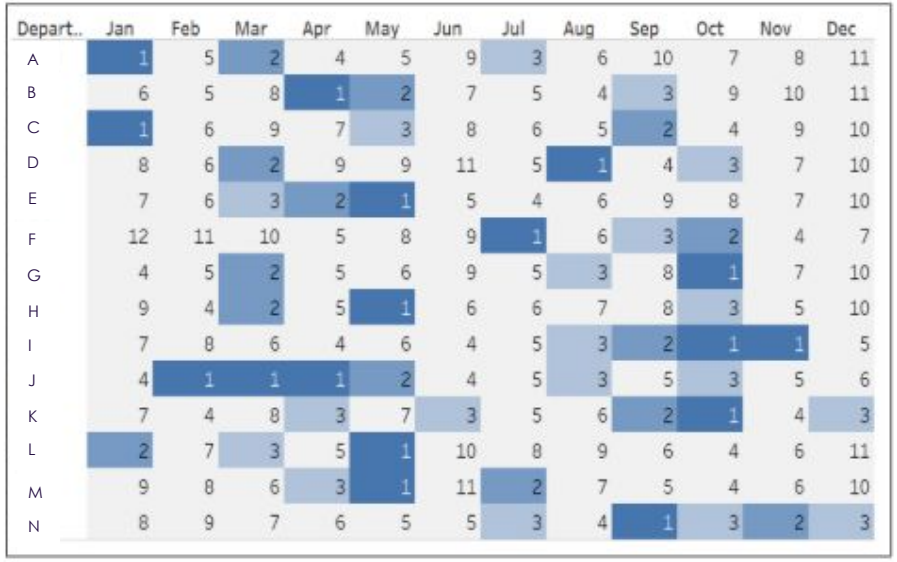
Use color to grab attention

- 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

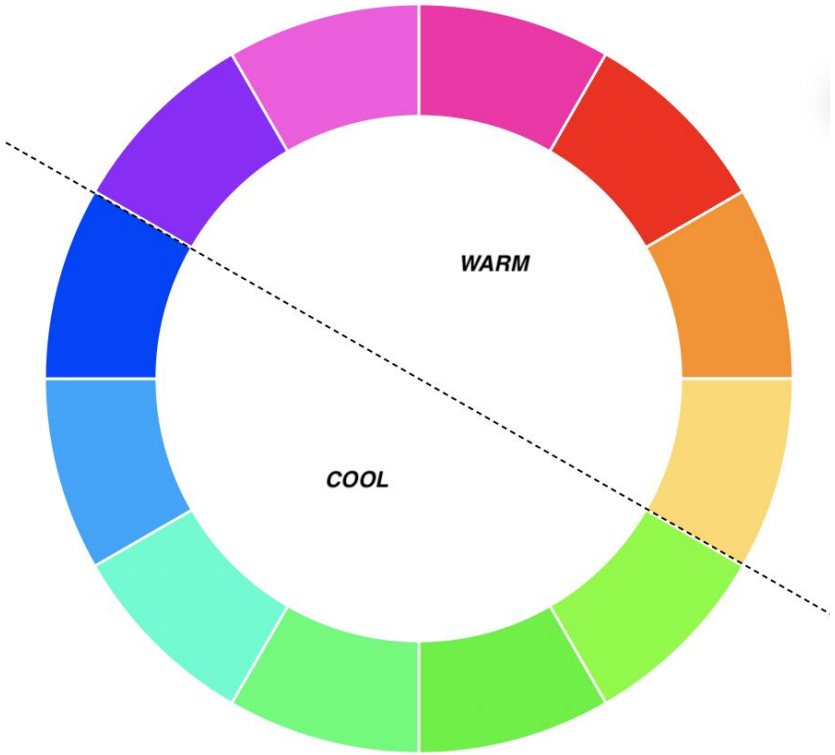
Tip 8

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

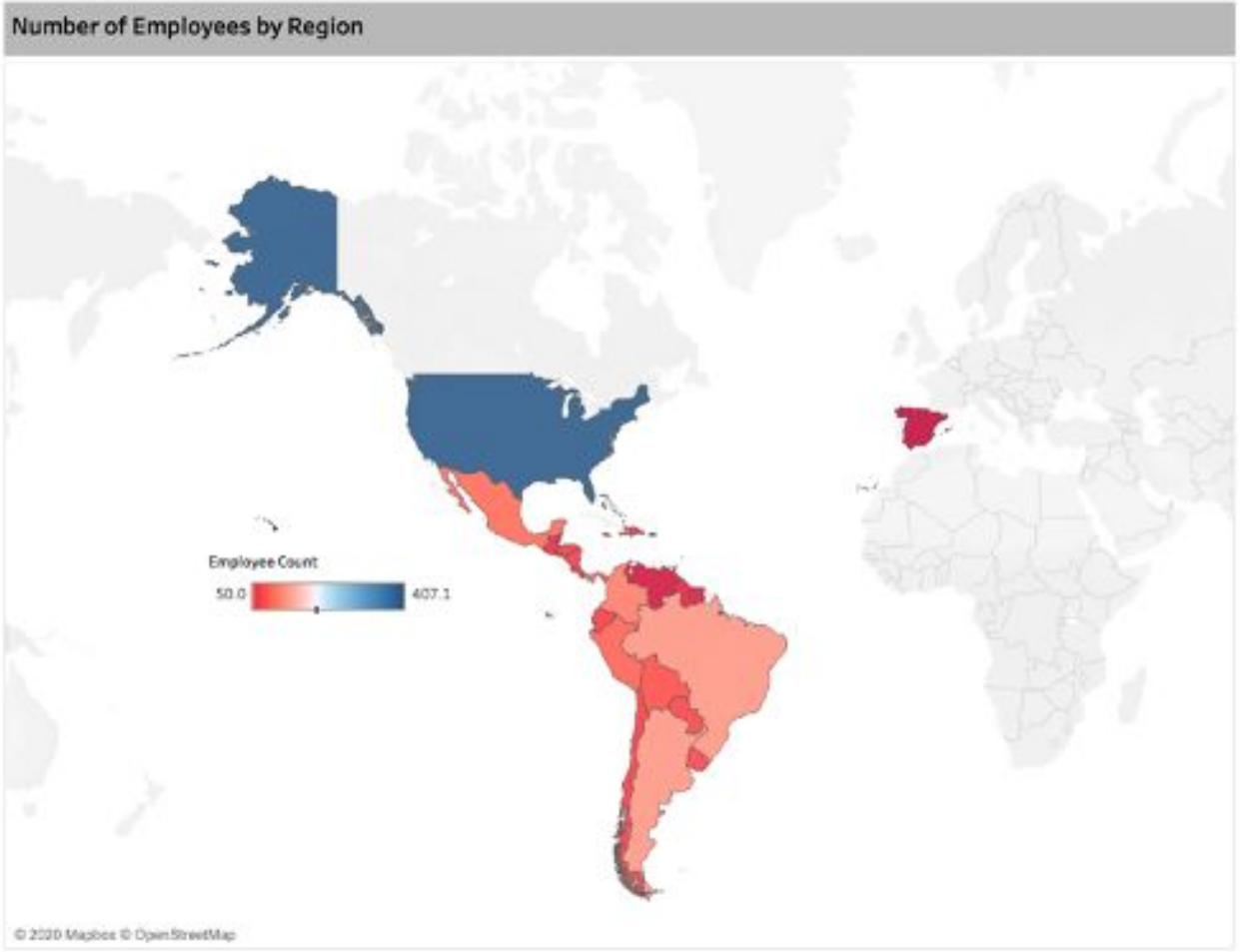
- Use a sequential color scheme when the order matters.

- 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.

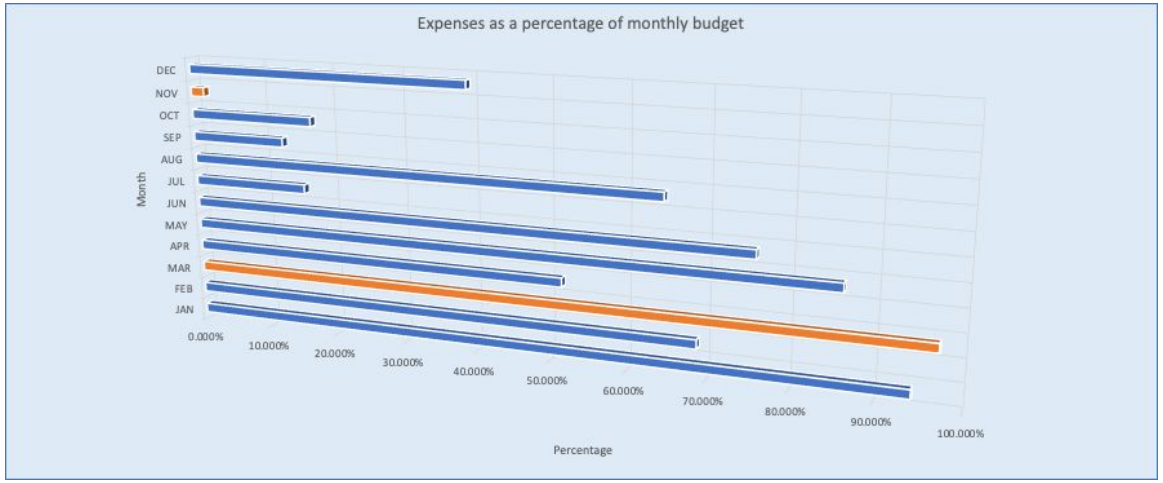
- These schemes use different hues with consistent steps in lightness and saturation.

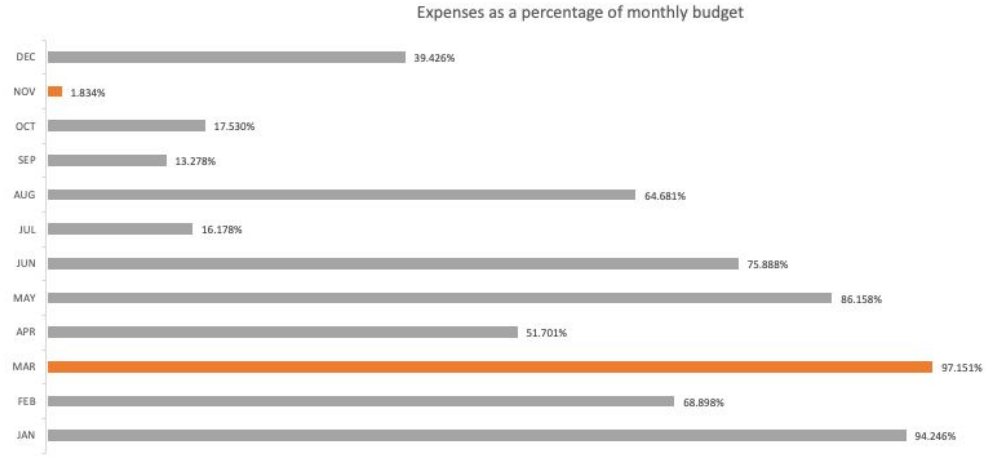
Tip 10

Reduce chart clutter

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

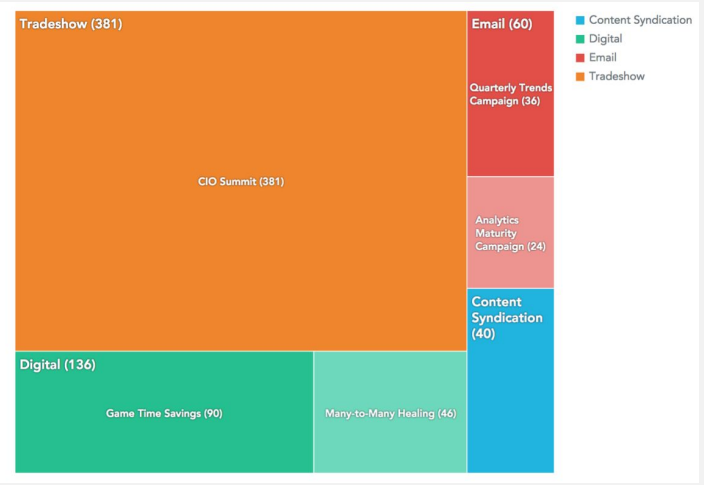
1. Remove special effects
2. Lighten the background
3. Remove chart borders
4. Remove gridlines
5. Direct label
6. Clean up axis titles and labels
7. Use consistent colors

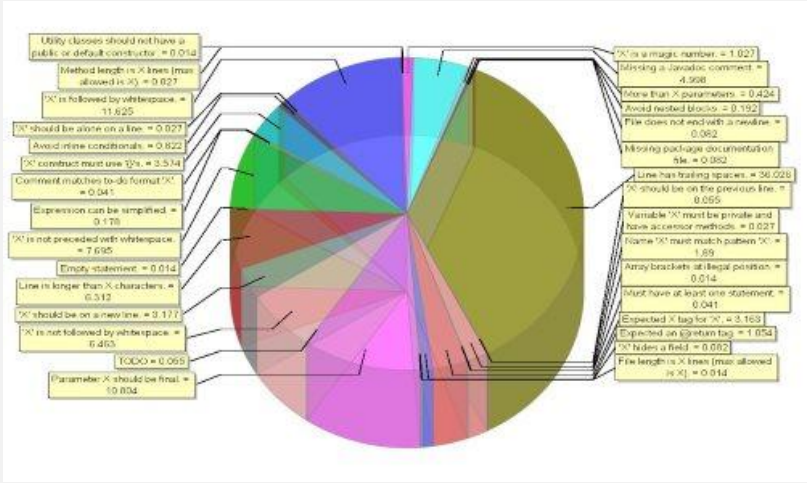




Activity: analyze visualizations

- Turn to **page 17** of your participant guide to find the **Analyzing visualizations** activity.
- You will be asked to assess 4 visualizations. Write down your notes.





Polling question

For each of the charts, select the best way to improve visual:

- Change colors
- Remove extra information
- Add more information

Which chart is the best?

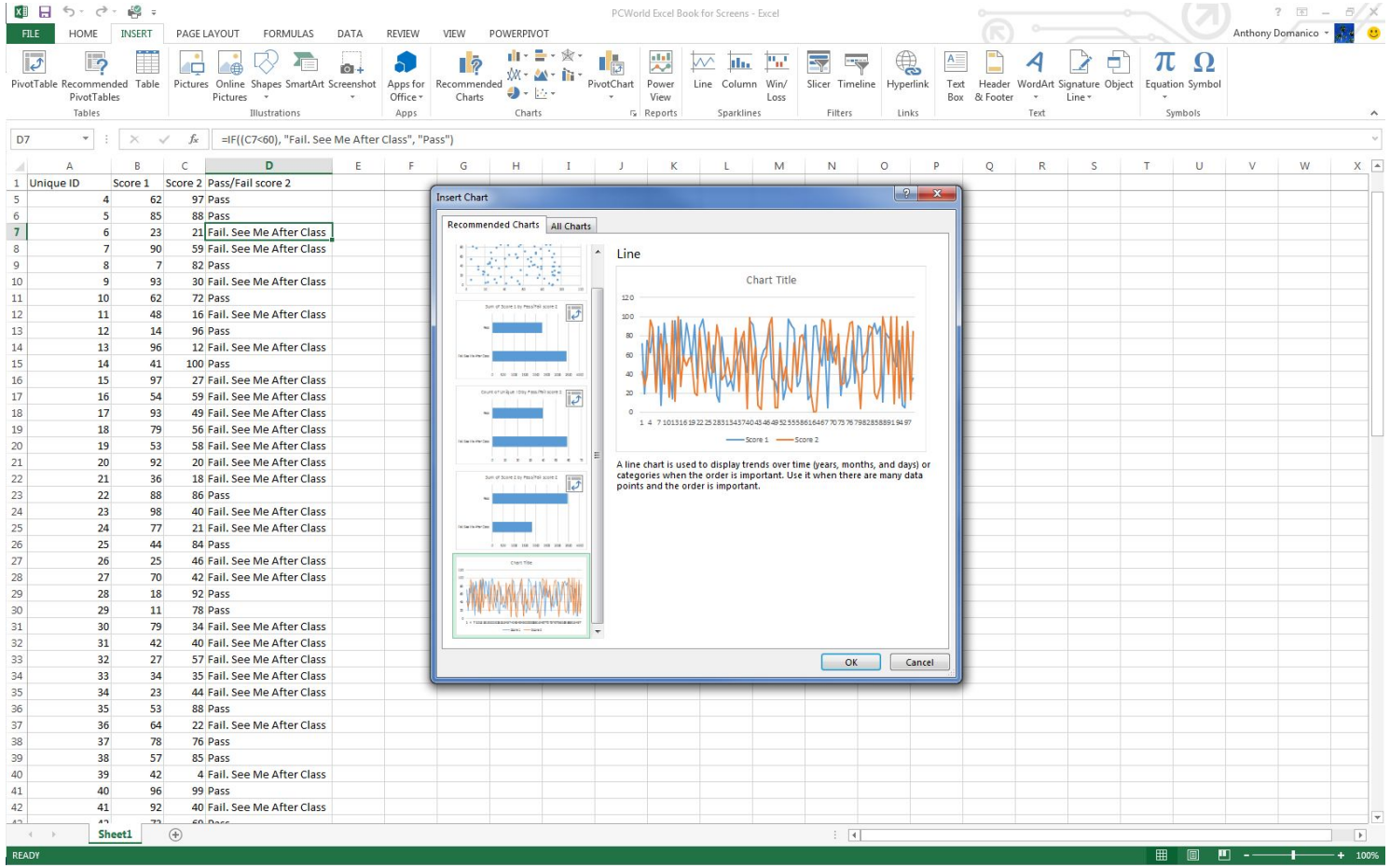
Polling question

What tools have you used to visualize data?

- Google charts
- Excell
- Tableau
- Python
- RStudio
- Power BI

Excel

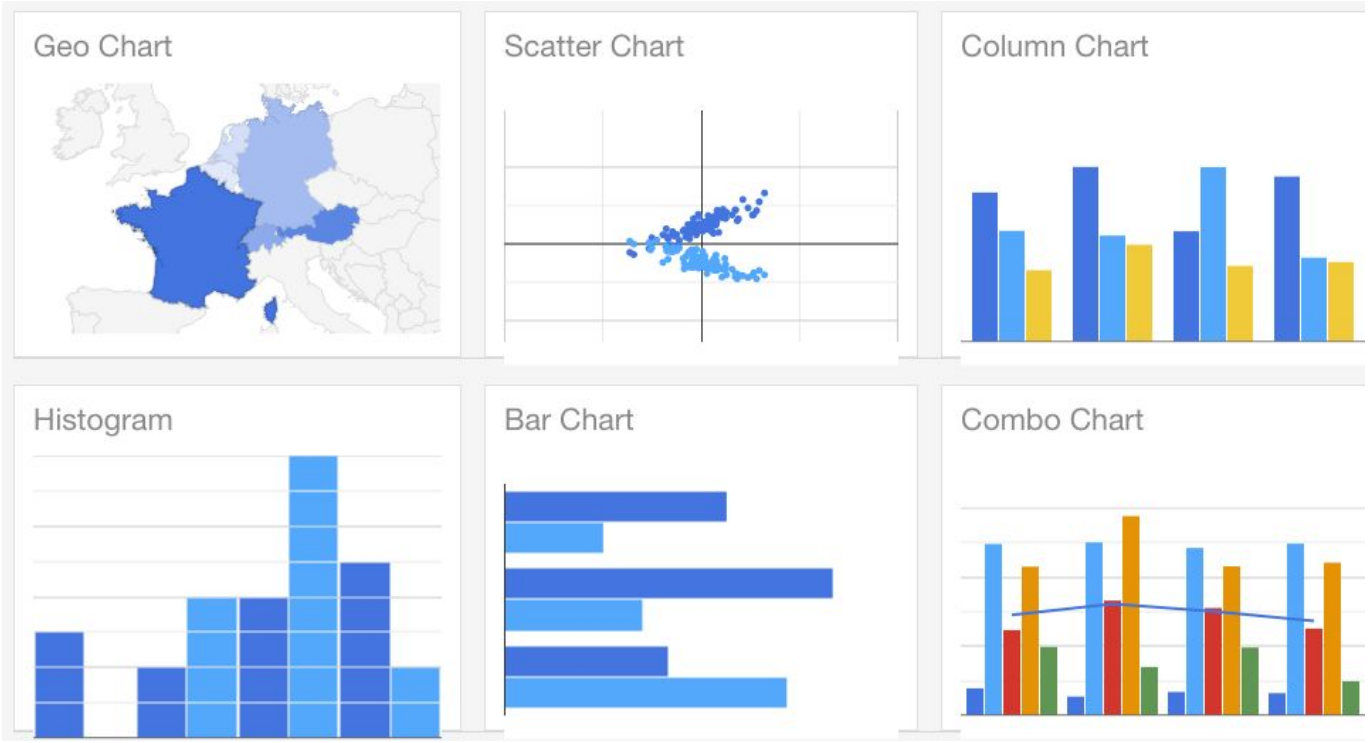
- Create basic chart types such as pie, line, bar, scatter, and more.
- Charts created in Excel can easily be ported to PowerPoint and Word.



Google Charts

- Free and open source, which includes a rich gallery, fully customizable, controls and dashboards, and HTML5

- Has more options than Excel; create interactive, animated and geospatial graphics

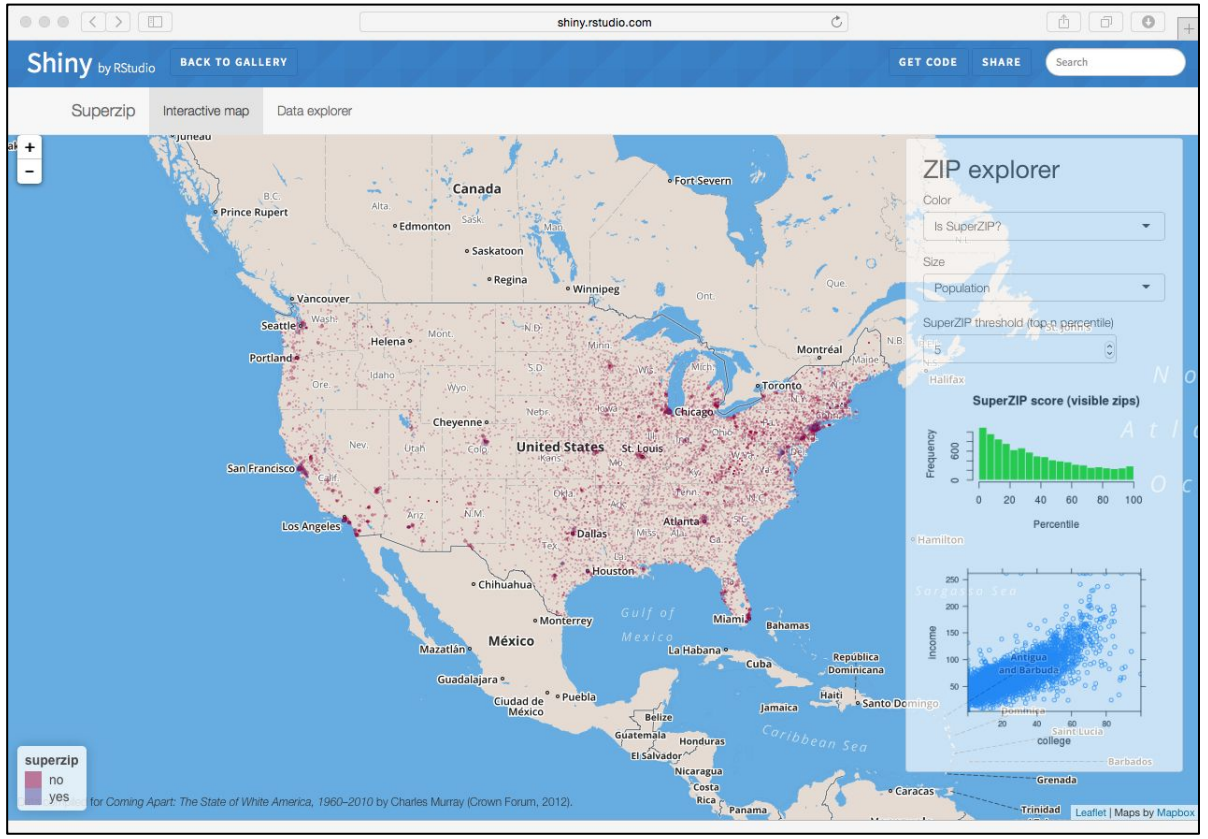


Tableau

- Tool for creating powerful and insightful visuals
- No programming required; drag and drop
- Share and collaborate on premise or in the cloud
- Platform can be used department or organization wide

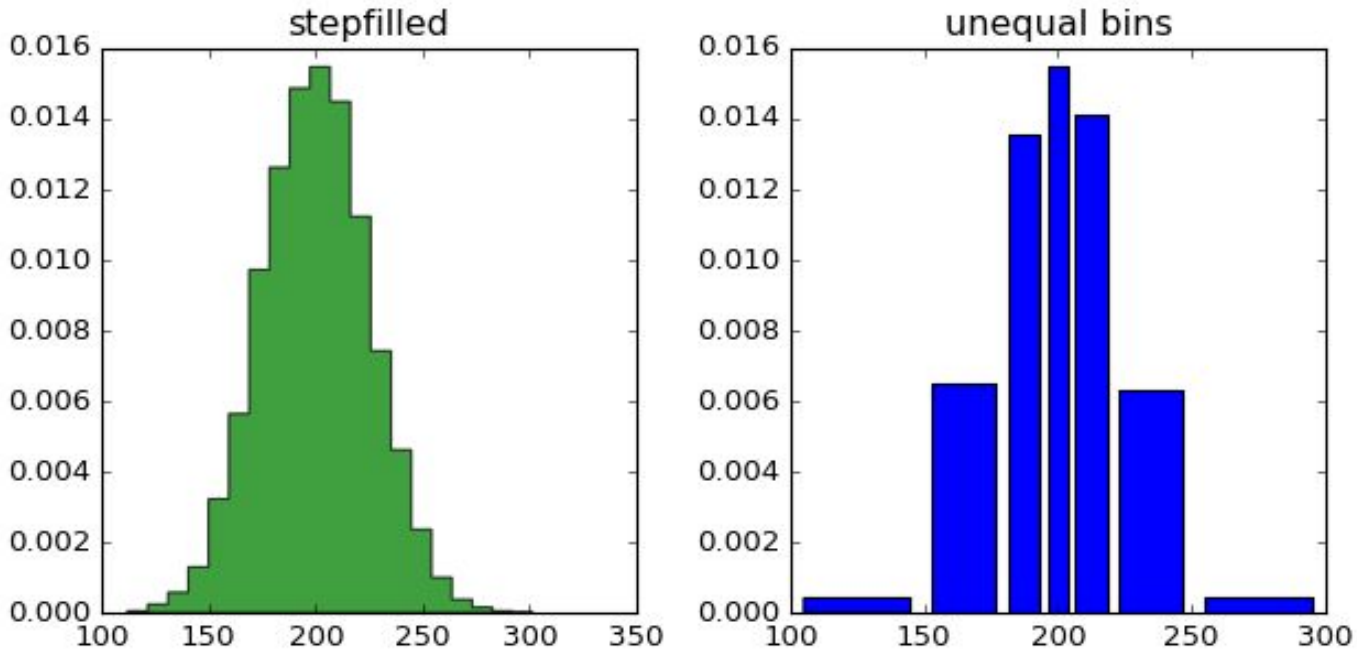
R and RStudio

- Programming tool
- Mainly used for statistical analysis
- Offers functions and libraries to build visualizations and present data
- Open source and free



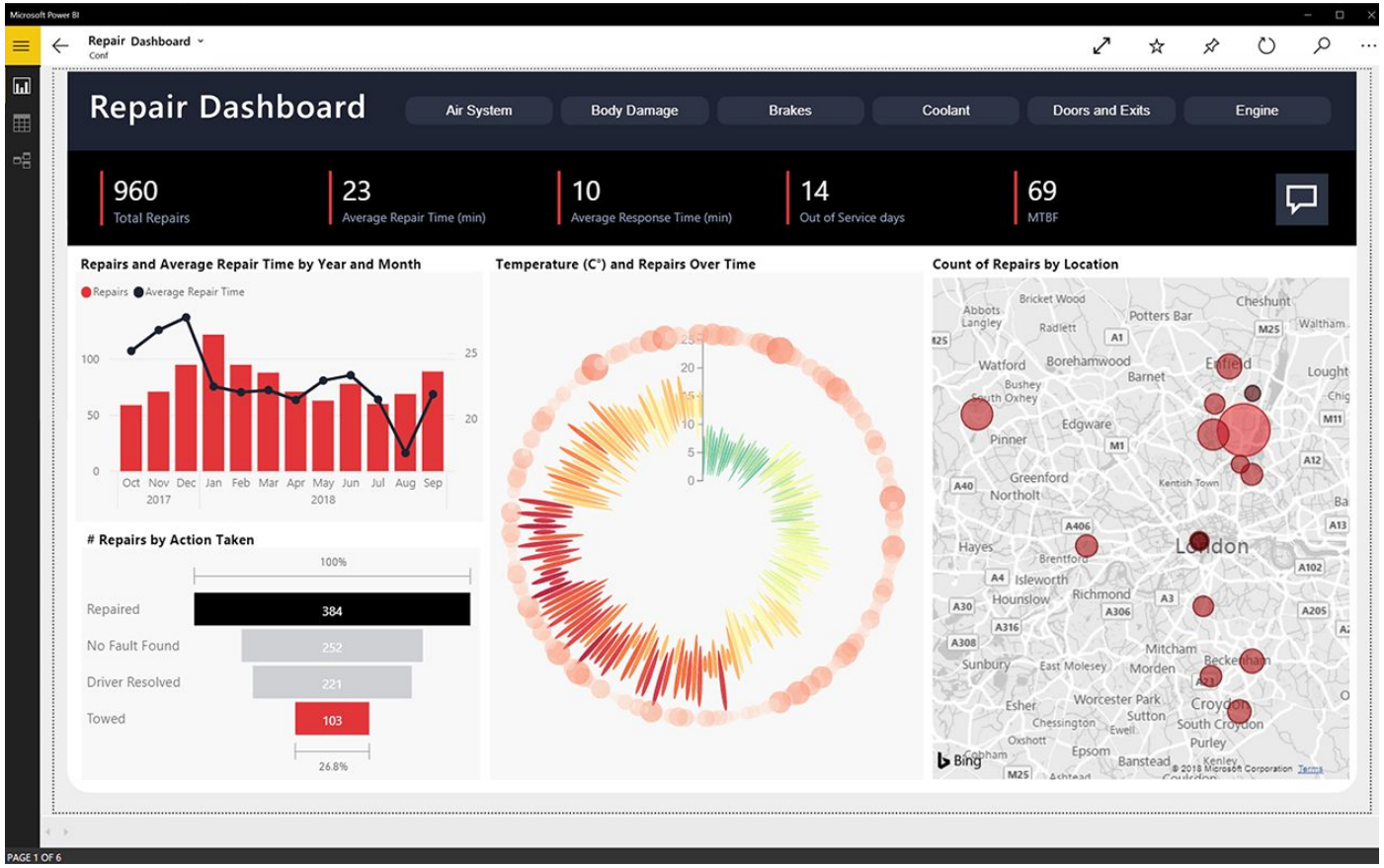
Python

- Programming tool
- You'll find libraries for practically every data visualization need
- Free and open source



Power BI

- Interactive visualizations and business intelligence capabilities
- Simple interface
- Create dashboards



Break

Agenda

Day 4

- Data visualization
- Misleading statistics & visual distortions
- Data storytelling

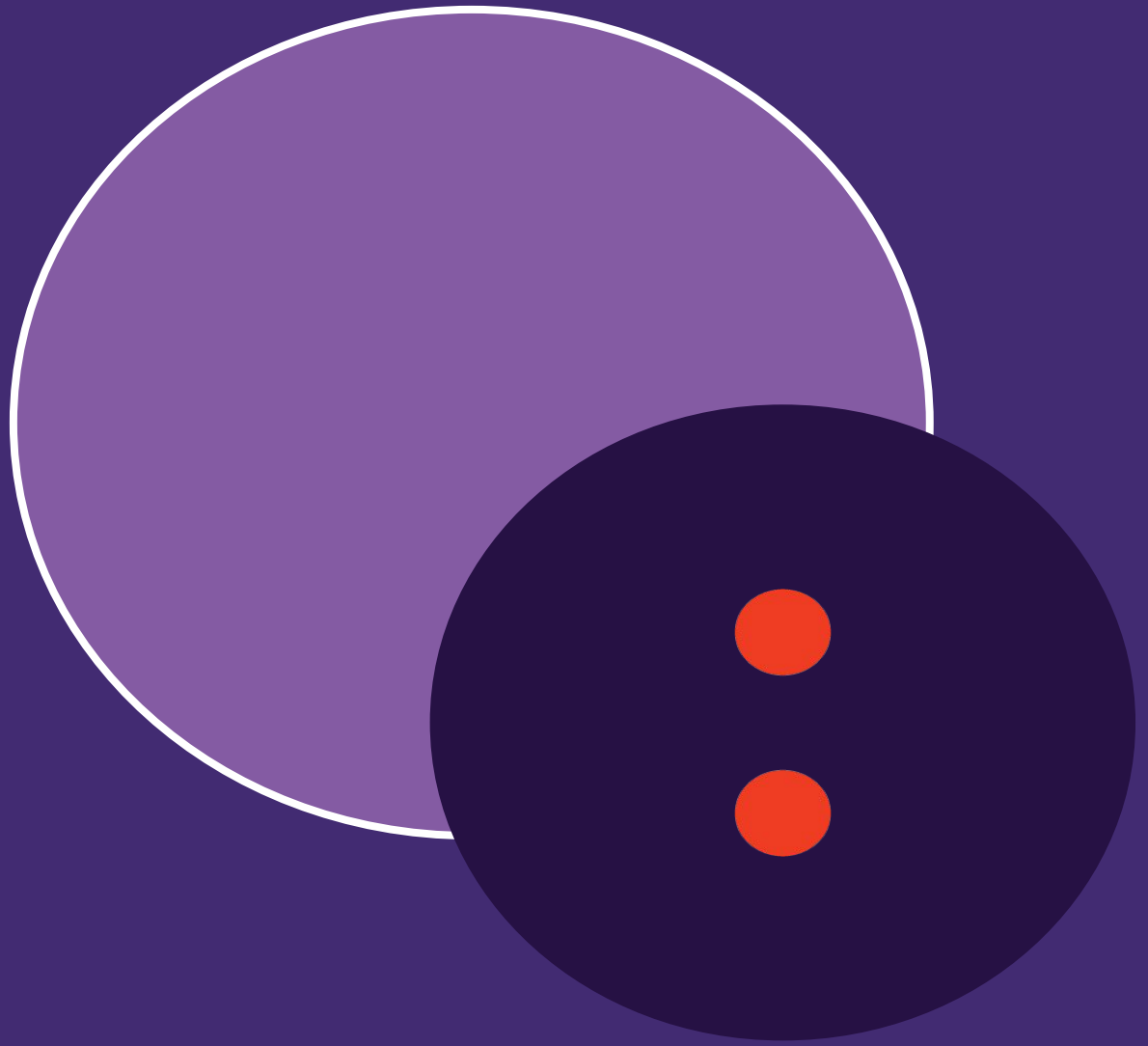
- What do I look out for when reviewing statistics or visualizations?

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.

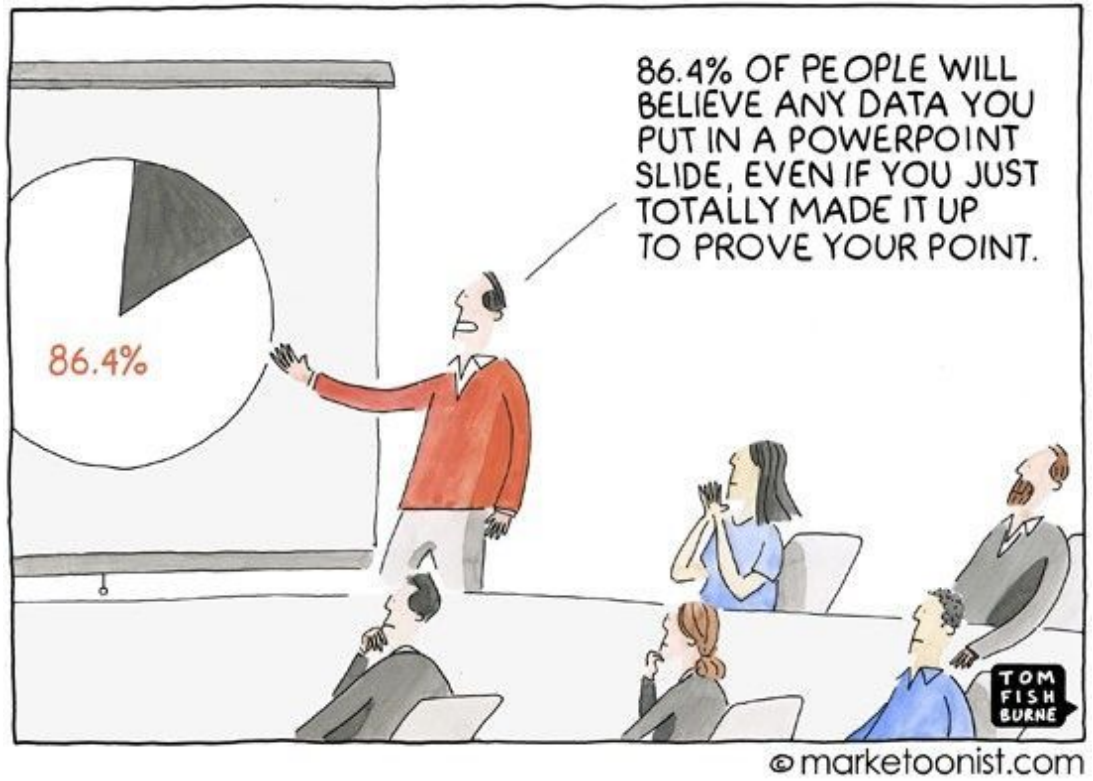
What's wrong with these statements?

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

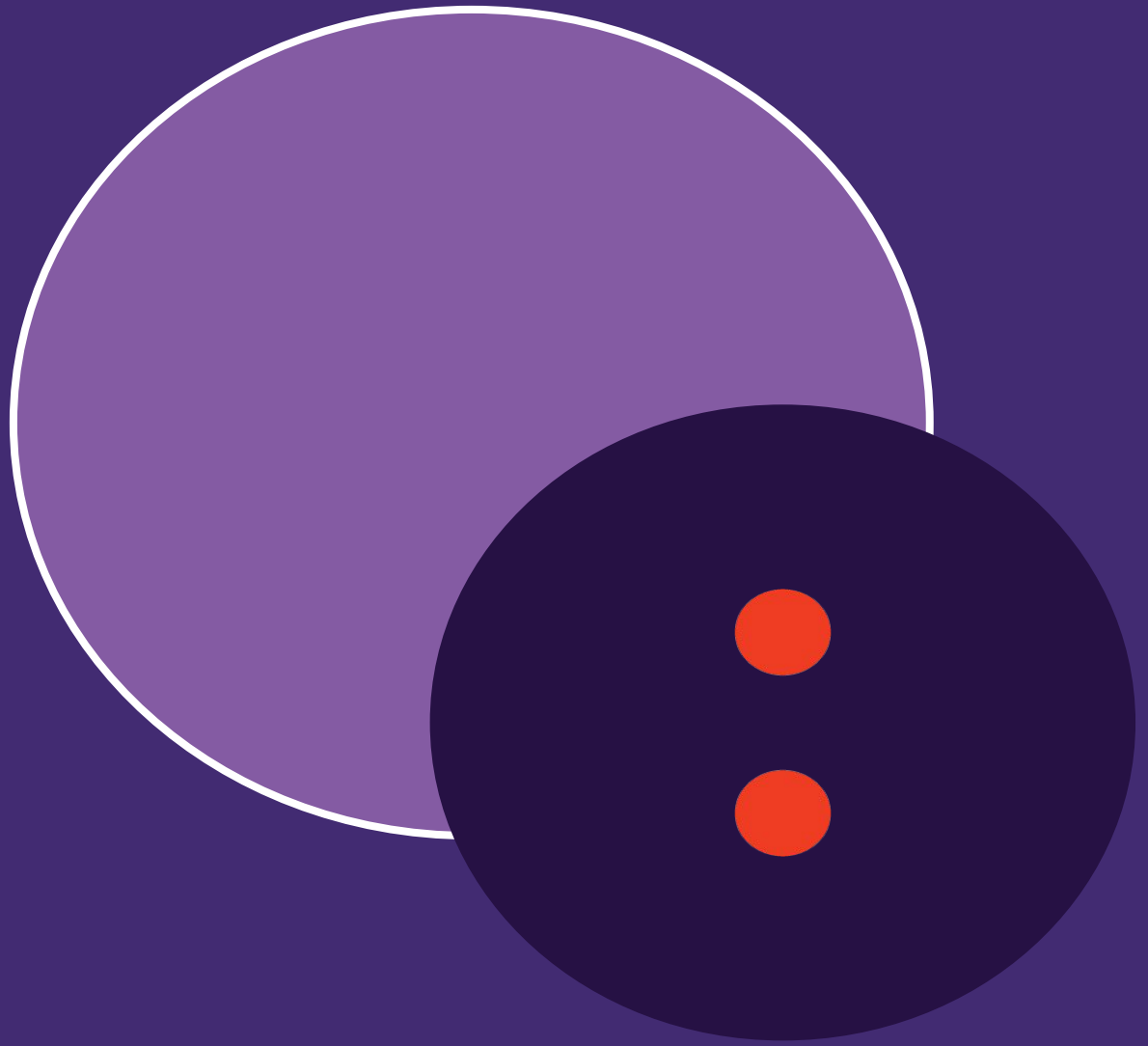
- Misleading statistics can be created through issues with:
  - data collection
  - data processing
  - data presentation

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?

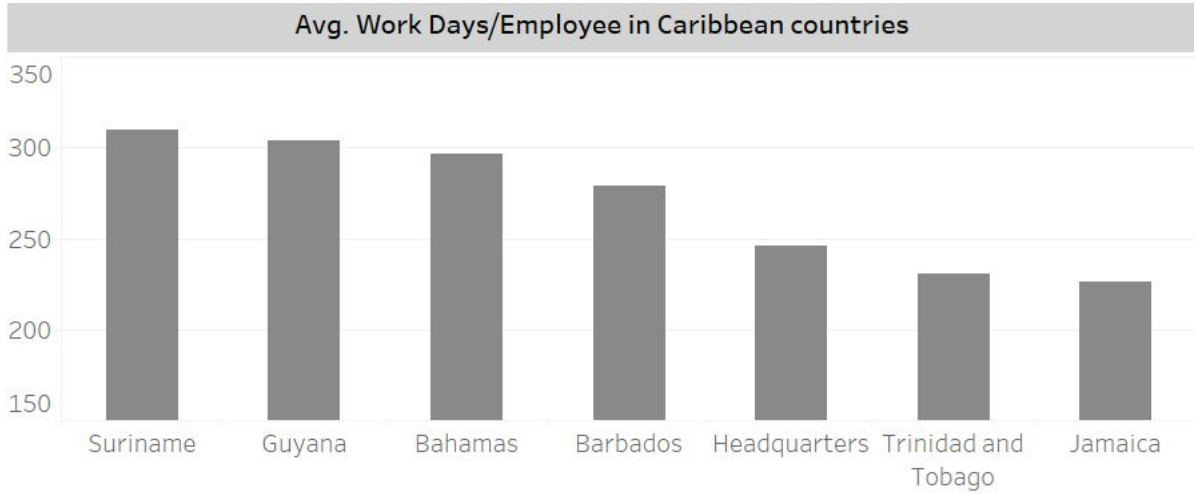
Visual distortions

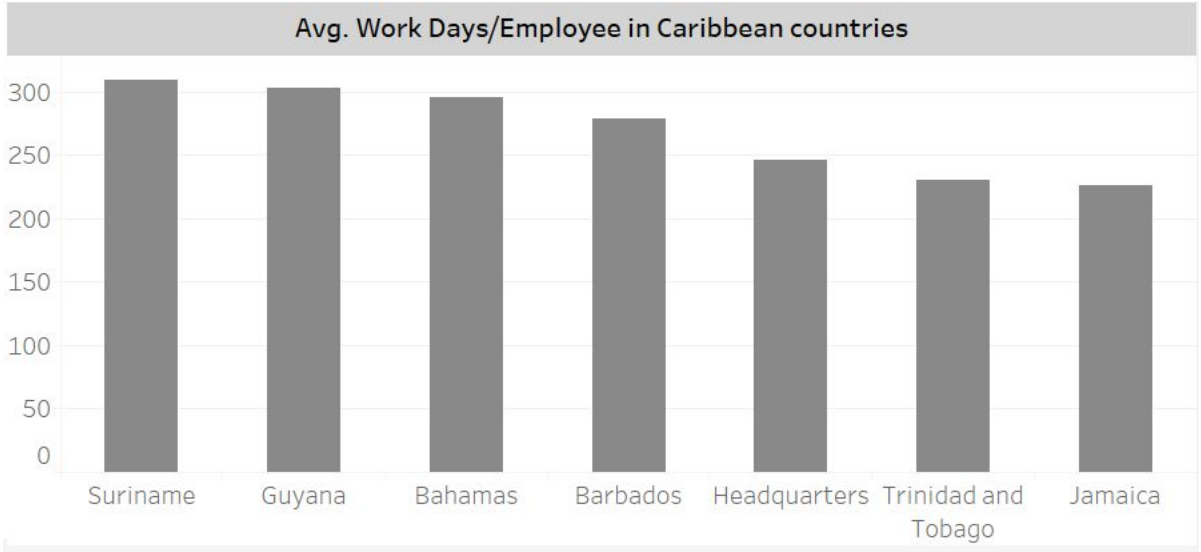


Visual distortions

- 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?

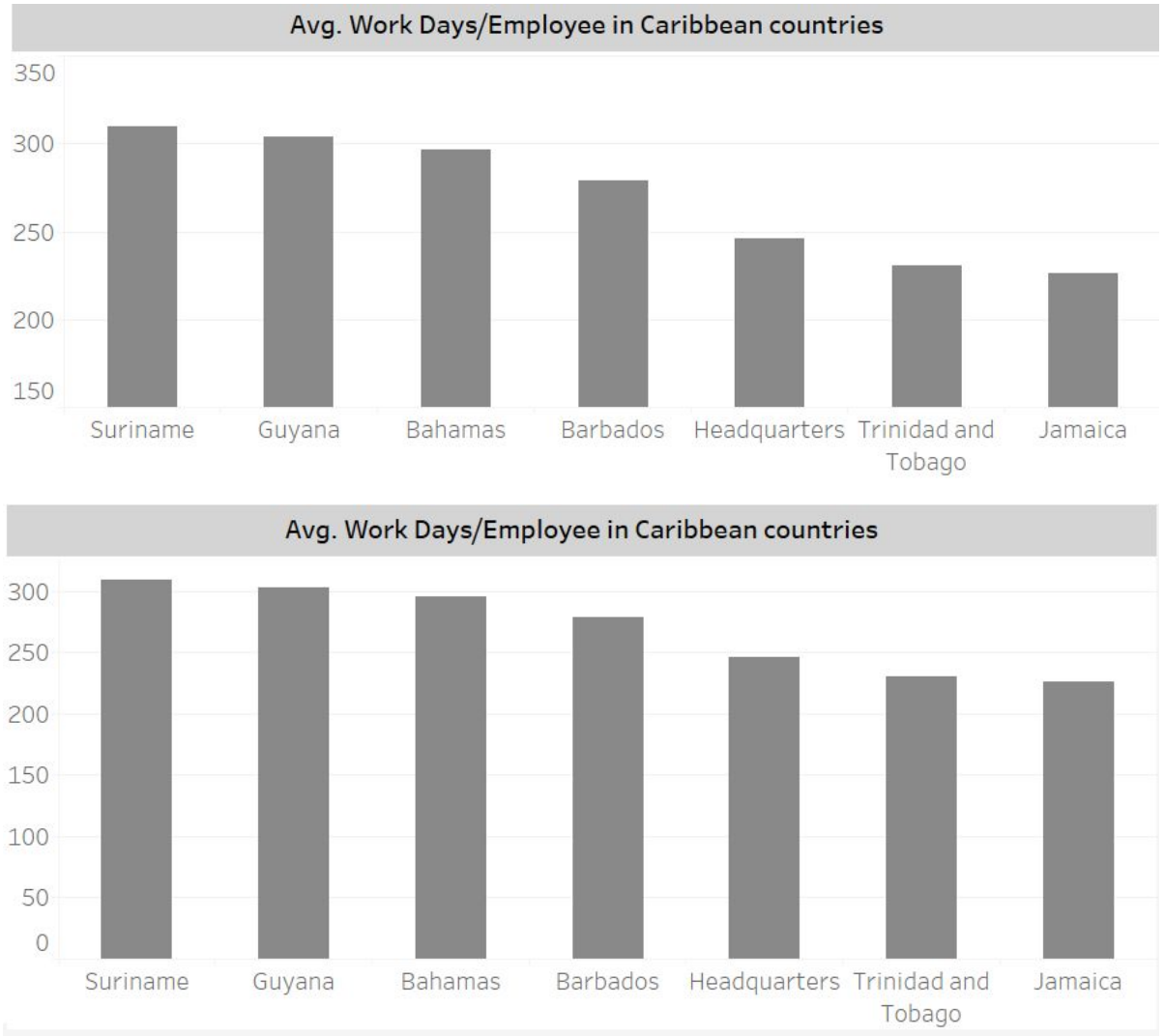




Truncated graphs

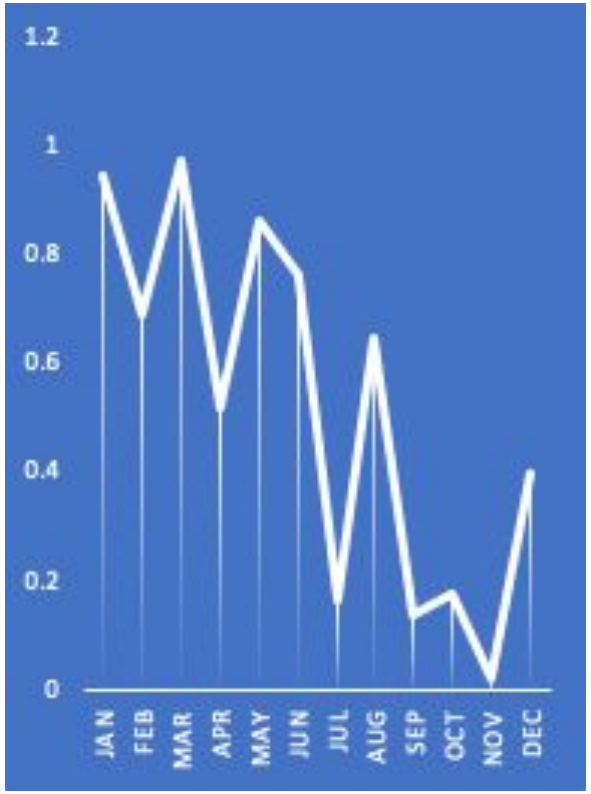
- One of the most common manipulations is omitting baselines or beginning the y-axis of a graph at an arbitrary number instead of 0.

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



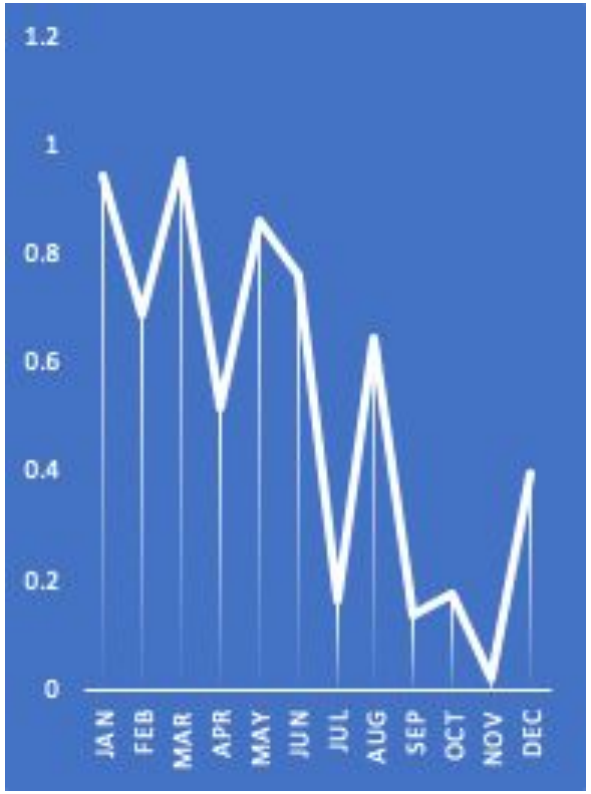
Visual distortions

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



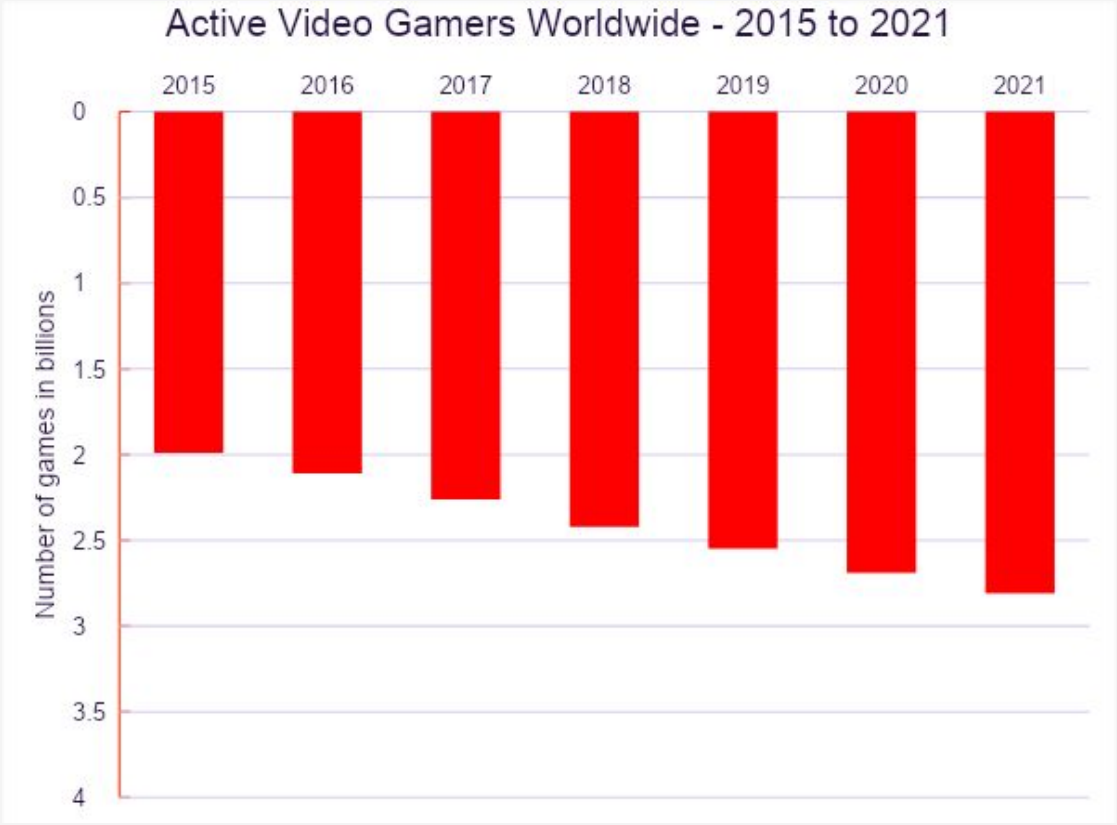
Exaggerated scaling

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



Visual distortion

How might this chart be misleading?

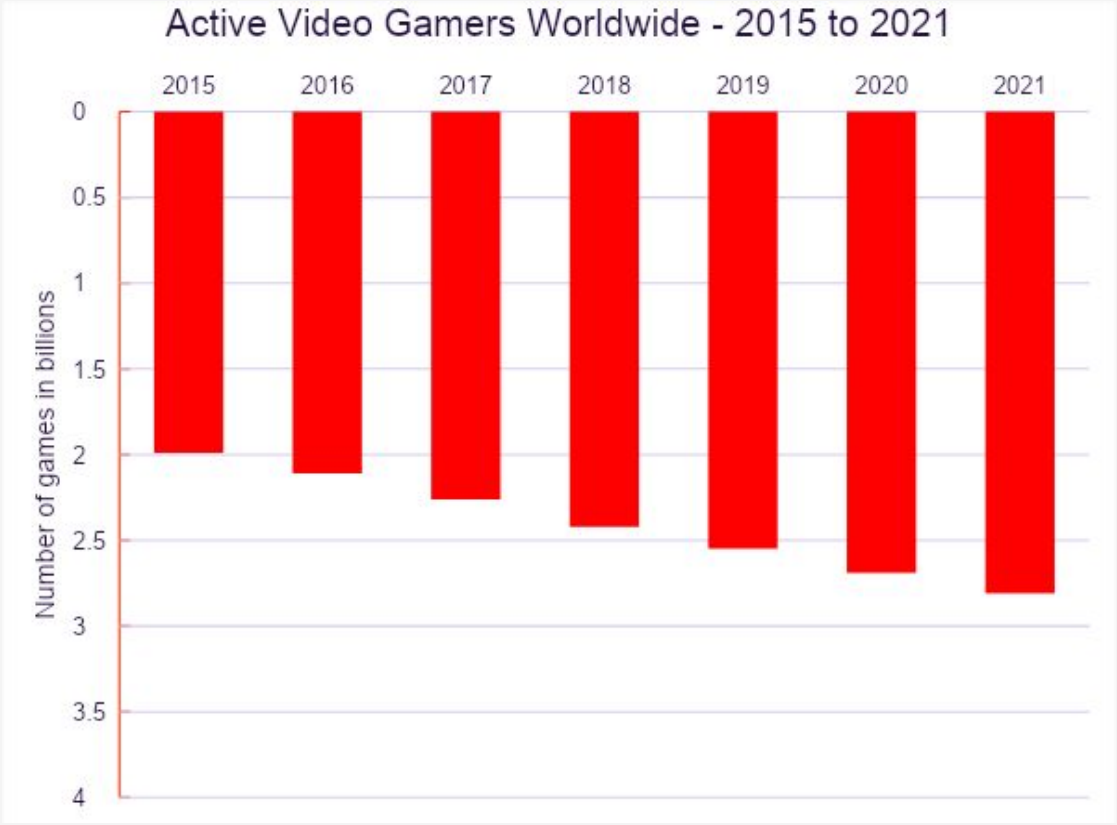


https://financesonline.com/number-of-gamers-worldwide/

Ignoring convention

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

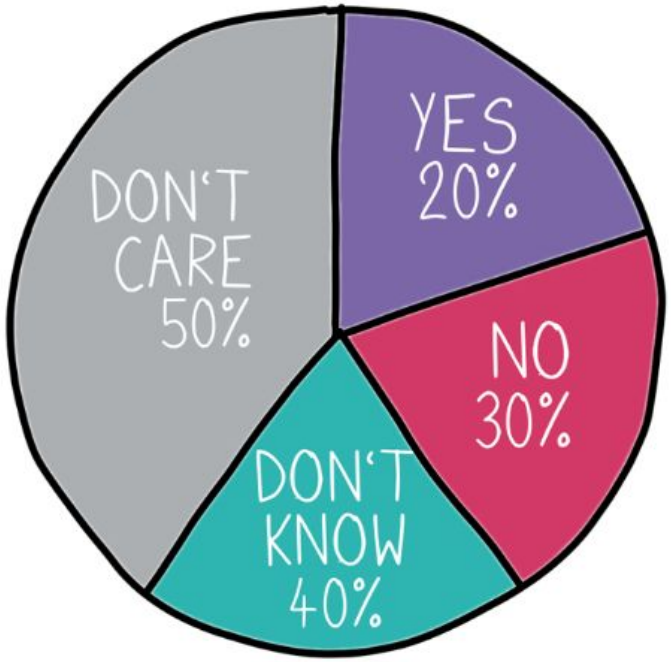
- In this example, the axis also moves downward, making an increase in gamers look like an decrease, at a quick glance.

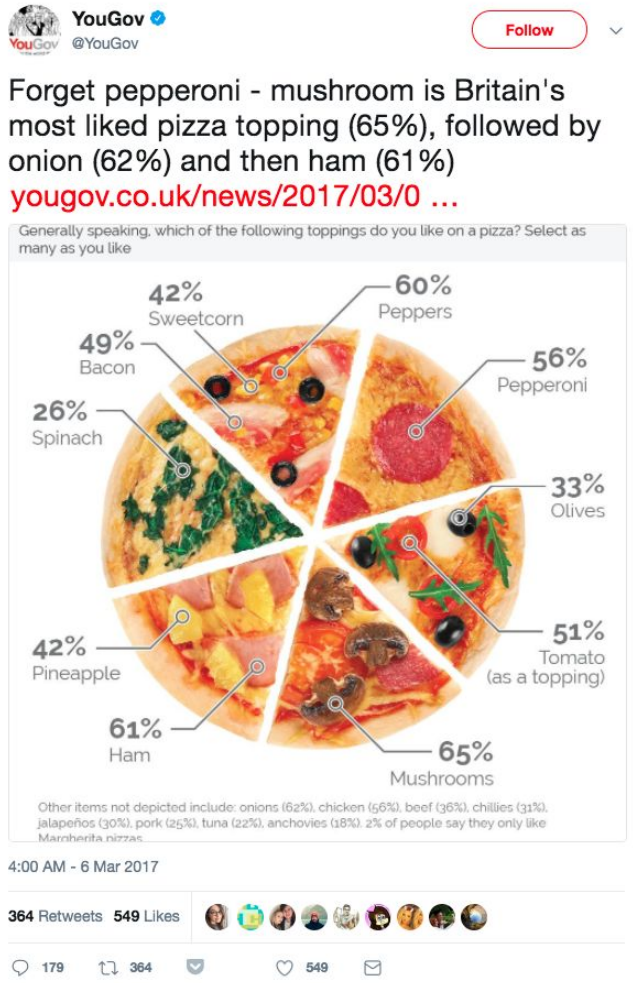


https://financesonline.com/number-of-gamers-worldwide/

Visual distortion

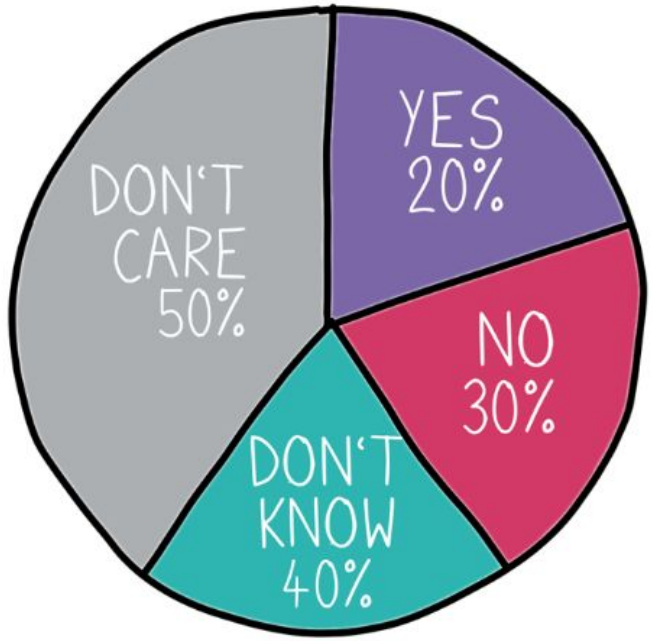
What do you notice about these pie charts?

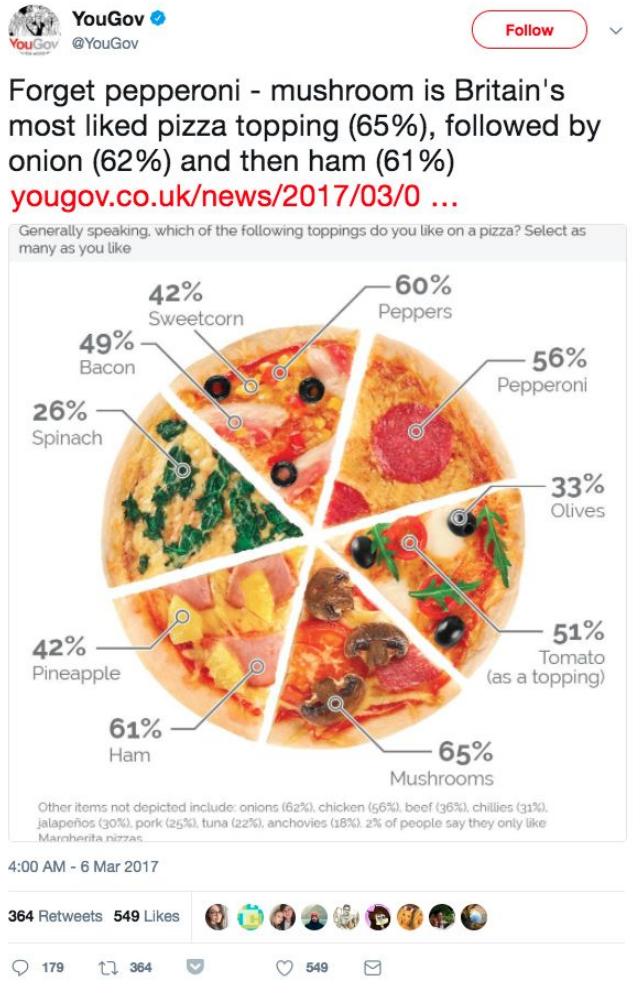




Numbers don’t add up

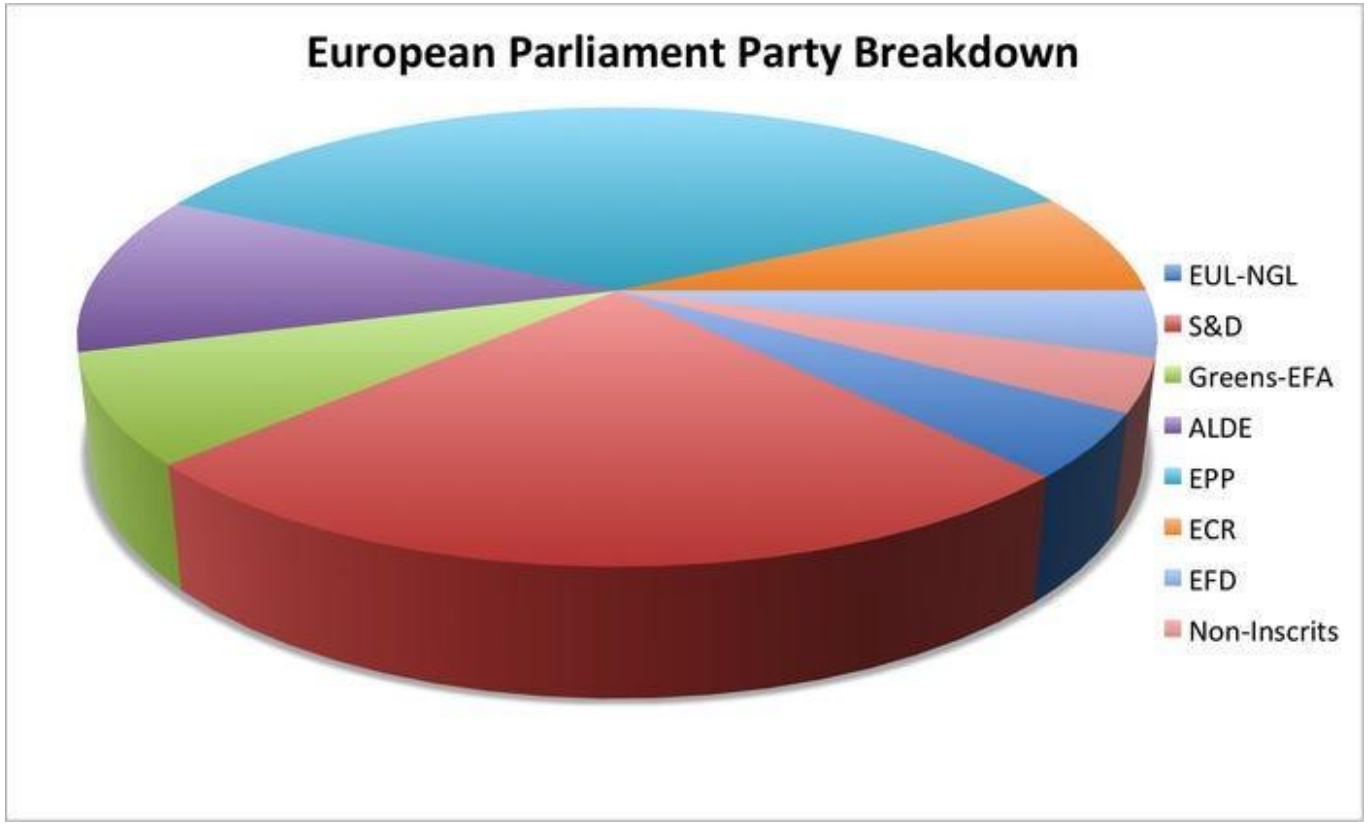
- With pie charts, the sum of each slice must add up to the whole. When the numbers don't add up, you know there's an issue.





Visual distortion

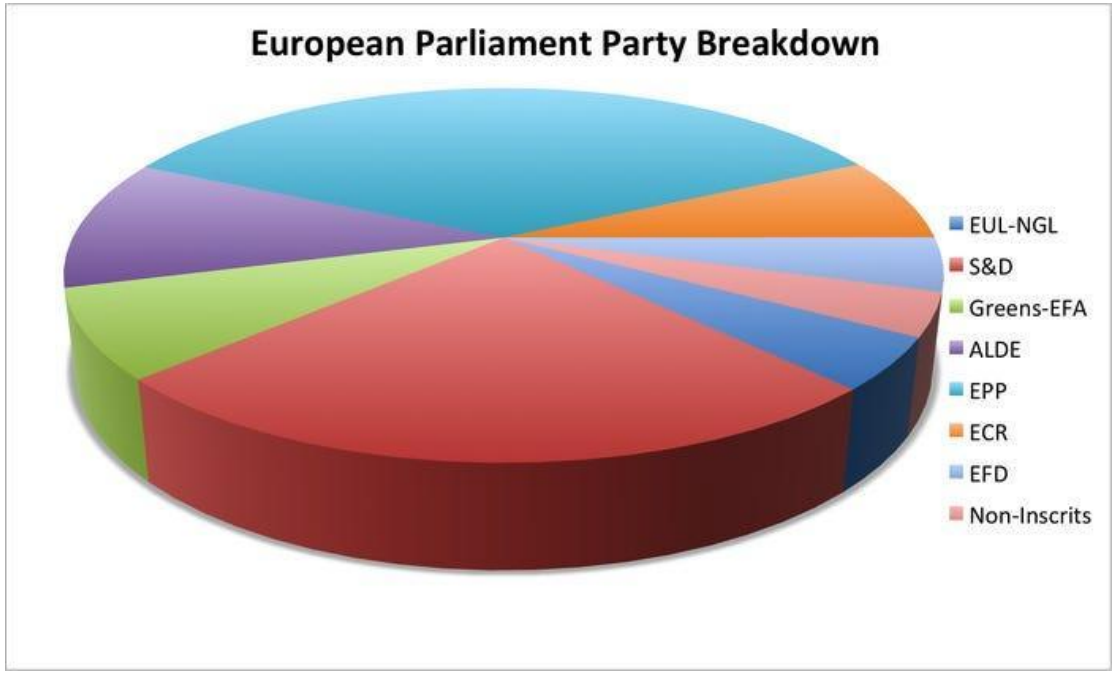
Does the S&D or the EPP party have more representation in parliament?

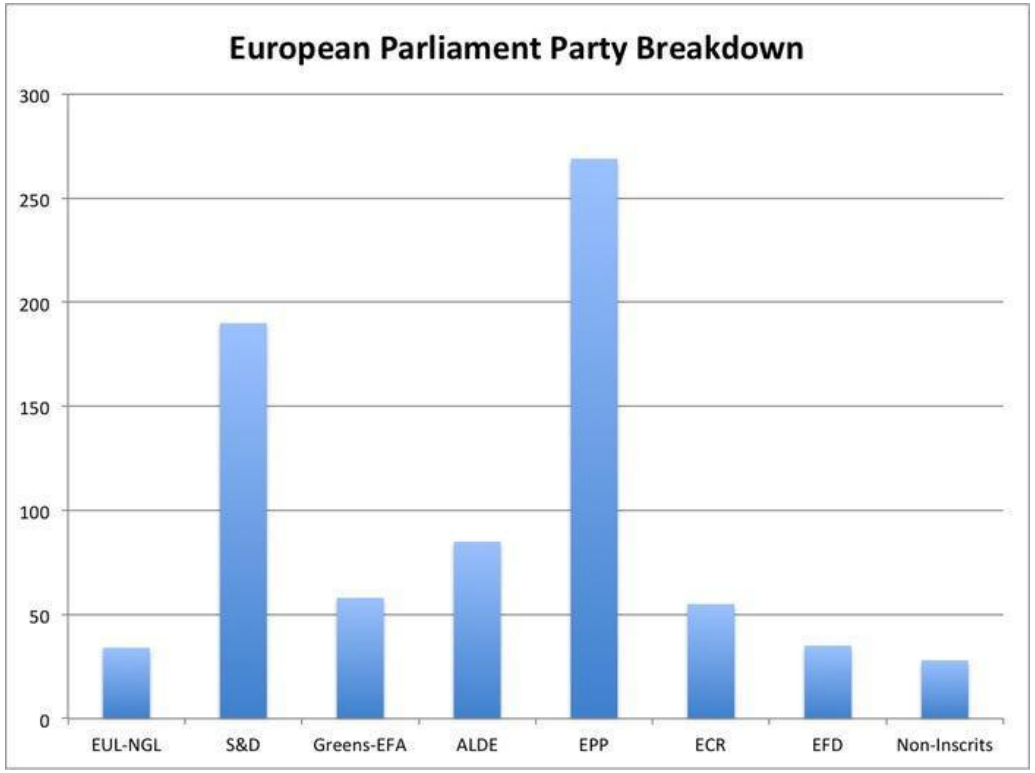


https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

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.

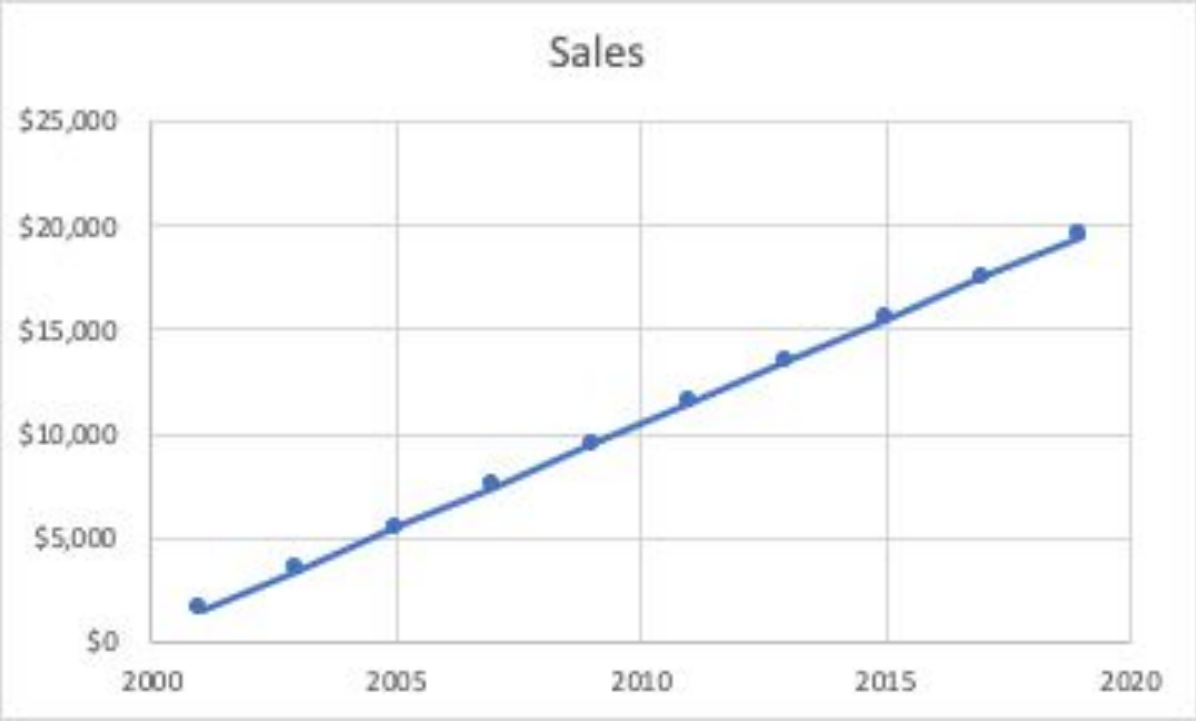


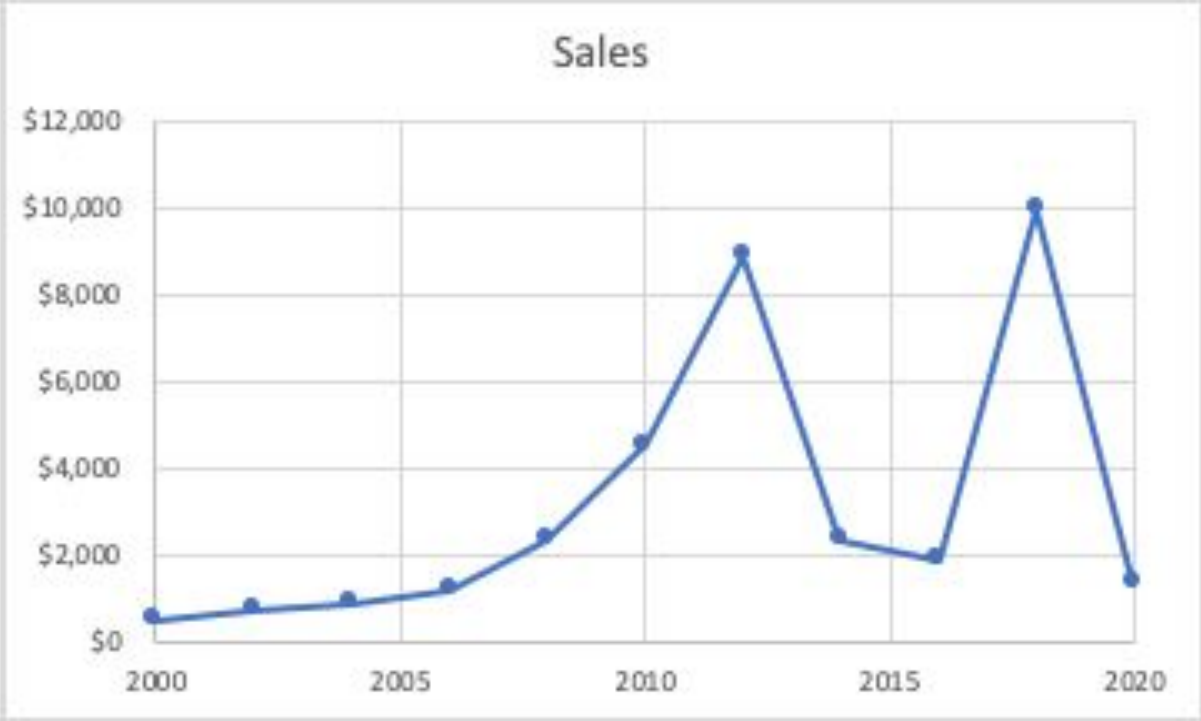


https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

Visual distortion

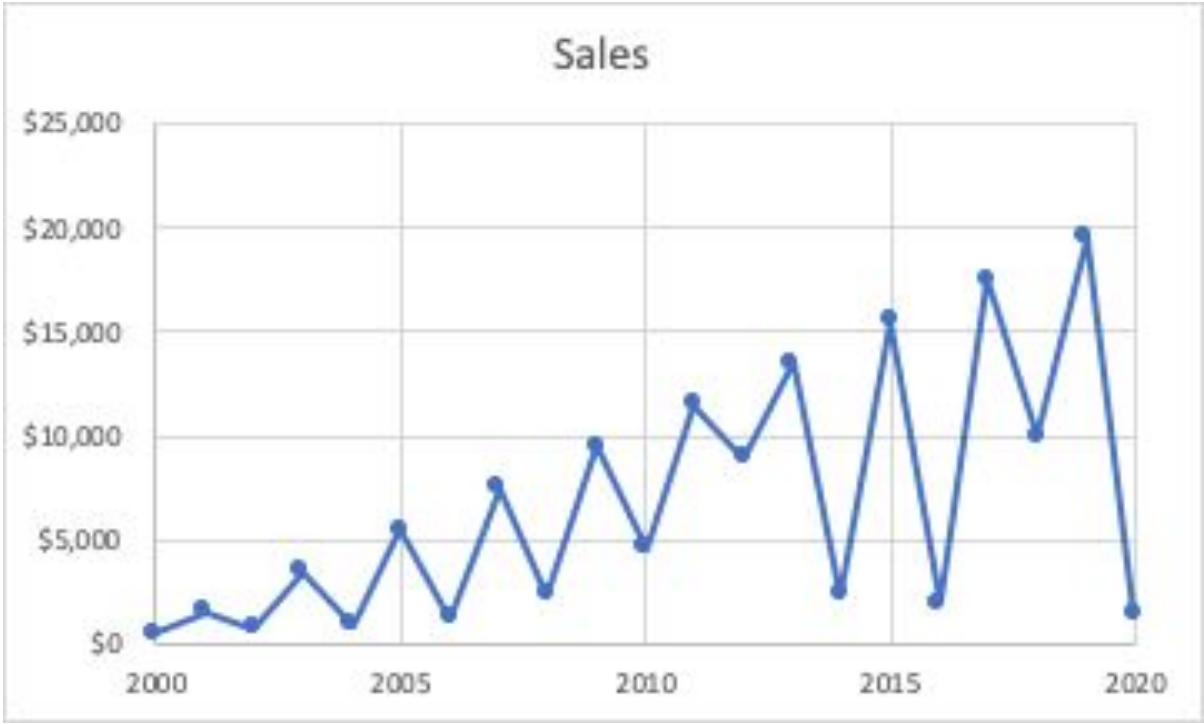
Which company has a better sales trajectory?



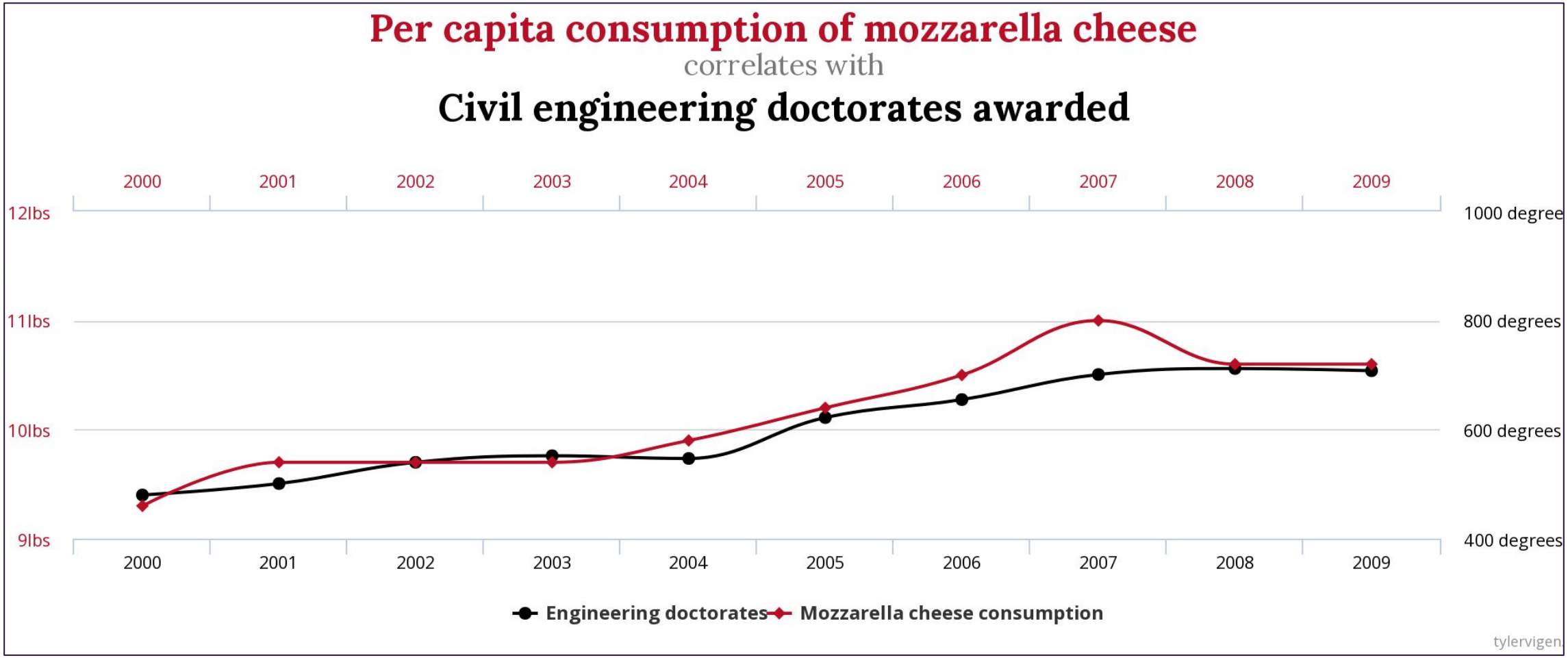


Improper extraction

- 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.



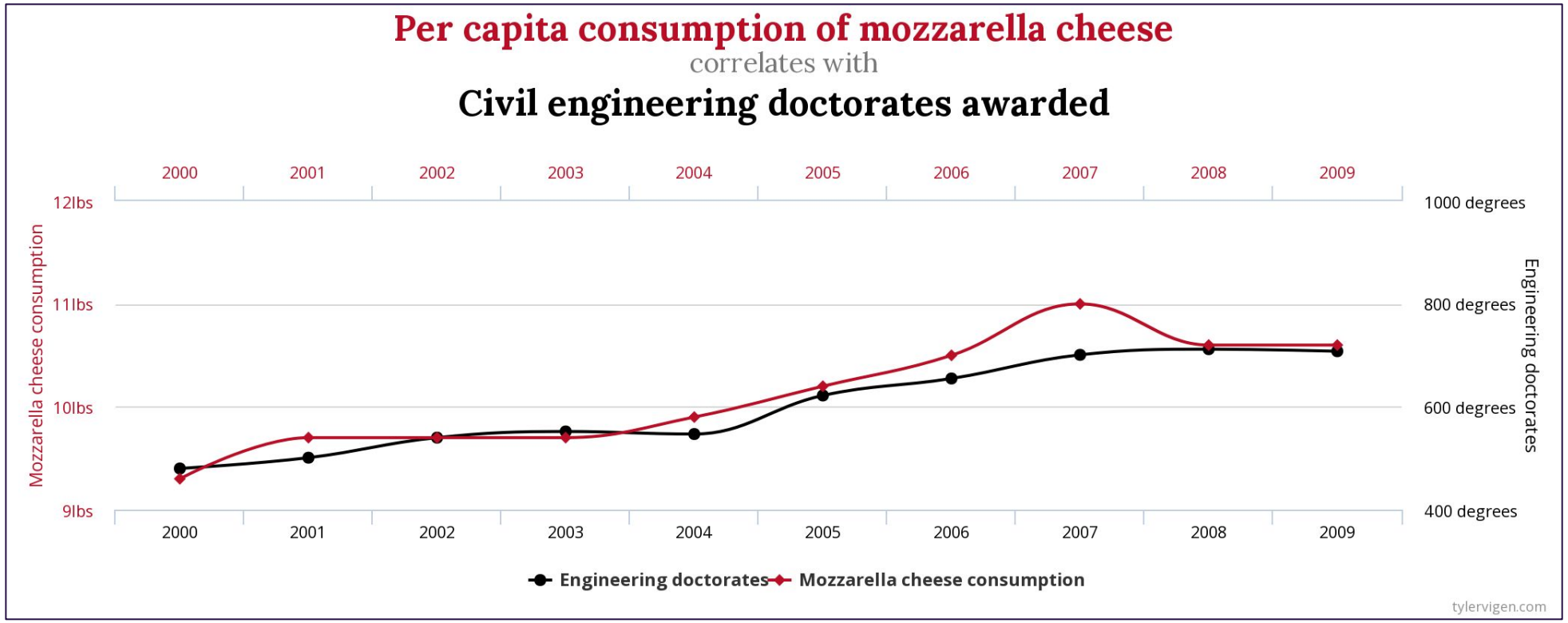
Visual distortion



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

Break

Agenda

Day 4

- Data visualization
- Misleading statistics & visual distortions
- Data storytelling

- Why are data stories useful?
- How do I craft a data story?

What is data storytelling?

- You focus on an insight and
- persuade an audience
- that the outcome of your analysis
- demands a course of action
- through narrative and visual communication.

Data stories and data visualizations

- A single data story may make use of multiple data visualizations.

- Data stories arrange visualizations into the linear sequence of storytelling: a beginning, a middle, and an end.

- Data story formats will likely incorporate other elements to explain and contextualize the visualizations:
  - prose text, either written or spoken
  - annotations, callouts, and labels
  - icons or graphics
  - images or photographs

Can't I just use a chart?

- Narratives are super effective, "sticky" content delivery mechanisms.
- Not everyone is a statistician, but they still want to make evidence-based decisions.
- Stories let you overview key findings quickly.
- Stories tap into both the logical and the emotional aspects of persuasion.

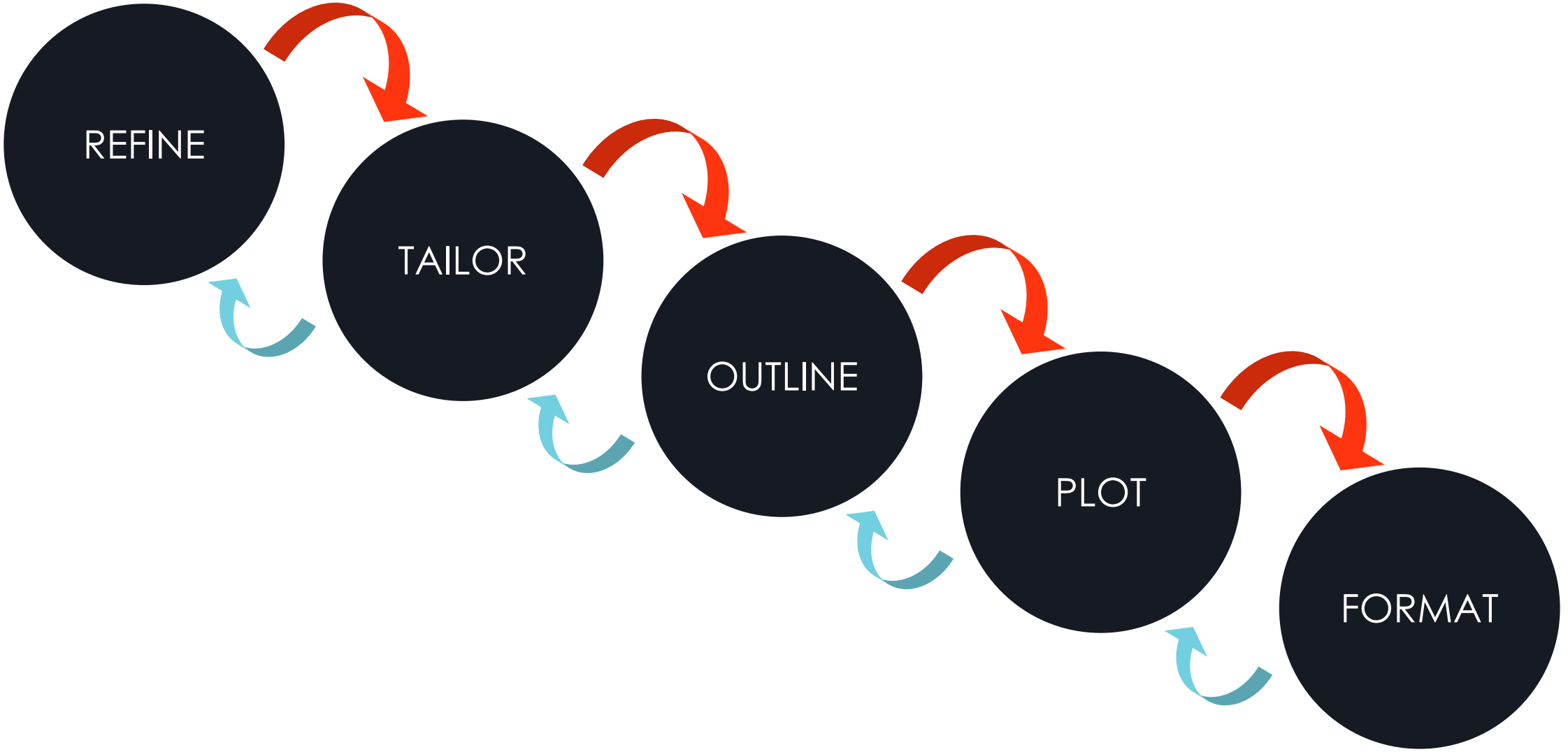
Why choose story?

If your insight is...	A story can...
Unpleasant	Help convince your audience that even unwanted results are actionable.
Disruptive	Encourage your audience to break with tradition, if the upshot is valuable enough.
Unexpected	Explain why a prediction or intuition failed, and offer some analysis and a solution.

Why choose story?, cont'd

If your insight is...	A story can...
Complex	Guide your audience to a more complete understanding in manageable chunks.
Risky	Embolden your audience to take responsibility for making a tough choice.
Costly	Compel your audience to consider a high-cost solution by underscoring the high value.

How do I craft a data story?



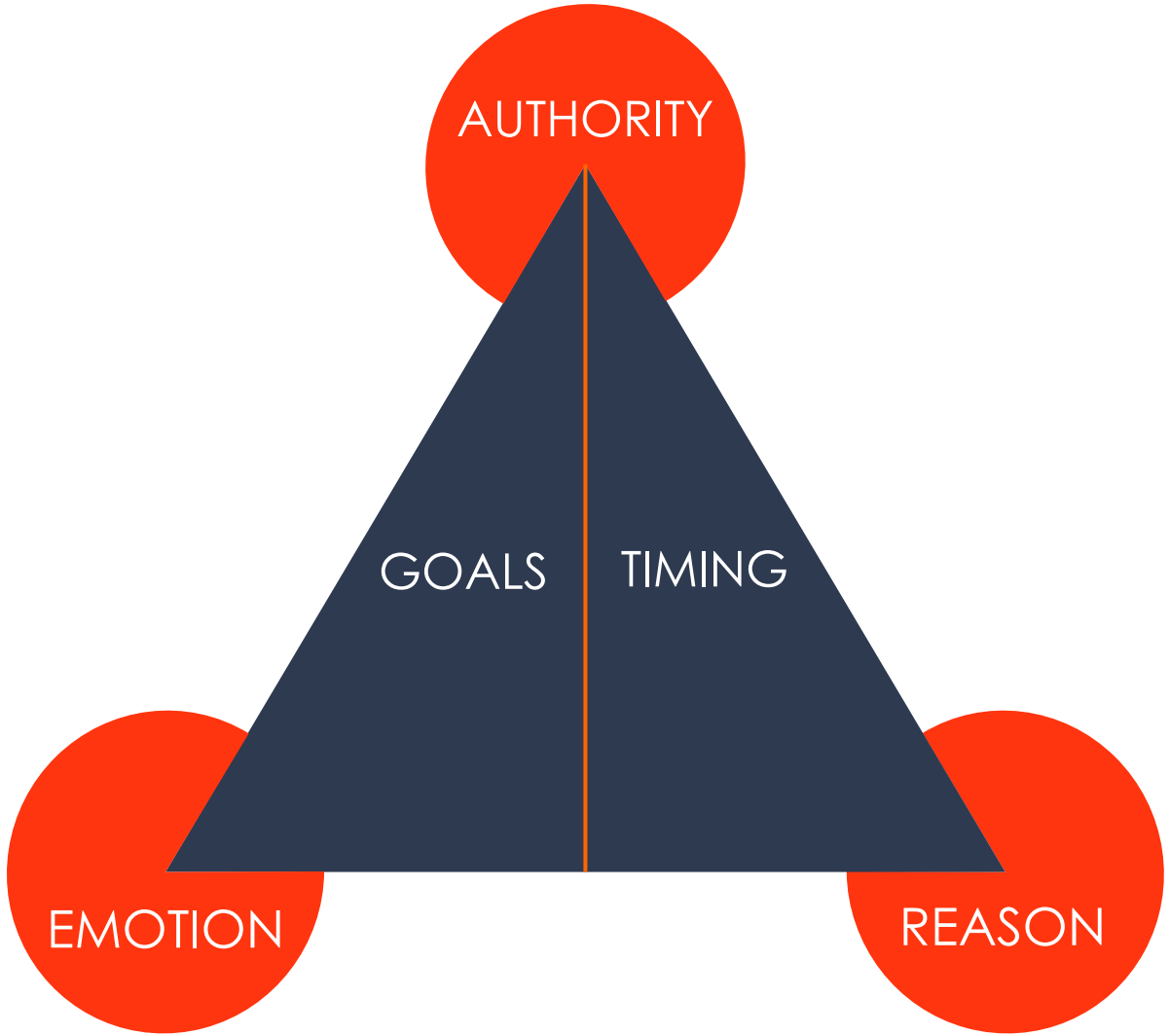
REFINE

Refining your insight

- In a data story, your insight is the most important piece.
- What will make your audience perceive your insight as maximally:
  - Valuable: an observation that seems to be rewarding
  - Relevant: an observation that seems timely
  - Practical: an observation that suggests a realistic and feasible course of action
  - Specific: an observation that clearly and completely accounts for a problem
- Make your insight as concrete and contextualized as possible

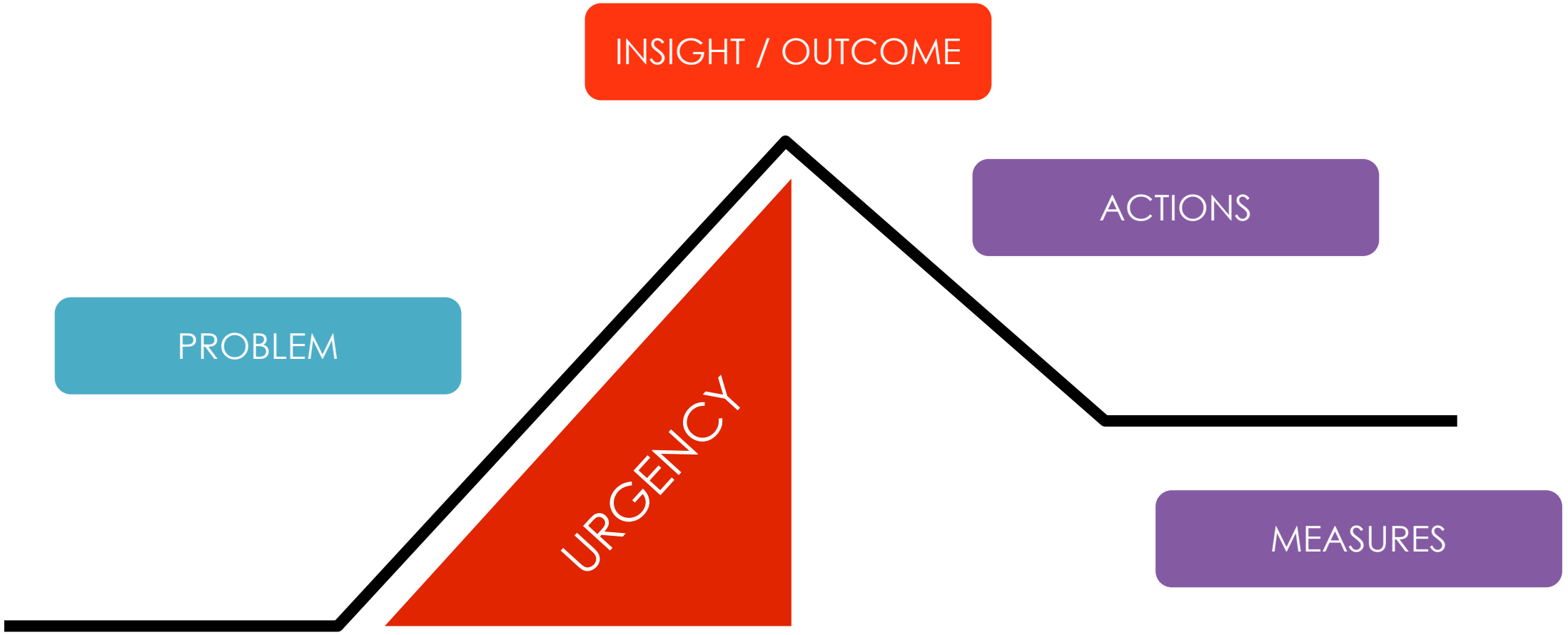
TAILOR

Tailoring to your audience



OUTLINE

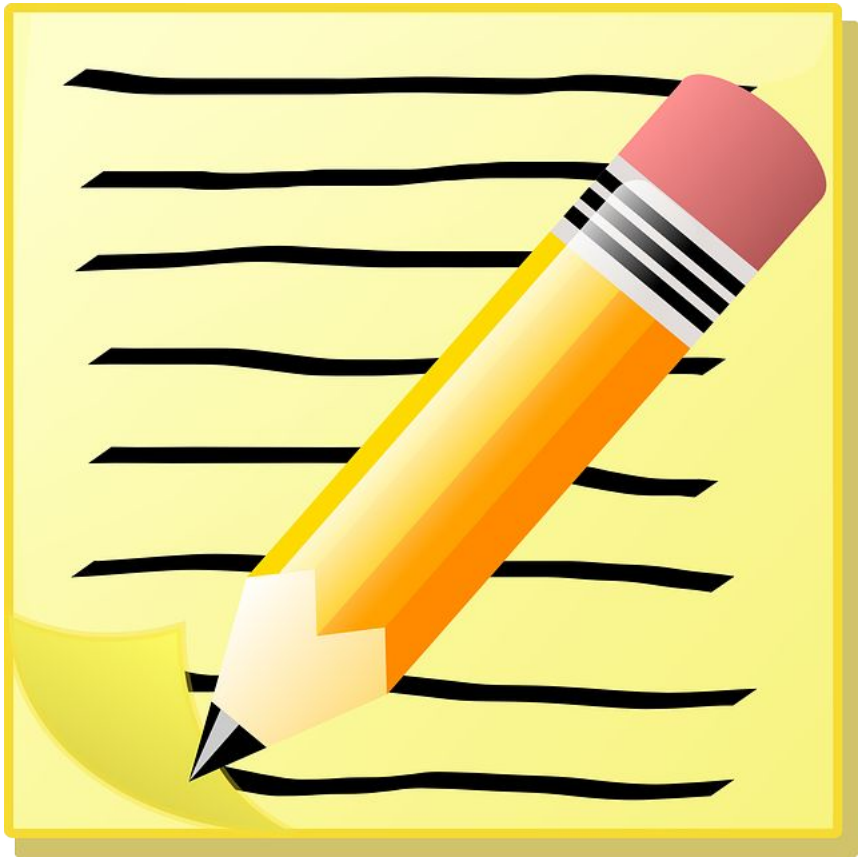
Outlining



PLOT

Plotting with a storyboard

- It's okay for your data story to remain flexible at this early stage.
- There are no right answers, only consideration and iteration.
- Focus on building the elements of the story first, on paper.
- Try out different versions quickly and don't get too attached.



FORMAT

Formatting for delivery

- You may find yourself needing to alter the way you tell your data story based on the affordances of the format.
- Sometimes the format is a given, but other times, it will depend upon your input and the use case.
- As with visualizations, the simplest storytelling format is often the best.

Slide Deck	Document	Interactive	Hybrid
Sequence of slides intended for real-time presentation	Illustrated text (report, infographic) to be read anytime	Digital object intended to align function with user experience	Blend / compromise of at least two formats

The Joy of Stats

Chat questions

- What data storytelling elements did you notice?

- Was this a good example of data visualization and storytelling? Why or why not?

End of Day 4

Data visualization
Misleading statistics & visual distortions
Data storytelling

DATASOCIETY:

Thank you and congratulations!