DATASOCIETY:

Fundamentals of Data Literacy

Finding the chat box

In the chat box, tell everyone how you take your coffee.

- **Zoom Client / Browser:** From the toolbar (probably on the bottom of your screen), select the button marked "Chat." The chat box should appear.
- **Mobile Devices:** The "Chat" button may be hiding under the "More" menu.

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
  - Q and A during the last 30 minutes of class

Best practices for virtual learning

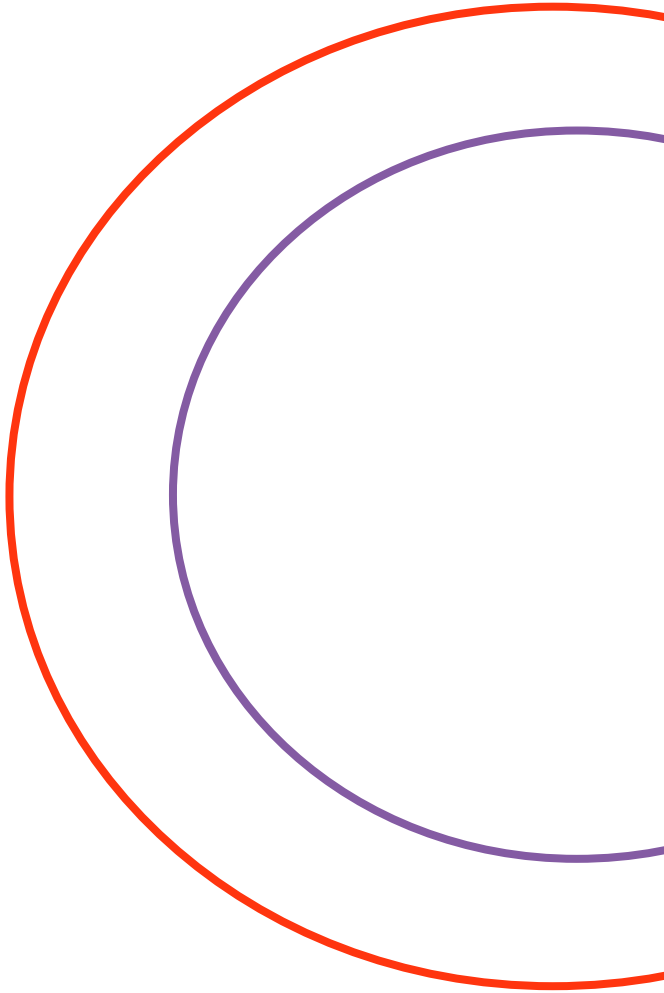
1. Find a quiet place, free of as many distractions as possible. Headphones are recommended.

2. Stay on mute unless you are speaking.

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

4. Participate in activities and ask questions. This will be interactive!

5. 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 discussion prompts, activities, a data science glossary, information about popular data science tools, and more!

Chat: introductions

1. Your name
2. What you hope to get out of the class
3. What you rate your current data literacy level on a scale of 1-10

1	5	10
No knowledge or awareness; may read articles or documents that contain percentages	Has combined data from multiple sources; has made basic charts/graphs; understands data limitations	Works with Big Data; has deployed machine learning

Agenda

- **Day 1**
  - The benefits of data
  - Data analytics overview
  - Data governance
  - Data tools
  - Data teams
- **Day 2**
  - Data-driven cultures
  - The data science process
  - Putting together a project
  - Foundational data science methods
- **Day 3**
  - Clustering problems
  - Classification problems
  - Regression problems
  - Working with text data
  - Working with network data
- **Day 4**
  - Neural networks
  - Refining a data project
  - Intro to data visualization
  - Best practices in data viz

Agenda

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

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

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.

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

Internal

- Information Resources Management (IT)
- Global Talent Management systems(HR)
- Comptroller and Global Financial Services systems (Finance)
- Supply chain records
  - e.g., ILMS, Ariba, FleetWave
- Service records
  - e.g., myServices

External

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

…and more!

Chat question

What data sources do you interact with most regularly?

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

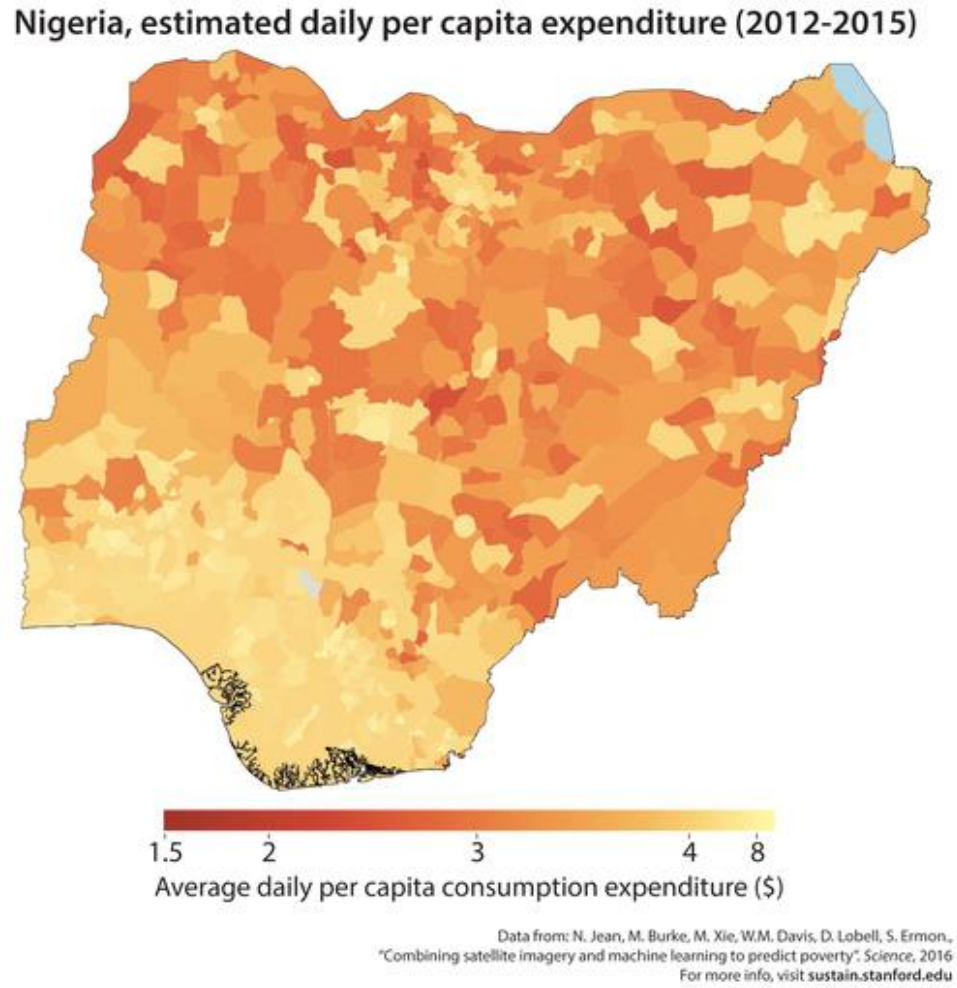
6. Enable experiments

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 Ti 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.

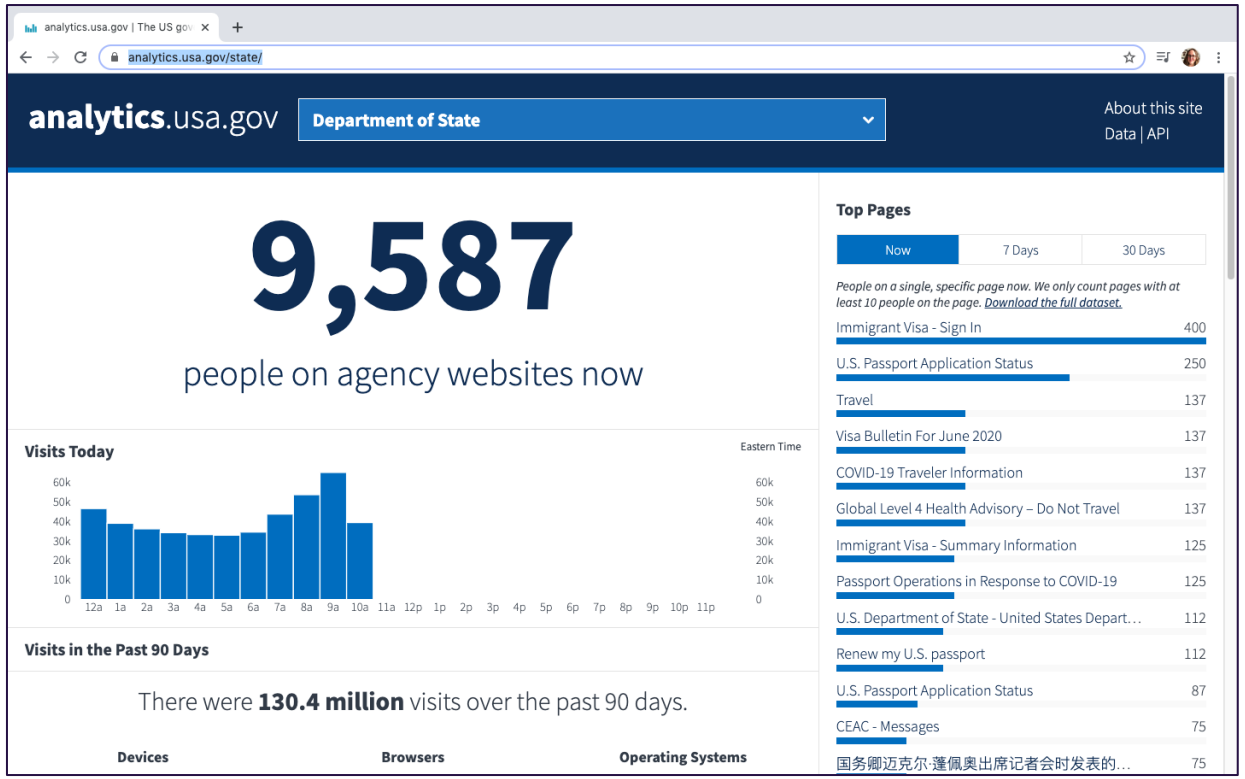
OLD

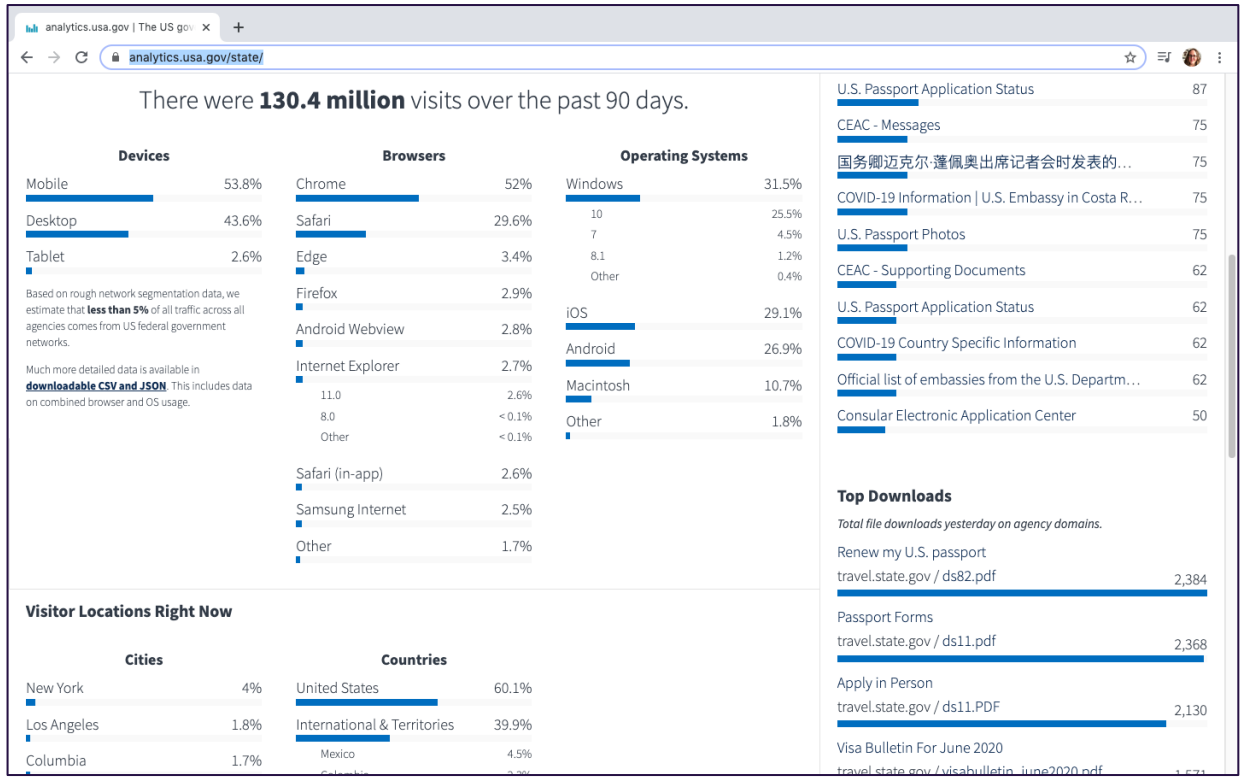
NEW

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

Chat question

How might this data from analytics.usa.gov/state be useful?





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

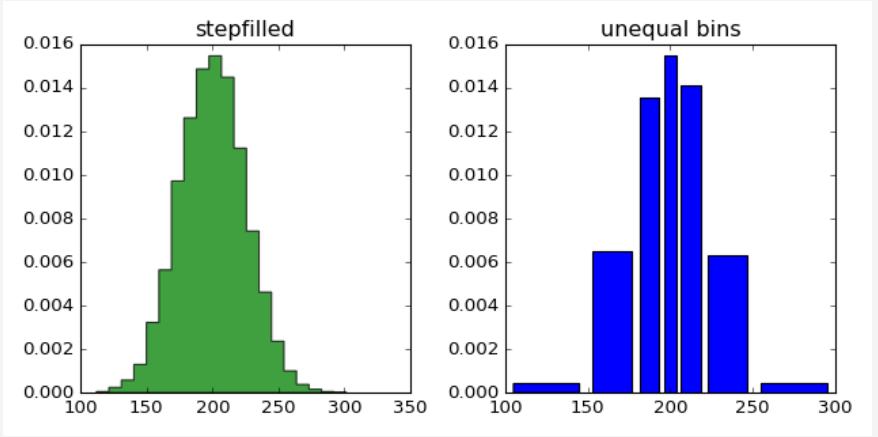
- The benefits of data
- Data analytics overview
- Data governance
- Data tools
- Data teams

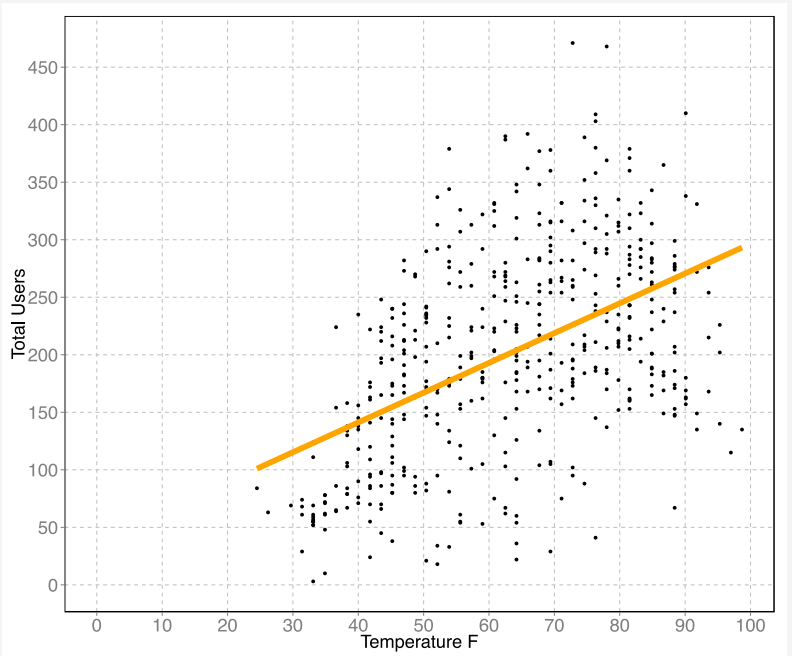
What is data analytics and how can it be used?

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.

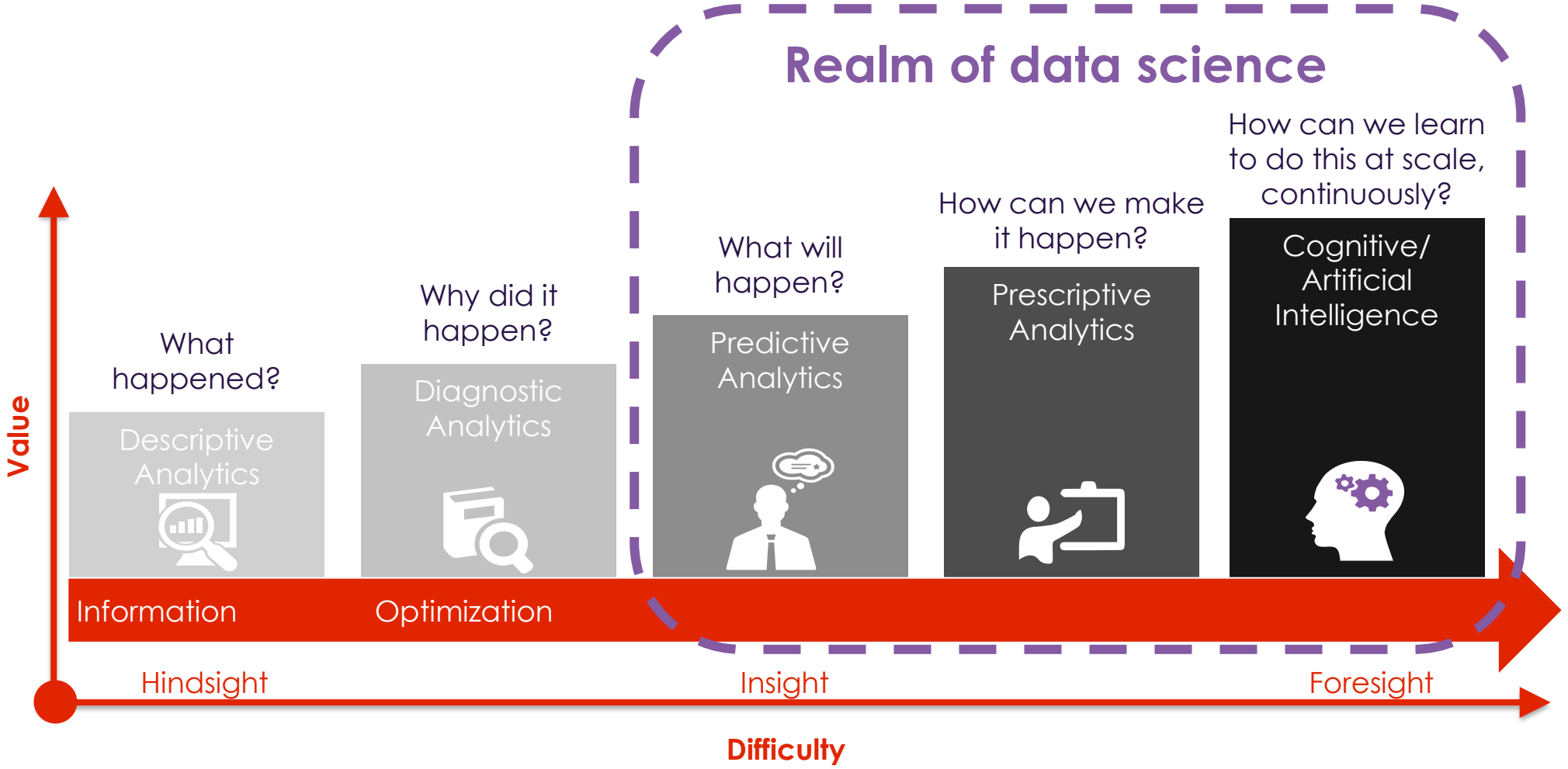




Chat question

What are some ways you currently use data analytics within your organization?

Data analytics maturity model



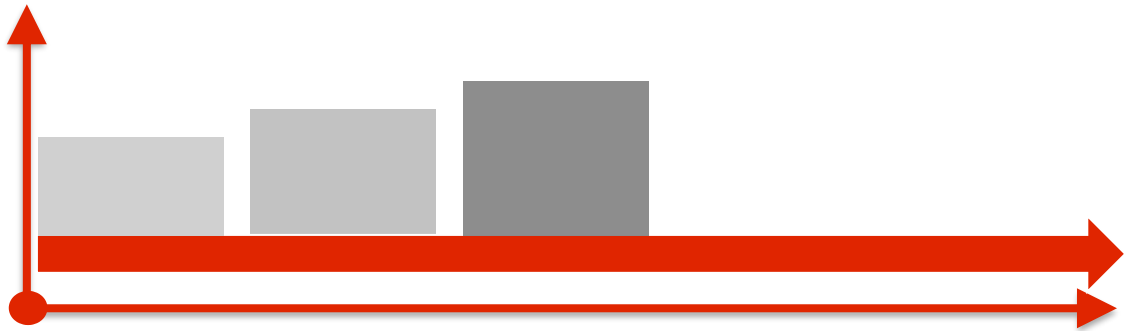
Stage 1: descriptive analytics

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

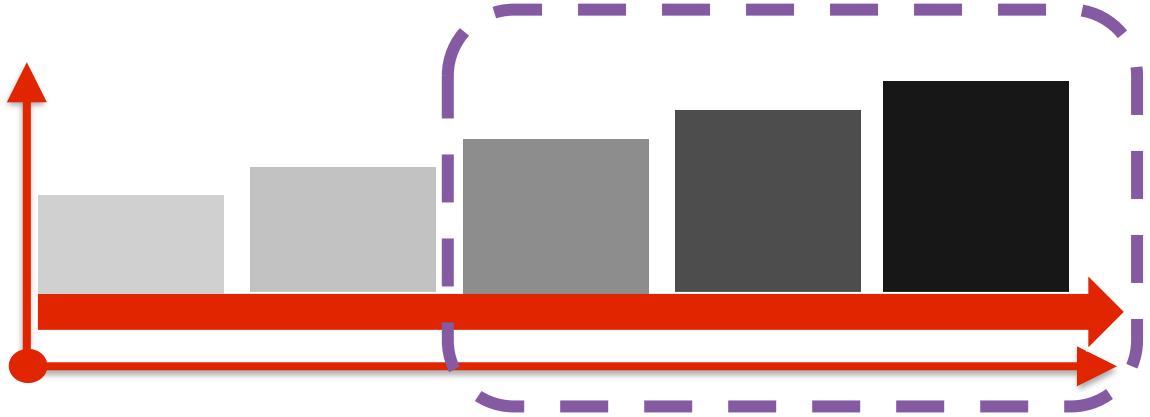
What questions does it answer?	Why did something happen in the past?
How valuable is it?	Provides insights into a particular problem, and can help you identify some root causes for past trends and behaviors
How labor intensive is it?	Requires detailed data, but doesn't have to be overly intensive

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

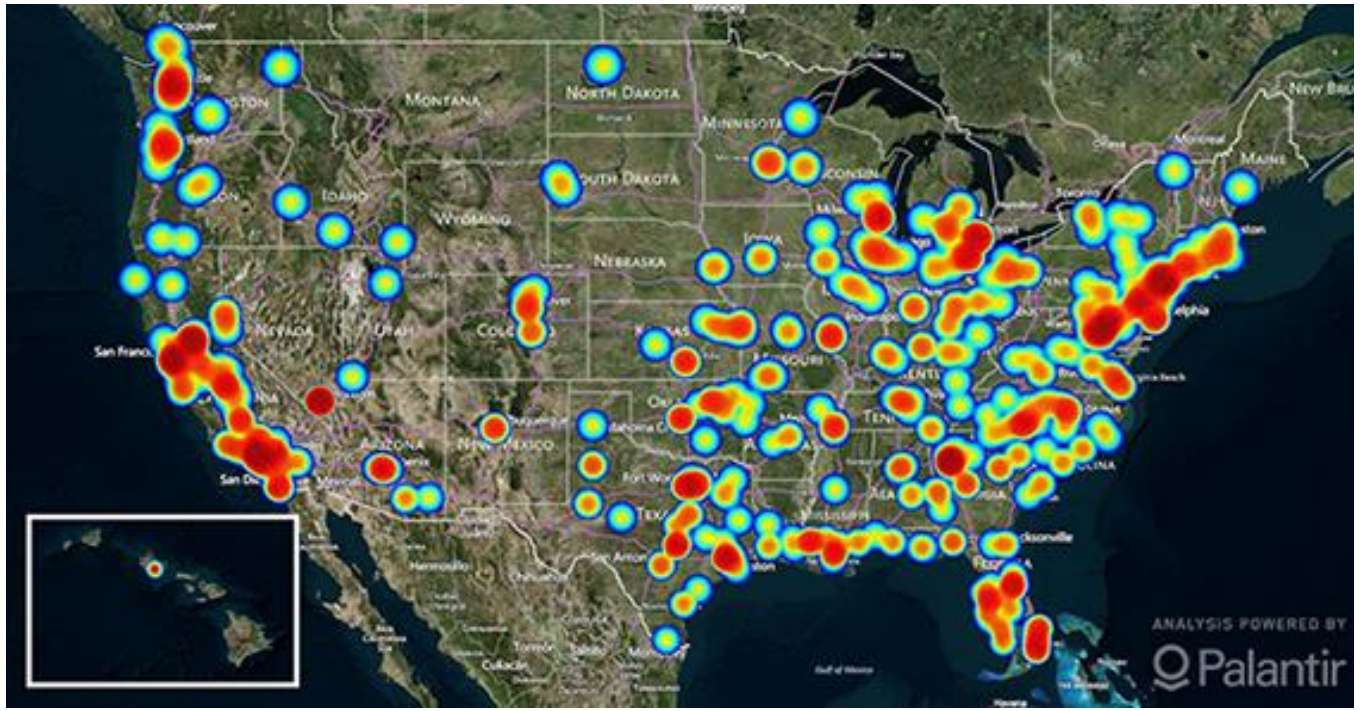
Stages 4,5: prescriptive analytics, AI



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 linked massage parlors and human trafficking.
- Once they find one owner of an illicit massage business by tracing business records, they often find that the owner has several other local businesses.
- 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/

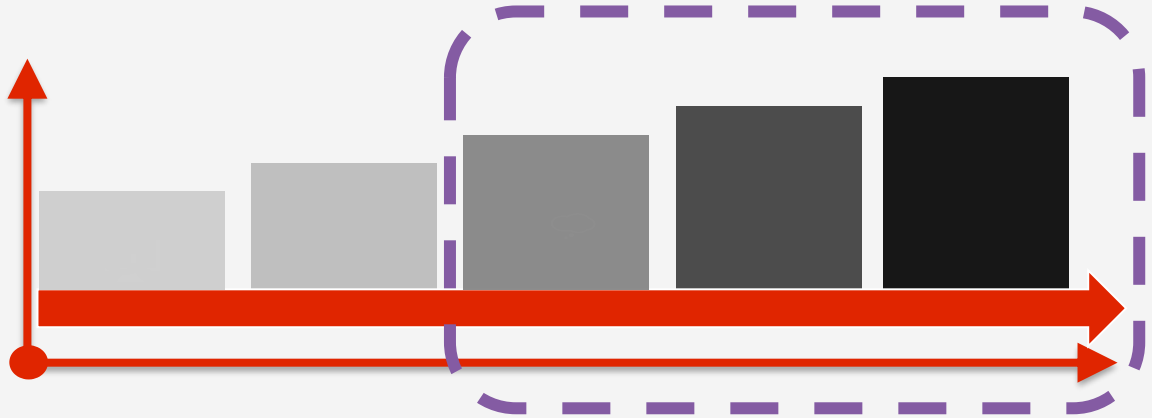
Polling question

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 up?

- 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

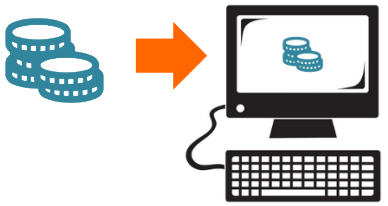
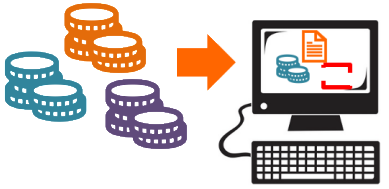
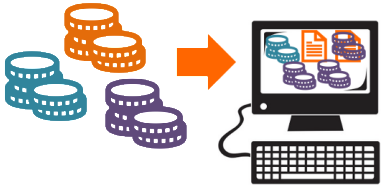
- The benefits of data
- Data analytics overview
- Data governance
- Data tools
- Data teams

What is data governance? Why is it important?
What principles, models, frameworks, and best practices can be used to ensure good data governance?

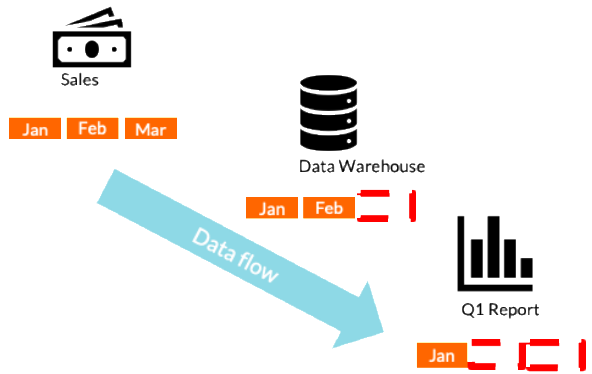
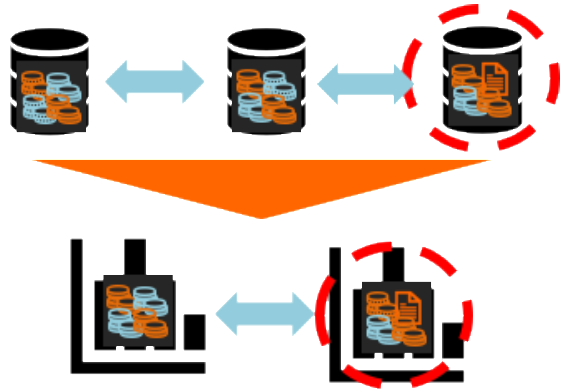
Chat question

What does “data quality” mean to you?

5 components of basic data quality

Component	Definition	Goal	
Accuracy	The data was recorded correctly	Data recorded reflects real life	
Completeness	All relevant data was recorded	Data recorded represents the entire population of outcomes	
Uniqueness	Entities are recorded once	There are no duplicate or indistinguishable records	

5 components of basic data quality

Component	Definition	Goal	
Timeliness	The data is kept up to date	Data is current for its intended use	
Consistency	The data agrees with itself	Databases and reports reconcile	

Quality data is “clean”

Clean data is:

- Valid
- Accurate
- Consistent
- Complete
- Uniform

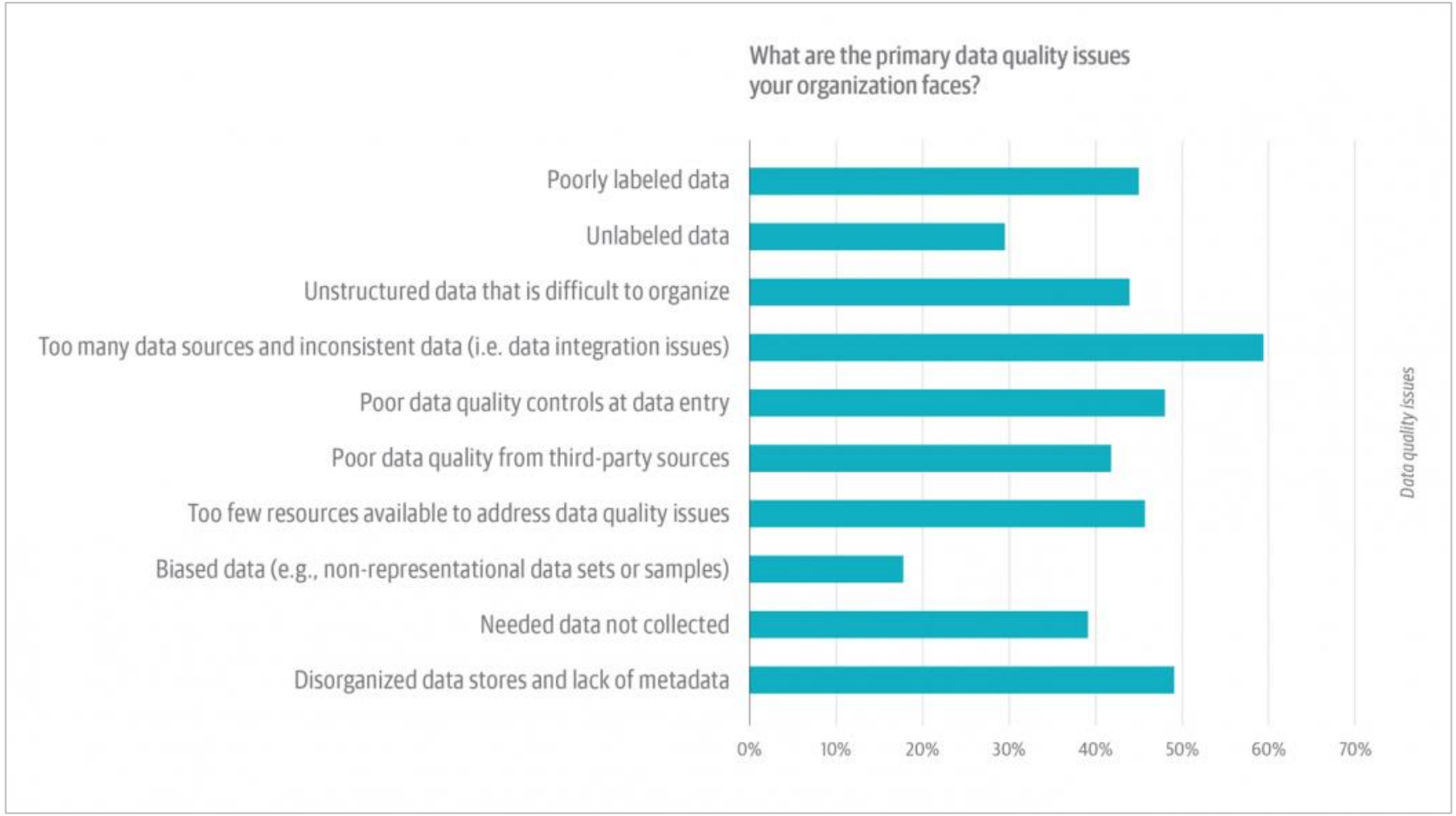
Clean data is **not**:

- Corrupt
- Incorrect
- Duplicate
- Incomplete
- Wrongly formatted

Accessing 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/



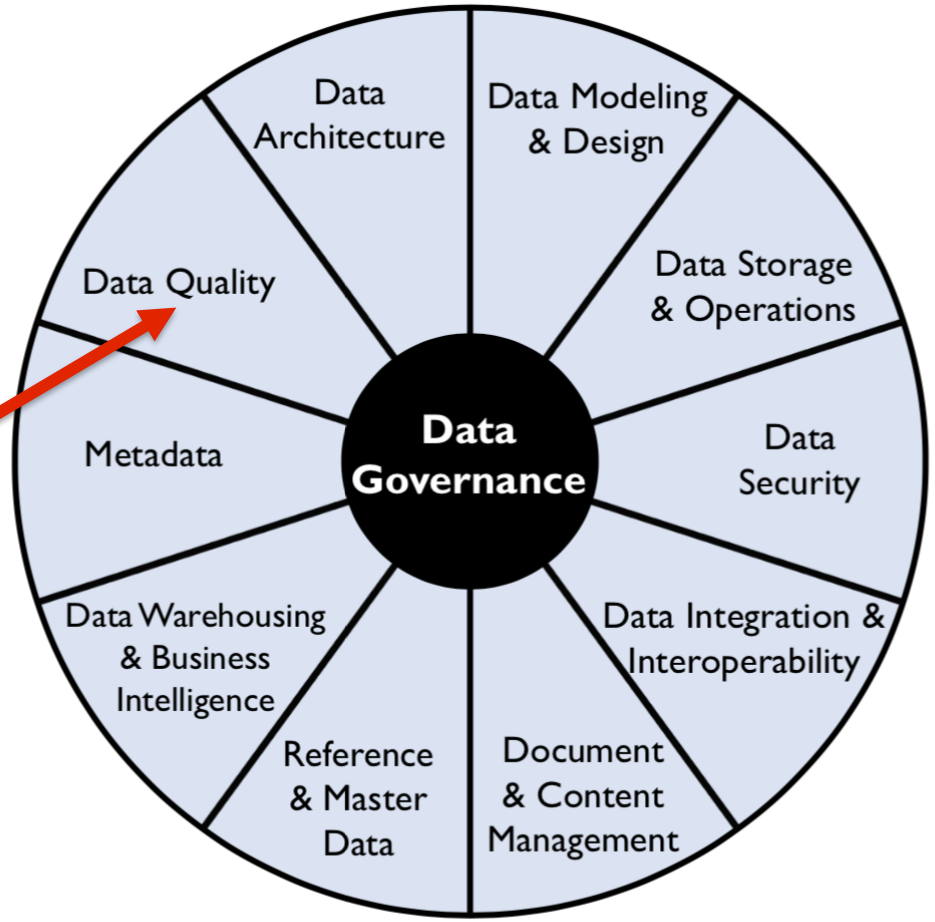
How to prevent data issues?

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

What is data governance?

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

- Data quality is just one component.



DAMA-DMBOK2 Data Management Framework

Why is data governance important?

1. Regulatory compliance – with increased regulation comes compliance that needs to be implemented and followed

2. Reduce risk – effective data governance enhances data security and privacy

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

Data governance principles

A data governance program should be:

1. Sustainable – it survives beyond the initial implementation

2. Embedded – data governance should be present in all processes related to data

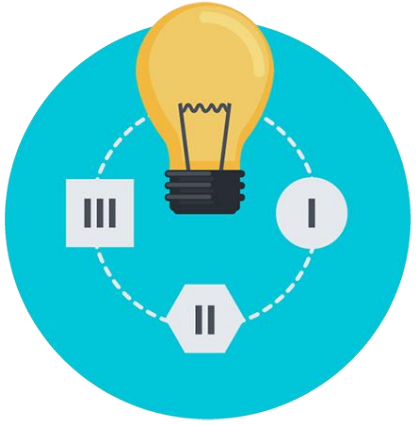
3. Measured – there should be some defined metrics to help demonstrate value to the organization

Data governance strategy

A data governance program might be documented using:

Charters

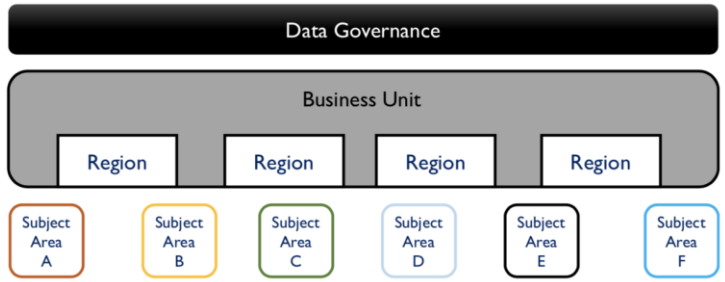
Implementation roadmaps



Operating frameworks / accountabilities

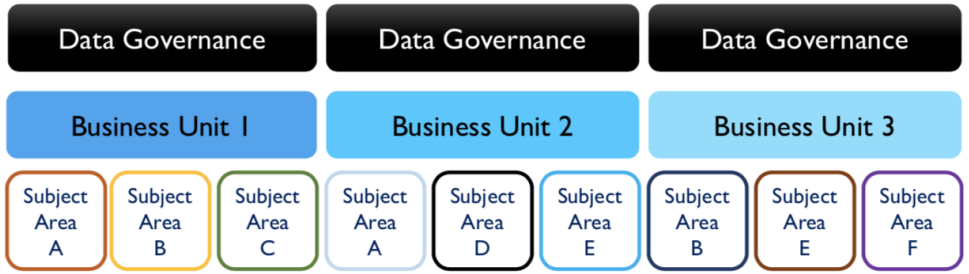
Plans for operational success

Data governance models



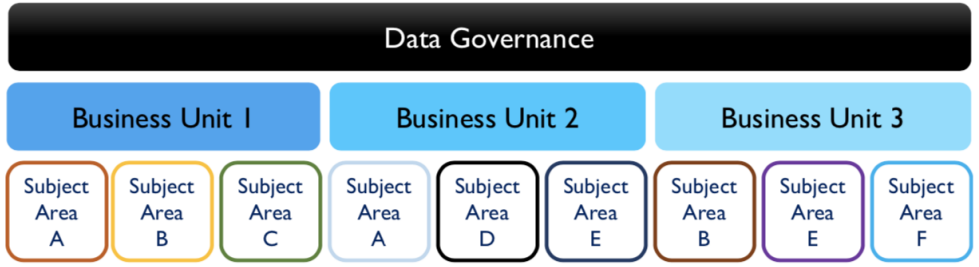
Centralized

One overarching data governance organization applies to all sectors.



Replicated

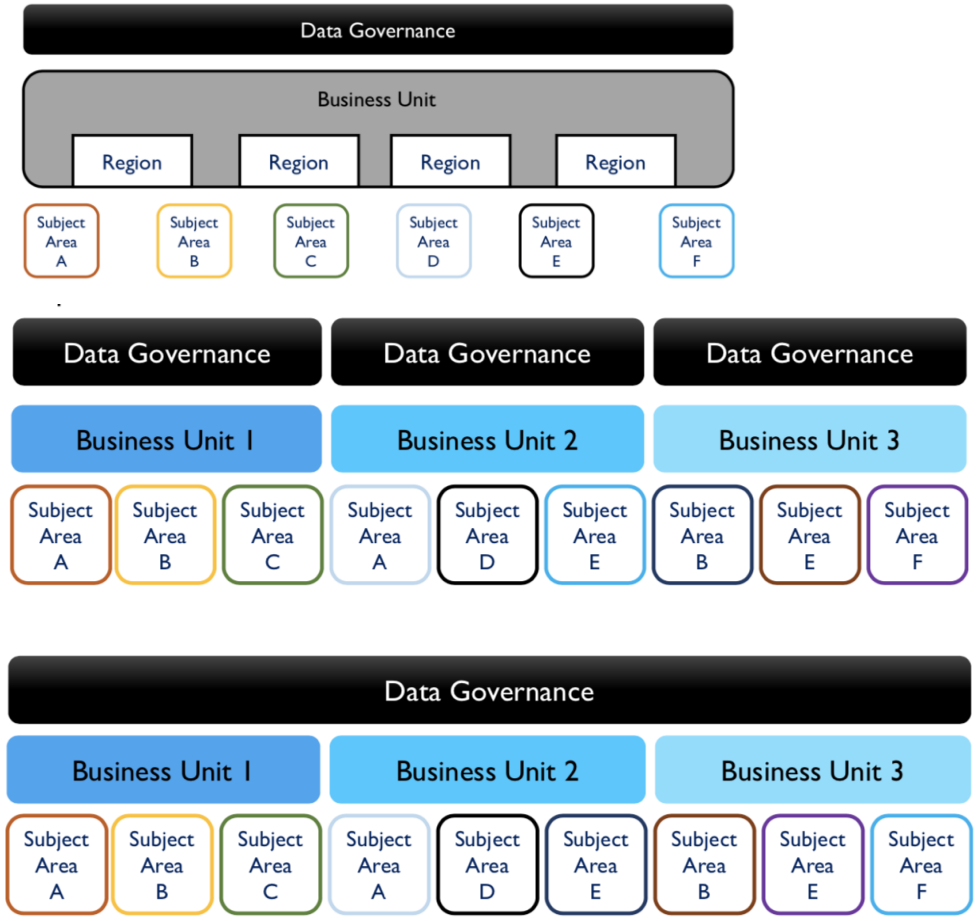
Each data governance section is repeated across departments but may have multiple governing bodies.



Federated

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

Chat question



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

Top: Centralized
Middle: Replicated
Bottom: Federated

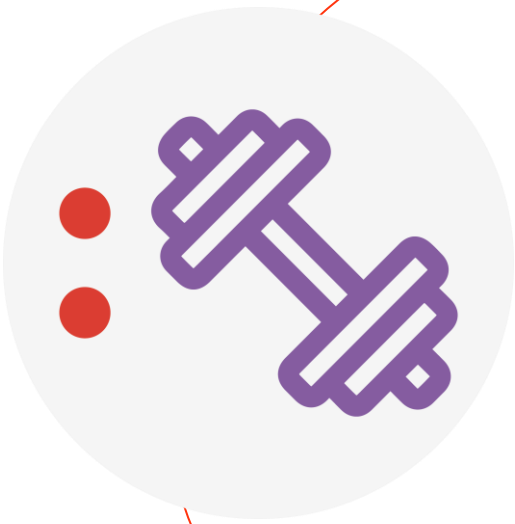
Poll question: data governance

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

Activity: evaluate yourself!

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



Agenda

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

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

Tools

R
git
Spark
GitHub
tableau
Microsoft Azure
THE C PROGRAMMING LANGUAGE
iCloud

Google Drive
MySQL
RStudio
AWS GovCloud (US)
amazon web services
MATLAB
sas
box

python
Microsoft SQL Server
hadoop
mongoDB
JavaScript
Scala
X
Microsoft Power BI

Chat question

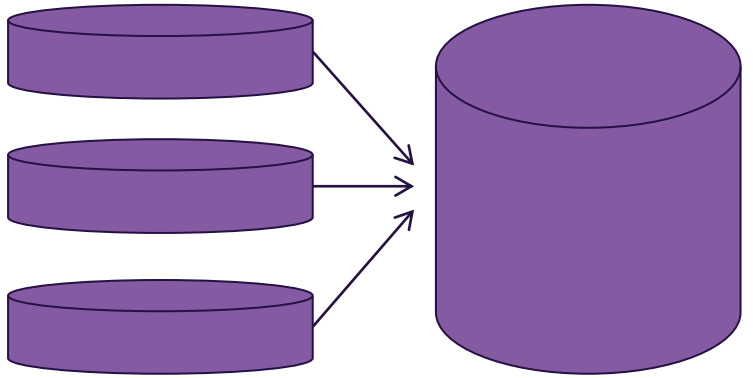
What data tools does your organization use?

Storage tools

- Databases
  - Relational
    - Structured data
    - e.g., Oracle, MySQL
  - Non-relational (NoSQL)
    - Unstructured and semi-structured data
    - e.g., MongoDB

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

- Data warehouses / Data lakes
  - Central repositories of (relational/non-relational) data from one or more disparate sources
  - e.g., Amazon Redshift, Azure Synapse, Snowflake



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.

- Example tools: Excel, R, Python

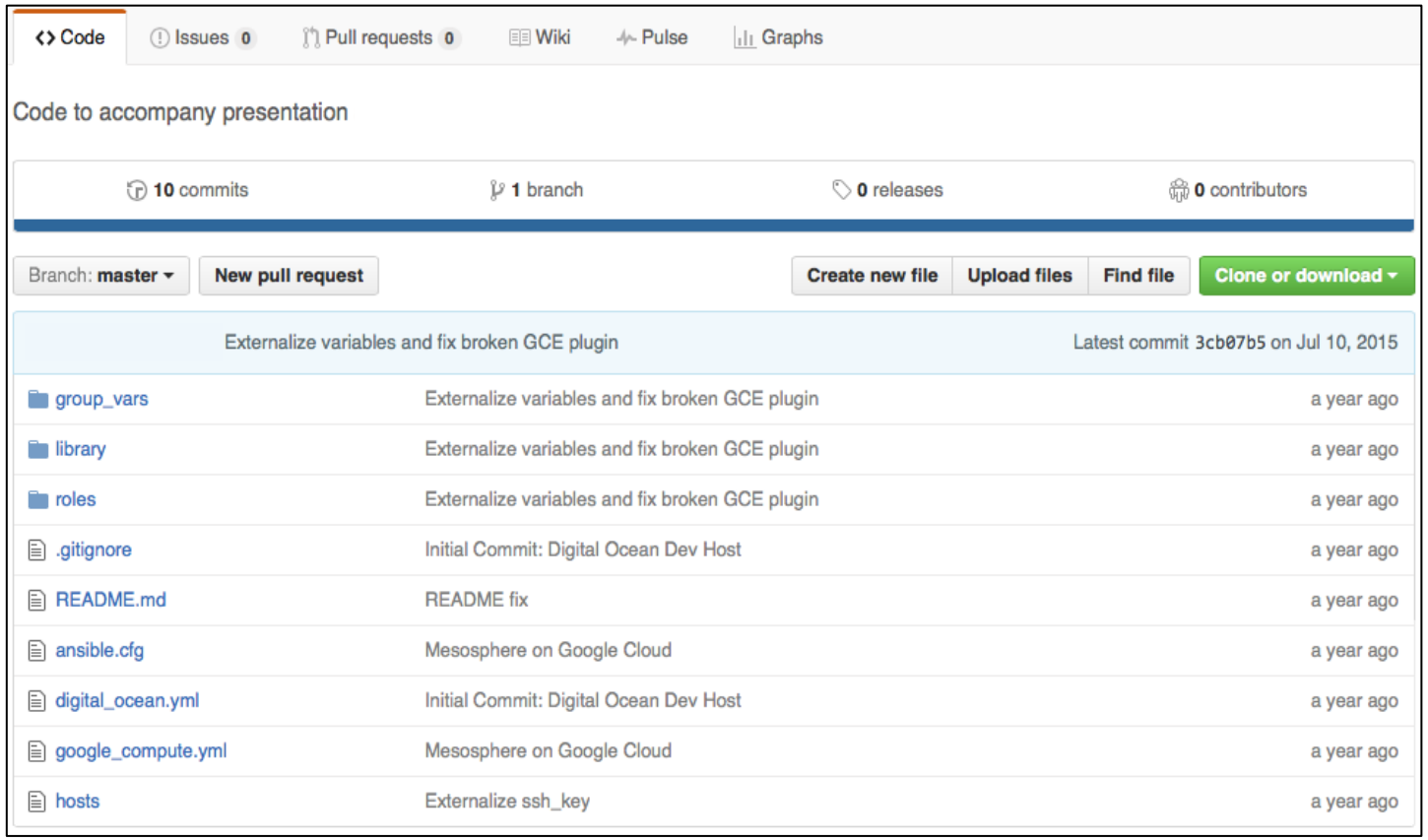
- Your participant guide contains a handout about the pros and cons of various tools on page 26.

Visualization tools

- Visualization gives a visual or graphical representation of data/concepts.
- Example tools: Excel, Google Charts, Tableau, R and RStudio, Python, Power BI

Collaboration tools

- Collaboration tools offer version control, workflow, bug tracking, task management, etc.
- Example tools: Git, GitHub



Questions to guide tool selection

1. Which steps are required in the data pipeline from ingestion to analysis?

2. Which technologies are available for working with data at various stages of the data pipeline?

3. How do different tools and technologies for working with data compare in their functionality, strengths and weaknesses?

4. Do you have staff who can be trained or know how to use particular tools?

5. Do you have budget constraints you need to be mindful of?

6. Is it on the approved software list?

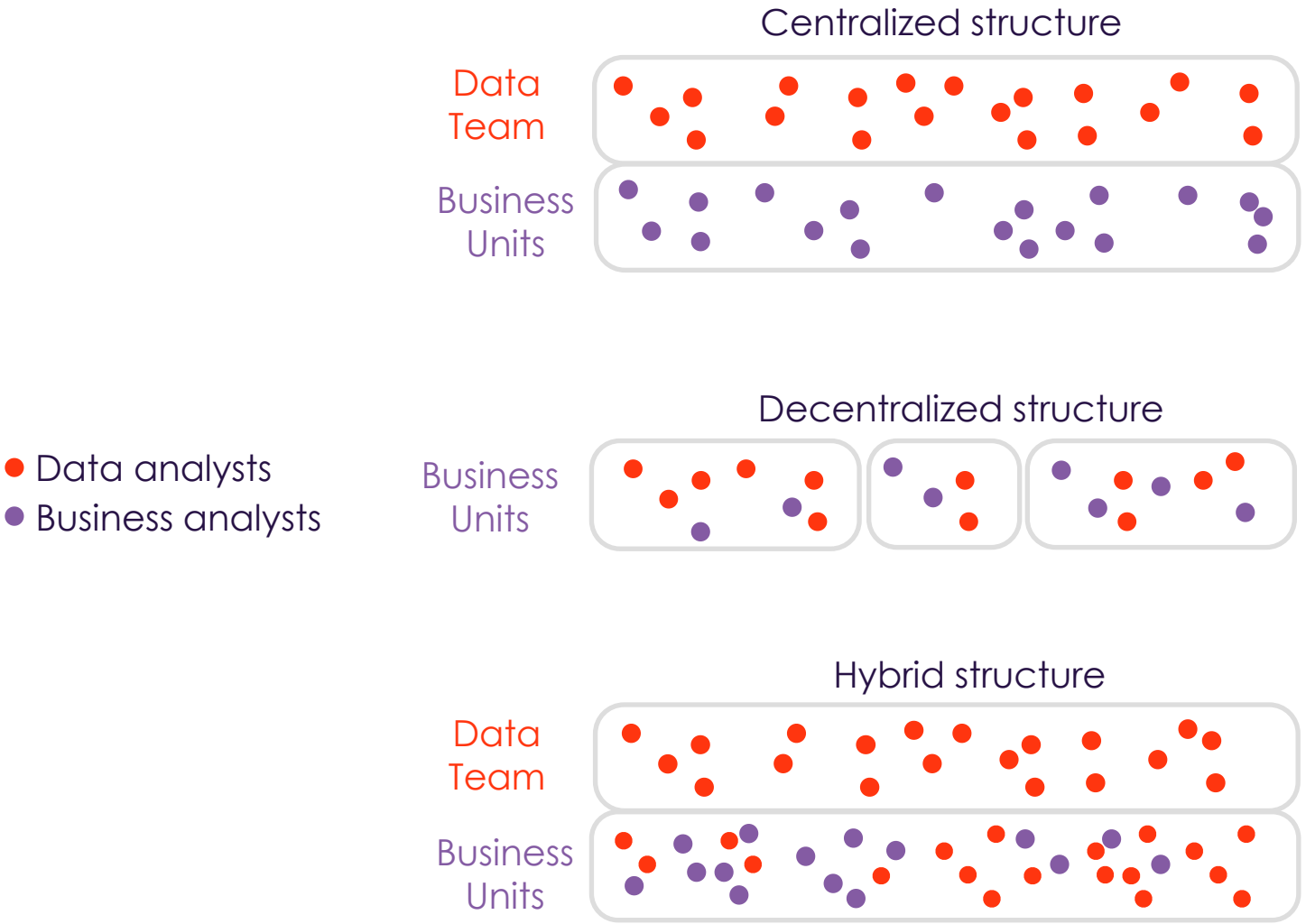
Break

Agenda

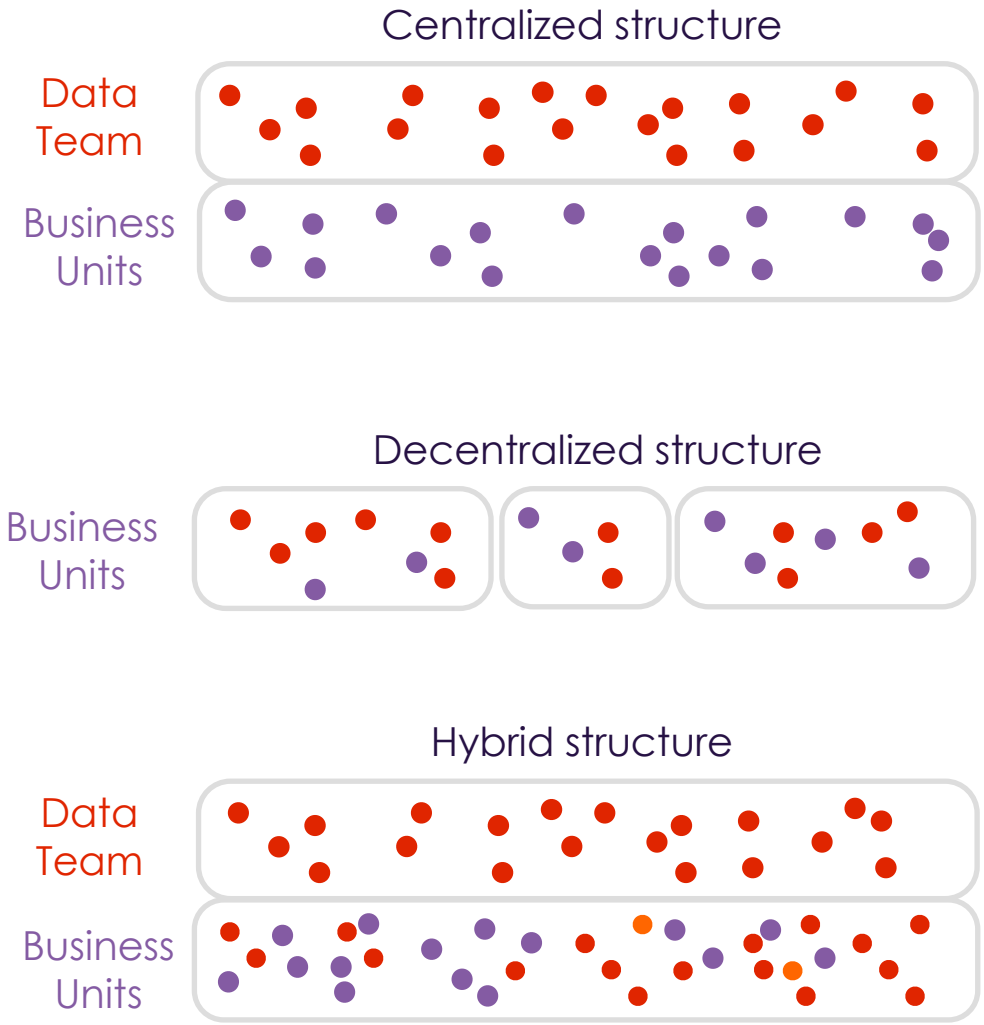
- The benefits of data
- Data analytics overview
- Data governance
- Data tools
- Data teams

How are data teams structured?

Team structures



Chat question

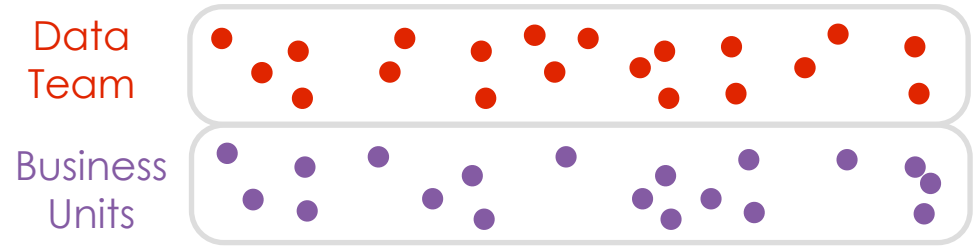


Which best describes the structure of the data teams in your organization?

- Centralized
- Decentralized
- Hybrid

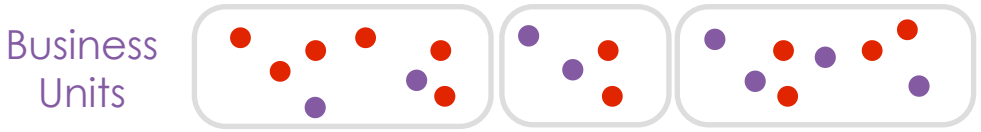
Pros and cons

Centralized structure



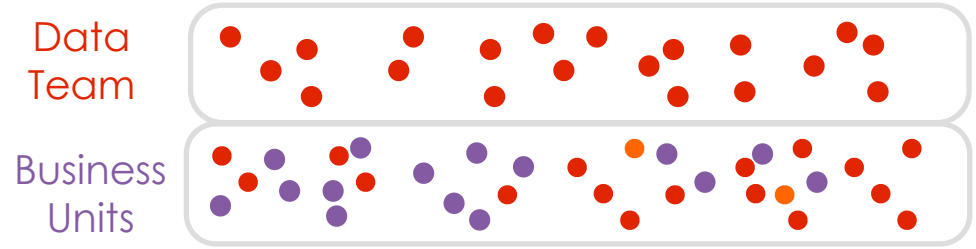
+ easier to standardize team processes
- harder to coordinate projects to meet strategic goals

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

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

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

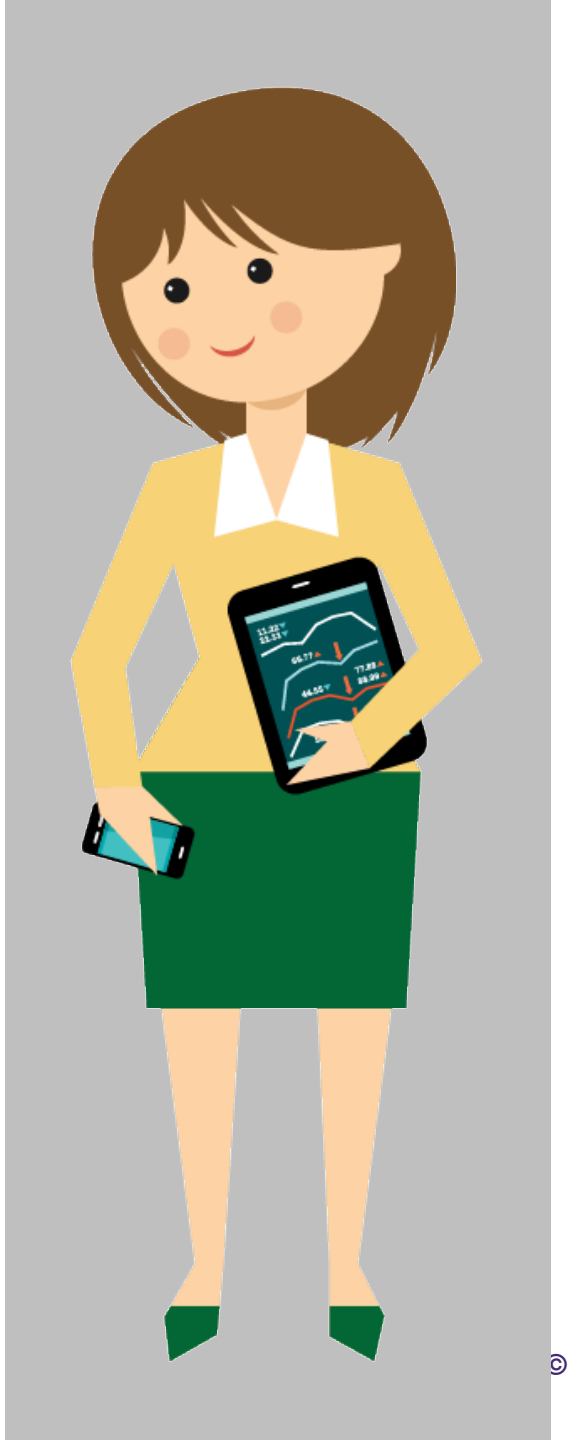
- Core skills to include: SQL, R / Python, Tableau / Power BI

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

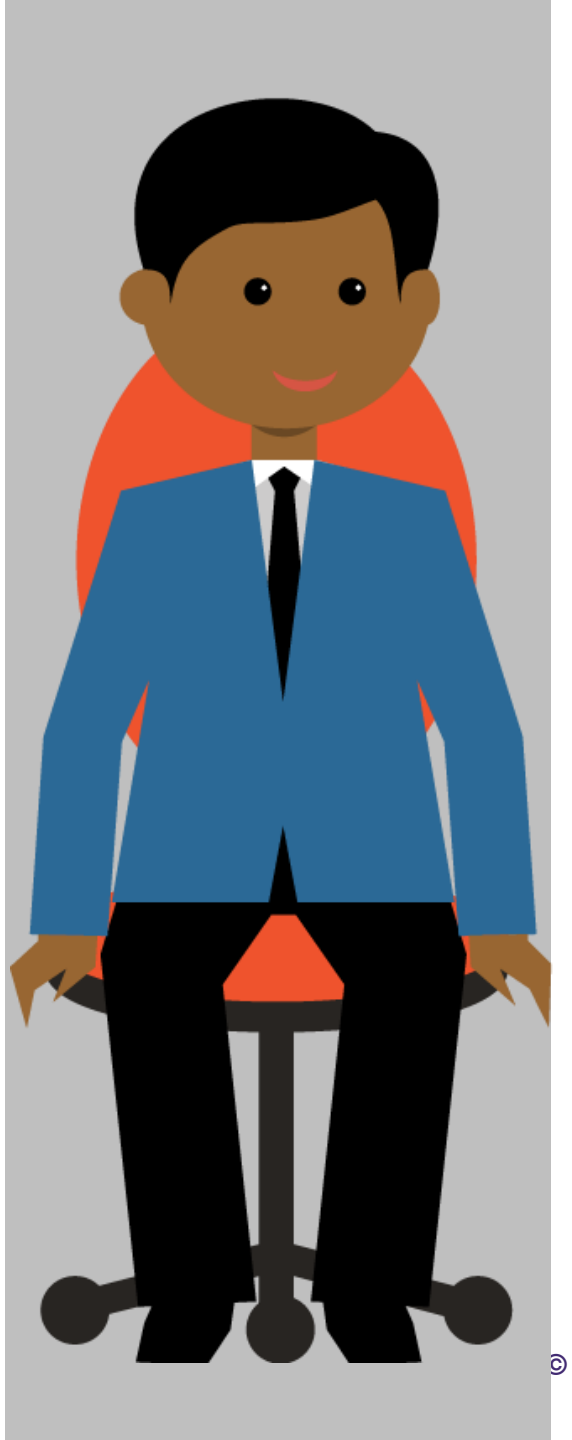


Data engineer

- 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

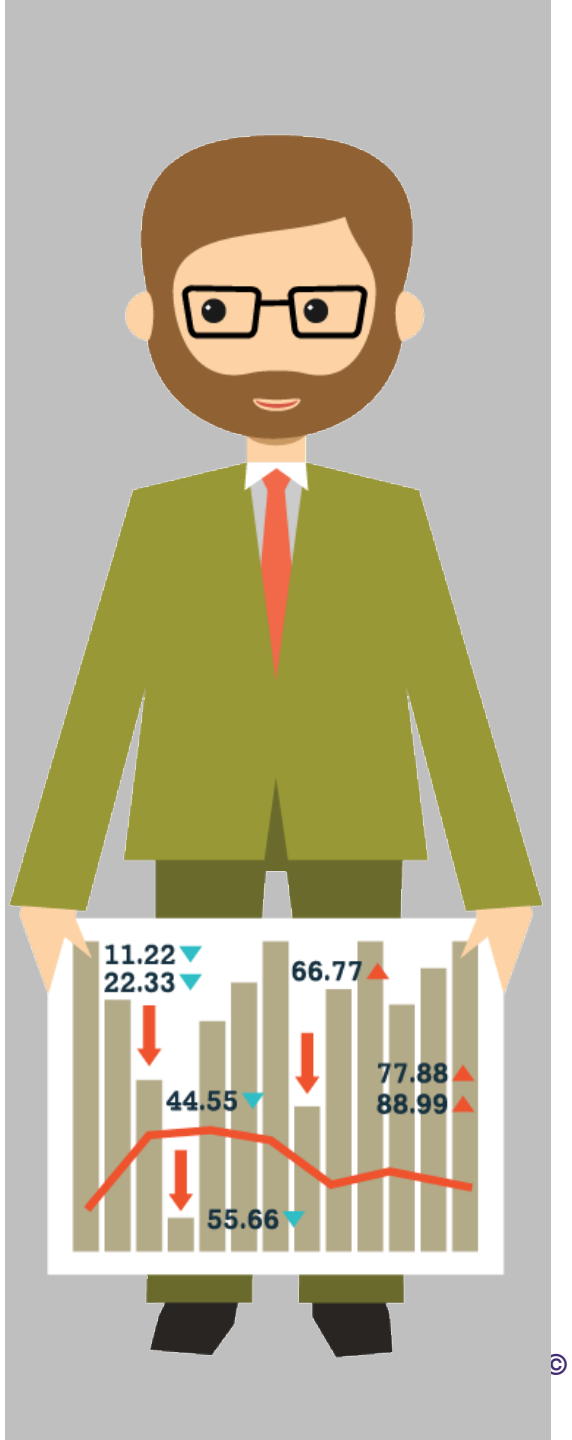


MLOps engineer

- 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

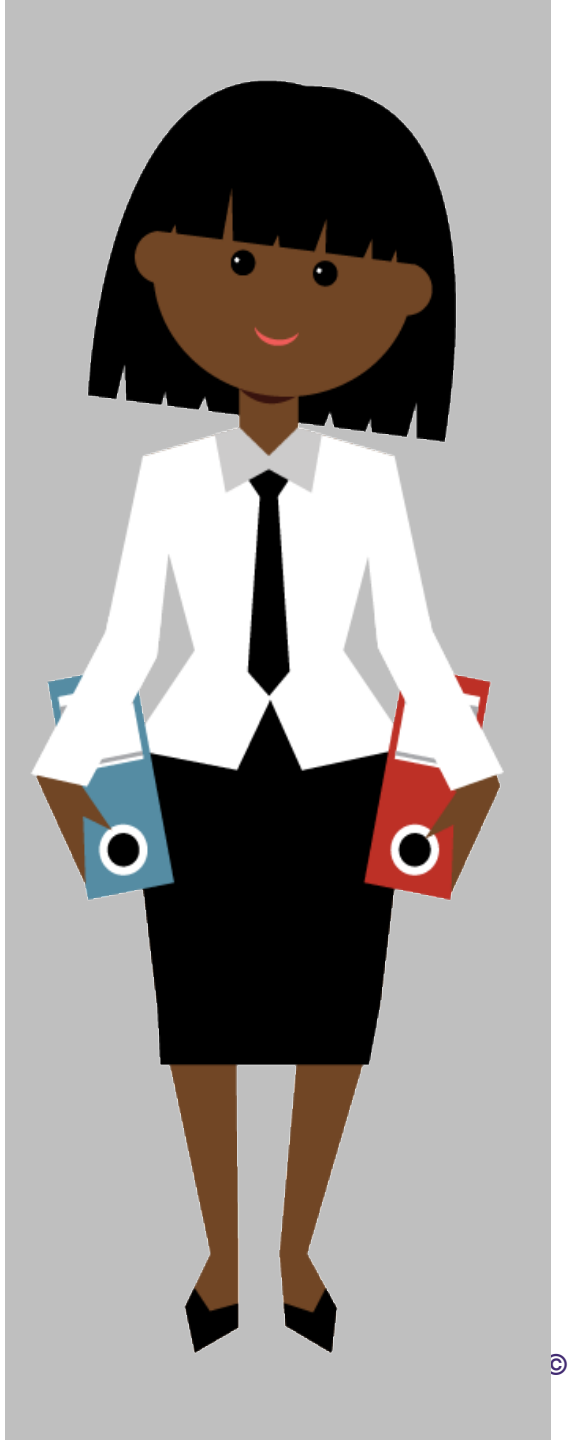


Data science manager

- Oversees and directs data science teams and projects and bridges 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



Chat question

Where do you fit in on your data team?

End of Day 1

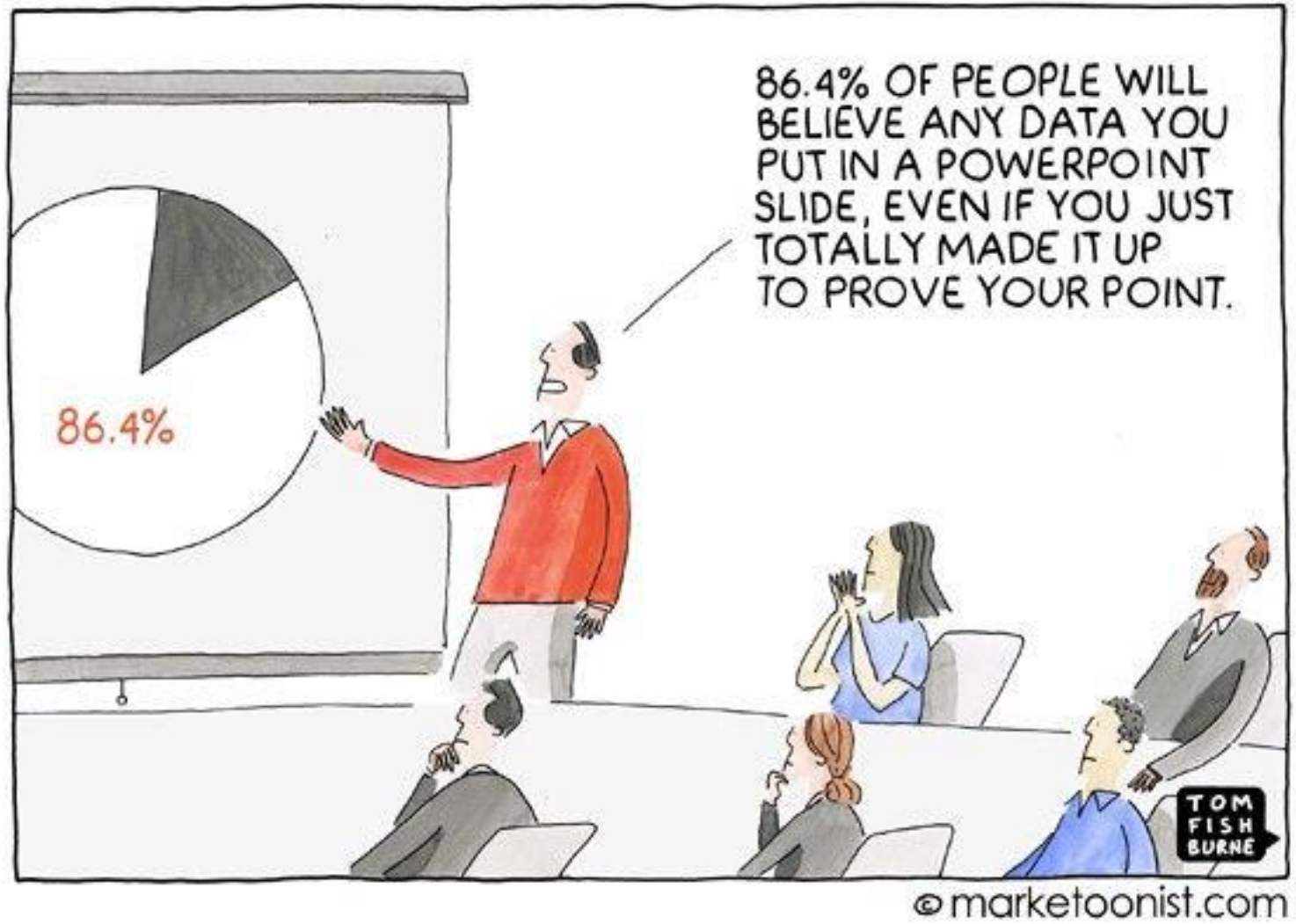
Questions? Comments?

DATASOCIETY:

Fundamentals of Data Literacy

Day 2

Welcome back!



Video: Netflix data driven animated gif campaign

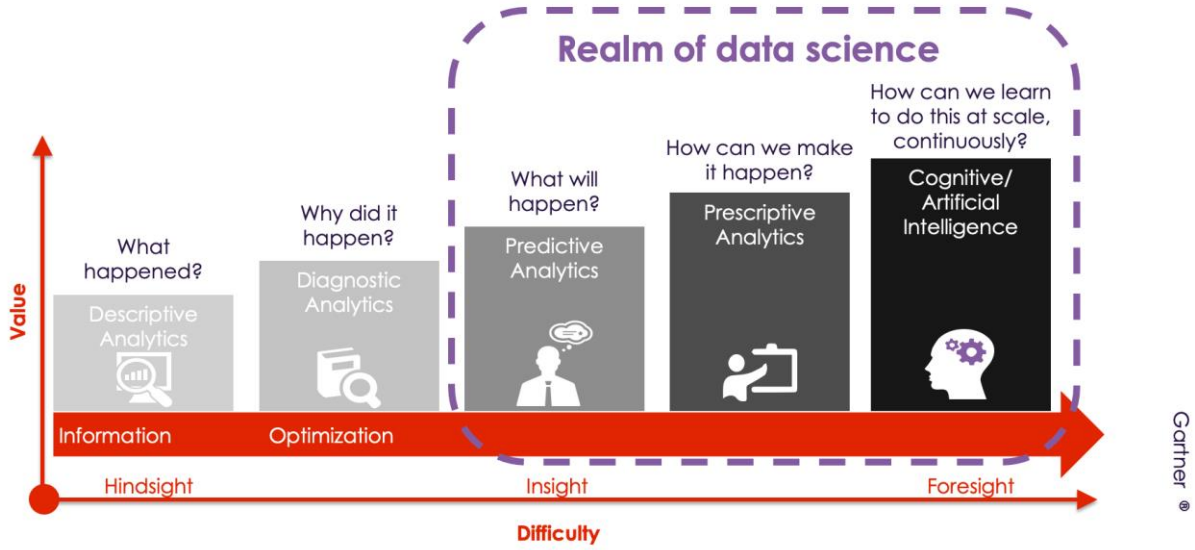
https://www.youtube.com/watch?v=tZkILxaANLU&ab_channel=WeAreNetflix

Chat question

How can you use data in your organization? Think of a specific project or a task.

Remember!

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



Agenda

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

What is a data-driven culture?
How can I help build one?

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.

What is a data-driven culture?

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

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

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.

Example: NASA, FAA, and BetterUp

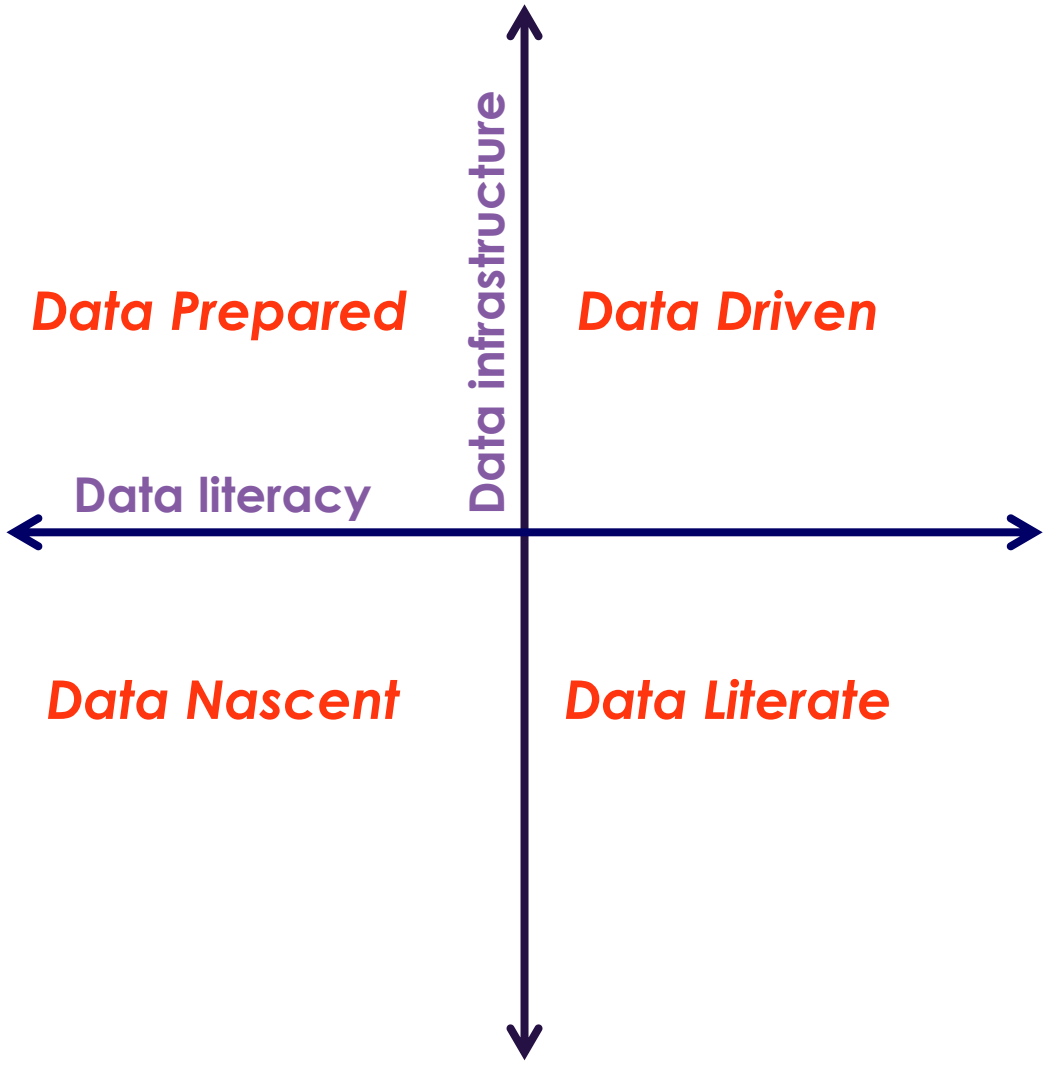
- Both NASA and the FAA are the first federal agencies to contract with BetterUp, an app-based workforce coaching platform grounded in behavioral analytics and machine learning.
- Algorithmic feedback helps tailor users' experiences and match employees with coaches.
- Executives and supervisors have access to a full data analytics dashboard providing insights their learning progress.
- The goal is to produce individualized professional development programs that best meet user needs.

Chat question

How else have you experienced the availability of data affecting your job's tasks and responsibilities?

Components of a data-driven culture

- A data-driven culture can be separated into two components:
  - Data infrastructure
  - Data literacy



Data infrastructure

- Components of data infrastructure include:

DATA ACCESS

Can staff access data easily and in a timely manner?

DATA STORAGE

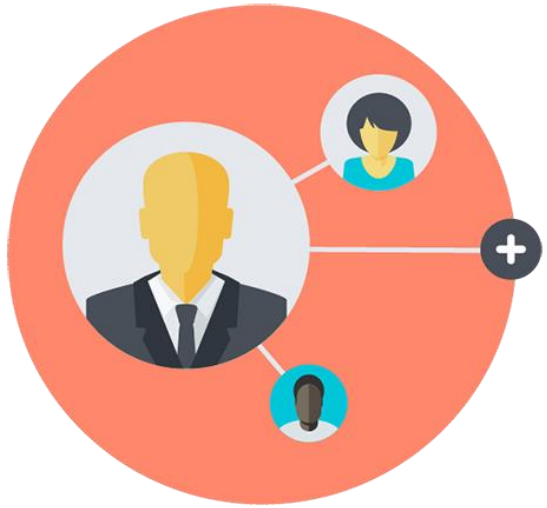
Is the data stored securely with a backup?

DATA COLLECTION

Is data collected in a timely and clean way?

Data literacy

- Components of data literacy include:



DATA LEADERSHIP

Do executives champion data usage?

DATA GOVERNANCE

Are staff aware of data standards and practices?

DATA KNOWLEDGE

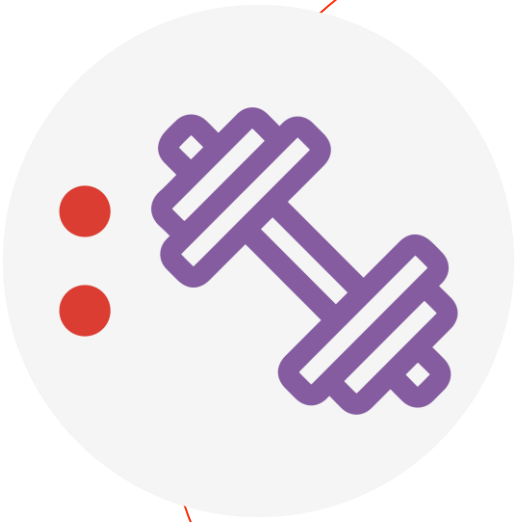
Does staff understand how to ask questions of data?

Activity: Are you data driven?

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

- In the chat, answer the following questions:

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



How to improve the culture?

- You can help build a data-driven culture no matter where you sit in an organization

- Steps you can take today include:
  - seeking out information on your organizations data policies and capabilities, including who maintains the datasets you need
  - modeling data-driven decision making in meetings
  - asking for the metrics and methodology behind conclusions and reports
  - highlighting successful data projects in newsletters or events
  - bringing in data experts for "lunch and learns"
  - encouraging colleagues to attend data trainings (like this one)

Chat question

What other strategies can you use to locate and connect with other data-driven members of your workforce?

Remember

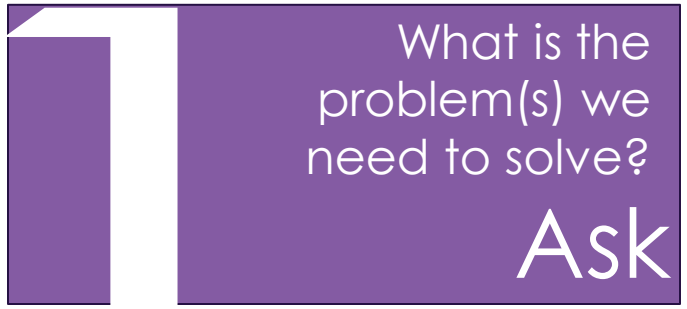
- Give people the opportunity to fail.
- This is an iterative process – it takes several tries to get it right.
- Be flexible.

Agenda

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

What are the six stages of the typical data science process?
How do I fit into the process as a non-data scientist?

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?

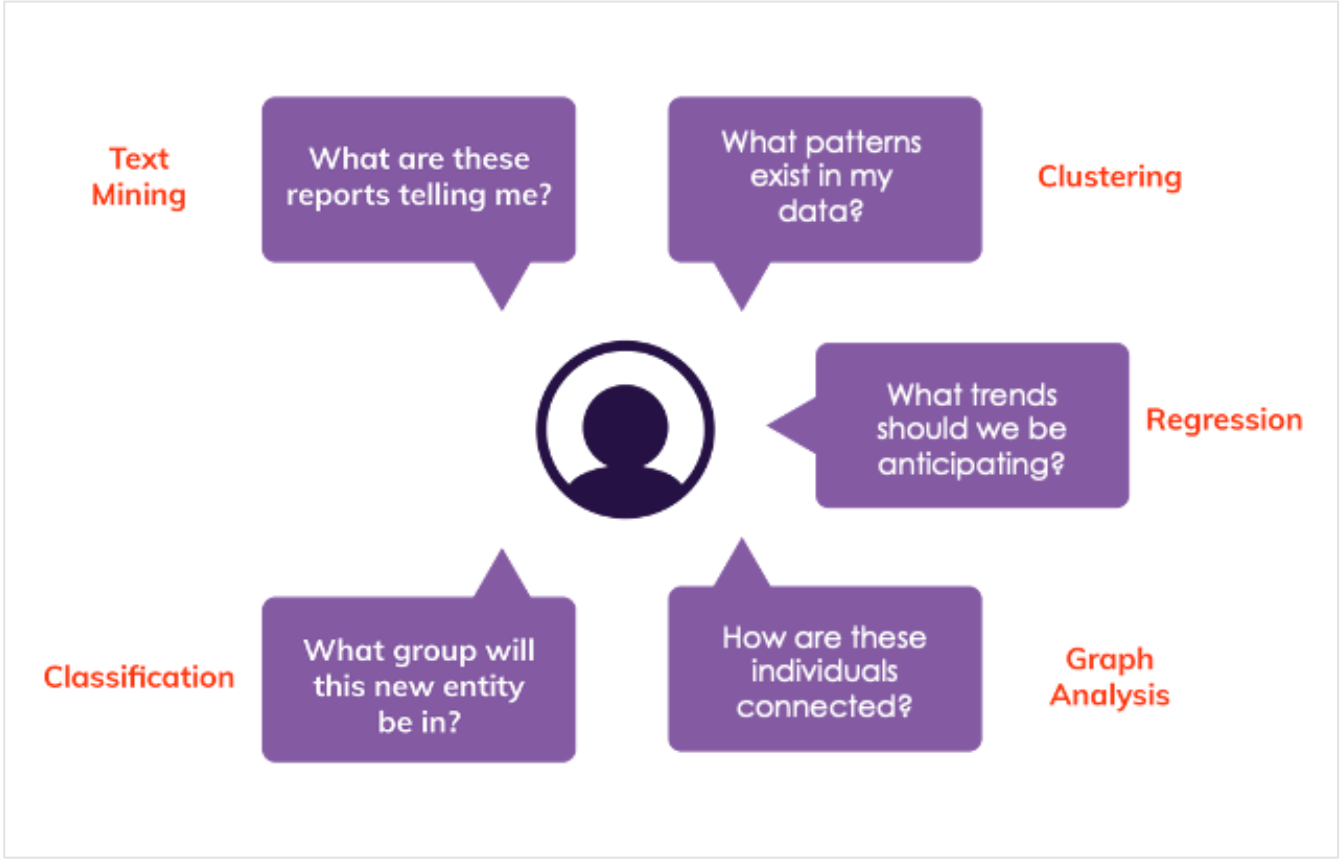
3 Which method(s) is appropriate to use? **Model**

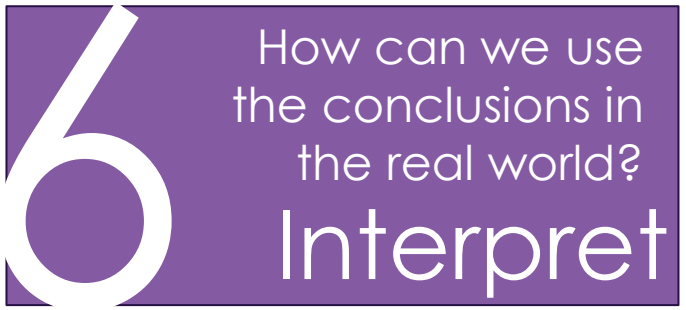
4 Do the model and assumptions work as expected? **Validate**

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

DATA SOCIETY

- 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.

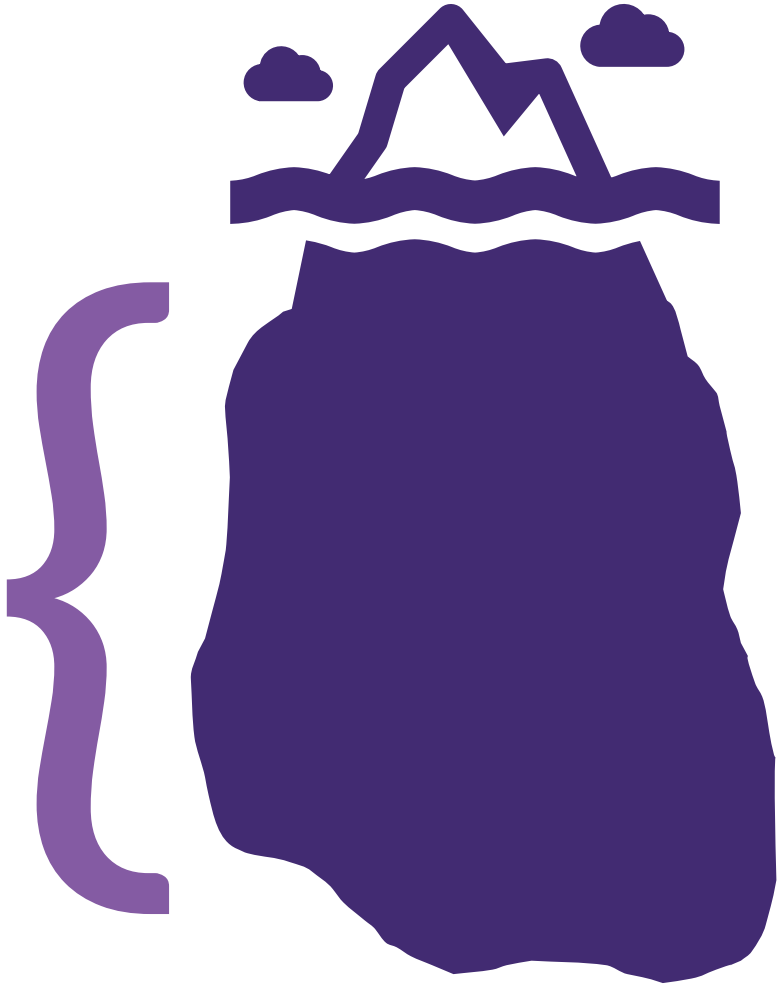
Example

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?

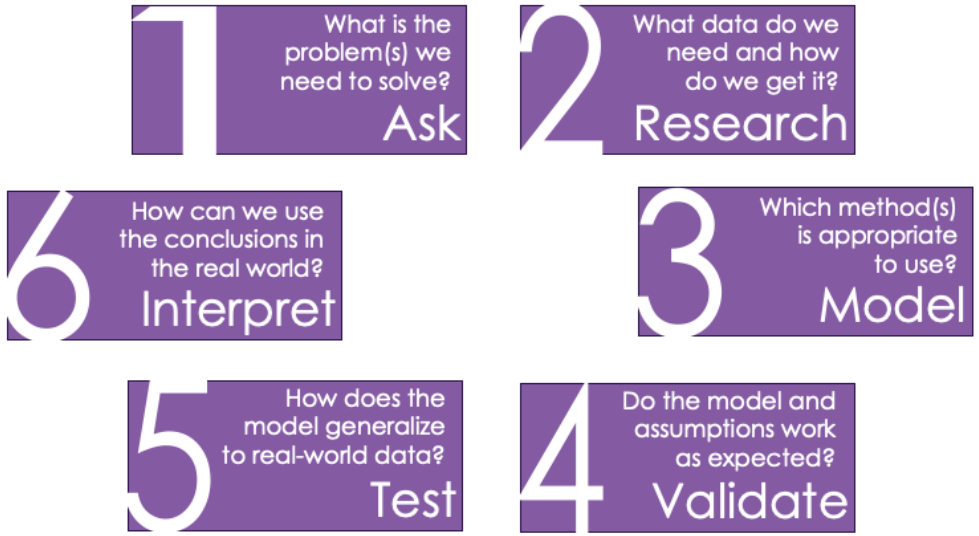
The iceberg of data analysis

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



Polling question

During which phase(s) of the data science process do you suspect data cleaning occurs?



How do I fit in?

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

Help decide what projects to tackle

- Is the question posed specific, measurable, and objective?
- Will the results be used by decision makers?
- Is the starting data good? Have internal data governance protocols been followed? Are you confident in any external data that you plan to use?
- Are there limits to what can be done with your data based on type or volume?
- Do you have the teams, tools, and processes necessary for the project?
- Is the project ethical?

Be clear about what is needed

- How will the results be used? What do you hope to achieve?
- Will you potentially want this reproduced with different parameters?

Ask questions about the process

- How did we obtain the data?
- How much data cleaning did you need to do?
- What programming language did you use?
- Where did you get the code from and how much did it get edited?
- How did you make sure your model is valid?
- How does the accuracy of this model compare to others used in this field or industry?
- Can this be reproduced?

Ask questions about the results

- What do these results mean?

  A good data scientist should be able to talk you through what the pretty chart or graph actually means.

- What would you suggest we do with the results?

  A good data scientist will have done their own research or met with a subject matter expert to develop a point of view.

Break

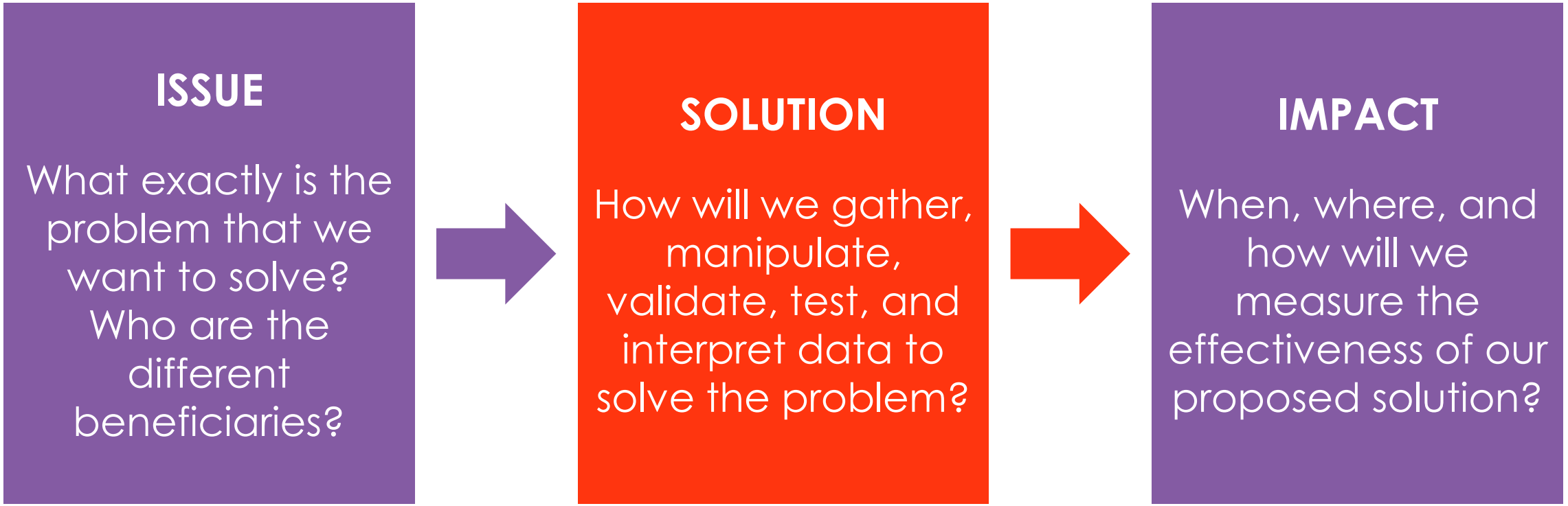
Agenda

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

How do I identify feasible and impactful data projects?

From issue to impact

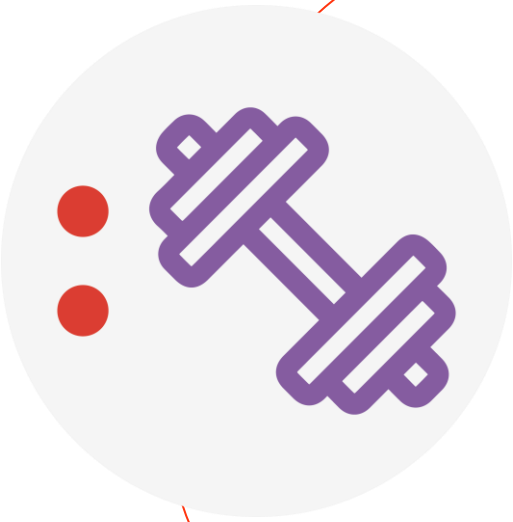
- When thinking about data science projects, it can be useful to consider the arc from **issue** to **impact.**



Activity: brainstorm ideas

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

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



Chat: data projects

1. Select one idea that you think is the most feasible

2. Craft the specific, measurable, and objective question that you'd take to the data science team

3. Identify what type of data you'd need to complete this project and tools you would use

4. Consider any challenges you may encounter

Agenda

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

How do I identify feasible and impactful data projects?

World's Smartest Home

Chat question

What two major problems were being solved?

What kinds of data did the various systems use?

World's Smartest Home

- **Problem #1:** Only allow children to turn on the television as a reward for when their rooms are clean.

- **Data:**
  - Set of 200 training photos of exact space to be monitored (100 "clean" photos, 100 "messy" photos)
  - Live video footage of rooms

- **Other needs:** Connect Alexa trigger to smart TV.

- **Problem #2:** Automate tagging, captioning, and upload of photos to family website to reduce labor.

- **Data:**
  - Less clear in this case – but what would you guess?
  - Probably involves training photos and a separate captioning AI.

- **Other needs:** Connect home AI to personal website and establish trigger conditions for upload.

One problem, two flavors

Classify novel images into various categories

- **Problem #1:** Only allow children to turn on the television as a reward for when their rooms are clean.

Is the room **clean** or is it **dirty?**

- **Problem #2:** Automate tagging, captioning, and upload of photos to family website to reduce labor.

Which **people** are in this photo?

What are they **doing**?

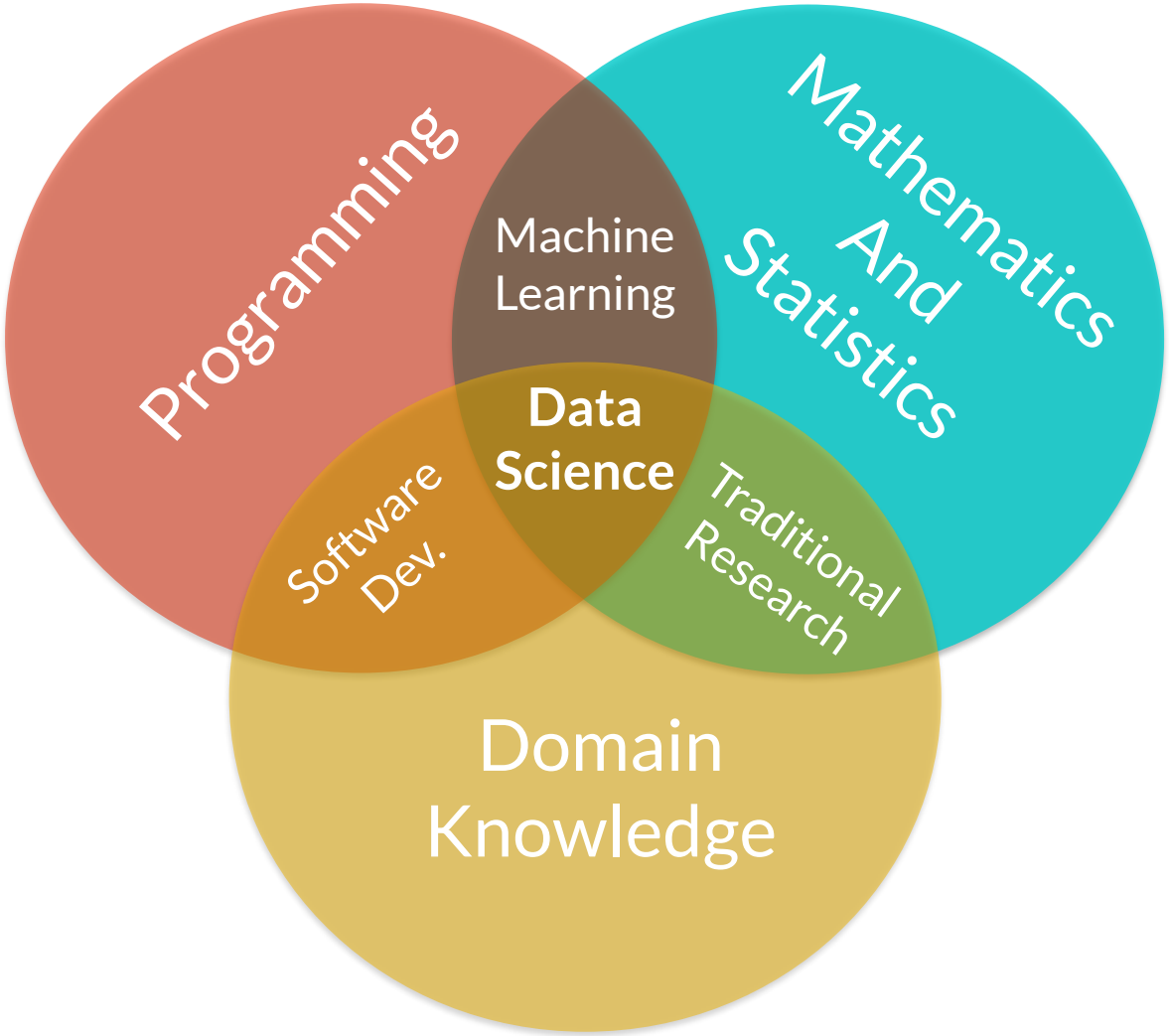
Why learn these terms and concepts?

1. To develop a common vocabulary with the data science team

2. To direct data science projects and make recommendations

3. To understand what options are available for finding new insights and becoming more efficient

Data science principles



What is machine learning?

- Uses **algorithms** to find patterns in massive amounts of data and predict future results with minimal human intervention

- 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 **unsupervised** or **supervised**

Unsupervised learning

- We call machine learning **"unsupervised"** when an algorithm is allowed to detect and learn patterns based on **untagged** input data.

- The goal is to model the underlying structure or distribution in the data in order to learn more about its composition.

- In other words, the machine looks for whatever patterns it can find, usually to **generate categories.**

- *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*

Supervised learning

- We call machine learning **"supervised"** when an algorithm is given **tagged and labeled** input data (identifying certain properties, characteristics, or classifications).

- Rather than generating categories, we're now tasking our algorithm with **mapping input to a desired output.**

- In the end, we want to create such a solid map that we can predict the output for any novel input data.

- *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."*

Chat question

"He looks at our rooms… and he can tell if they're dirty or clean. The red means it's dirty, the green means it's clean."

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

End of Day 2

Questions? Comments?

DATASOCIETY:

Fundamentals of Data Literacy

Day 3

Recap: data projects

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

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

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

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

Agenda

- Foundational data science methods (ctd.)
- Clustering problems
- Classification problems
- Regression problems
- Working with text data
- Working with network data

What does a data project look like in practice?
What kinds of considerations are there for working with complex data?

What is an algorithm?

Inference vs. prediction

- Data science methods use algorithms to answer two broad kinds of questions.

- **Inference:** Given a set of factors and an outcome, what kinds of associations can we deduce?
  - *Which segment of campaign spending contributed most to the election outcome?*

- **Prediction:** Given some input data, how accurately can we estimate the output?
  - *What is the likelihood of a certain election outcome based on voter demographic data?*

- Data science projects may involve questions of both types: *Once we have **inferred** a set of likely predictors, how can we improve the **predictive** power?*

Polling question

"Does level of education or gender have a more significant effect on financial literacy within a given population?"

Will this question result in an inference or in a prediction?

Polling question

"Does the level of education indicate the level of financial literacy within a given population?"

Will this question result in an inference or in a prediction?

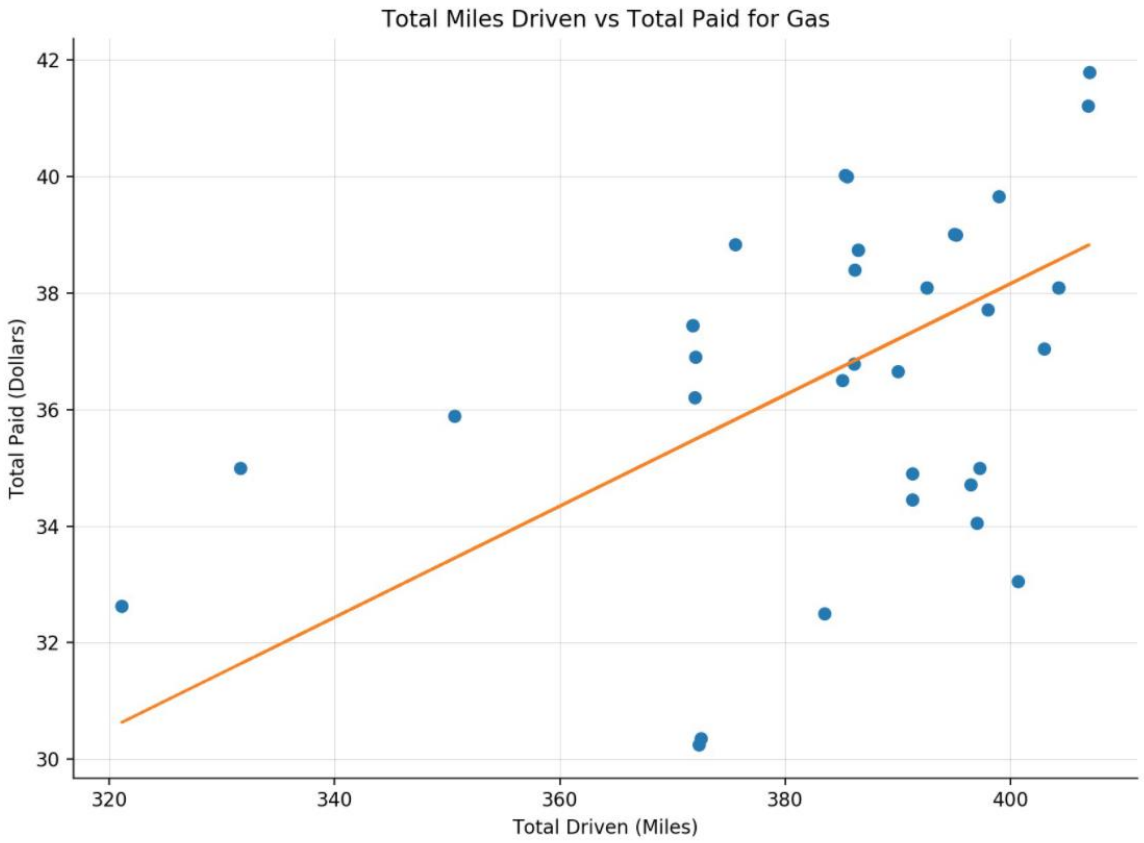
Classification vs. regression

- Depending on the kind of desired output, data scientists will select a method allowing them either to perform a **classification** or a **regression**.

- **Classification** sorts response variables into **categories.**
  - Based on chest X-ray images, we can categorize someone as either a COVID patient or a pneumonia patient.

- **Regression** generates response variables that are **numerical.**
  - Based on inputs concerning the number of COVID cases within a given area, we can forecast the likely number of cases in the short term.

Chat question



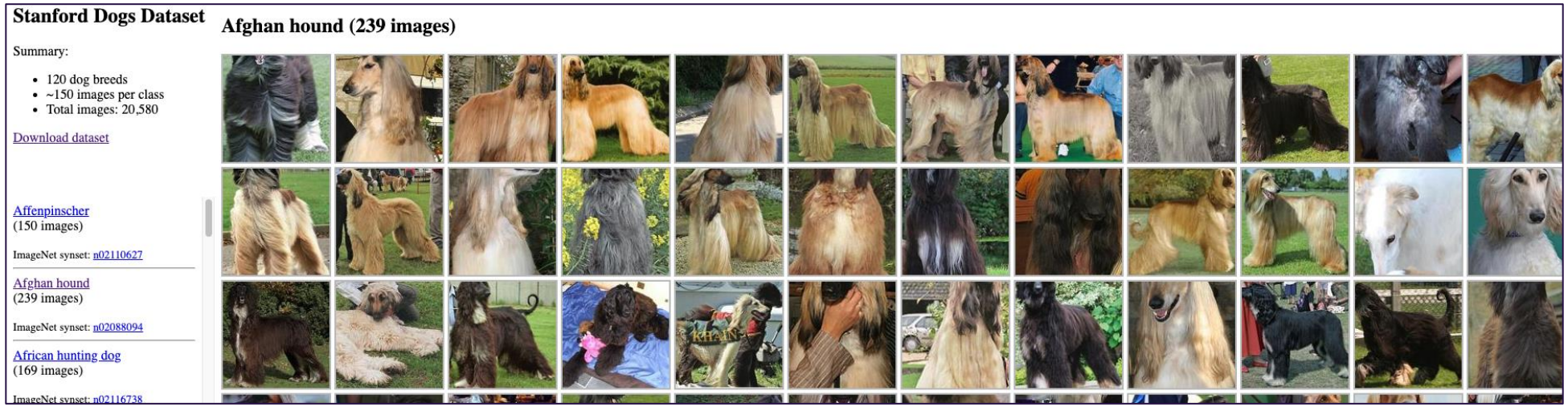
Suppose you have collected data about how far you can travel on a tank of gas and how much each refuel costs.

What will this data help you to predict for an upcoming trip?

Is this prediction an instance of classification or regression?

Polling question

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



Suppose we want to train an algorithm to recognize dachshunds using this dataset. Is this an instance of **supervised** or **unsupervised** learning? Is this an instance of **classification** or **regression?**

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.

Foundational methods

CLUSTERING

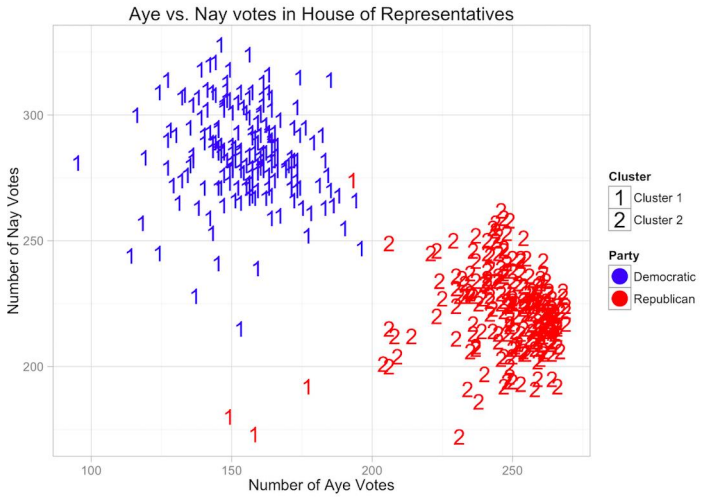
generating labels from unlabeled data

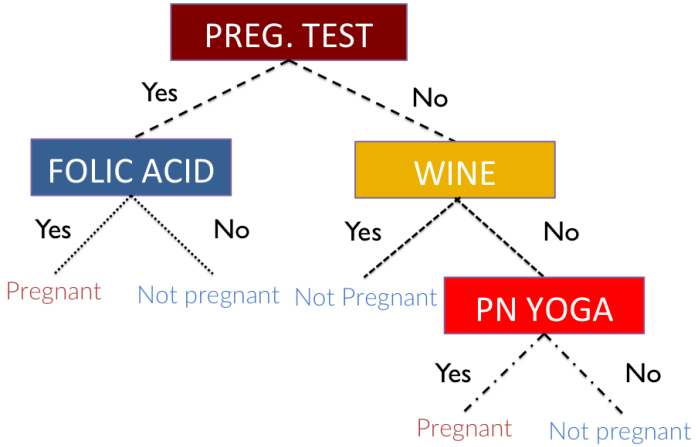
CLASSIFICATION

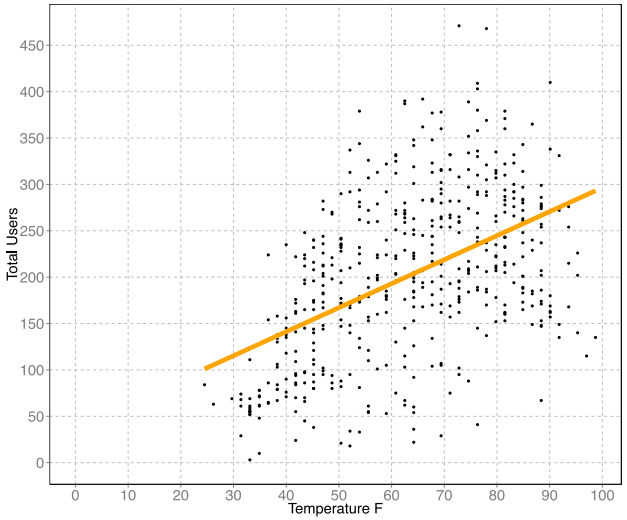
applying labels to novel data points

REGRESSION

assessing / predicting the influence of various factors







Agenda

- Foundational data science methods
- Clustering problems
- Classification problems
- Regression problems
- Working with text data
- Working with network data

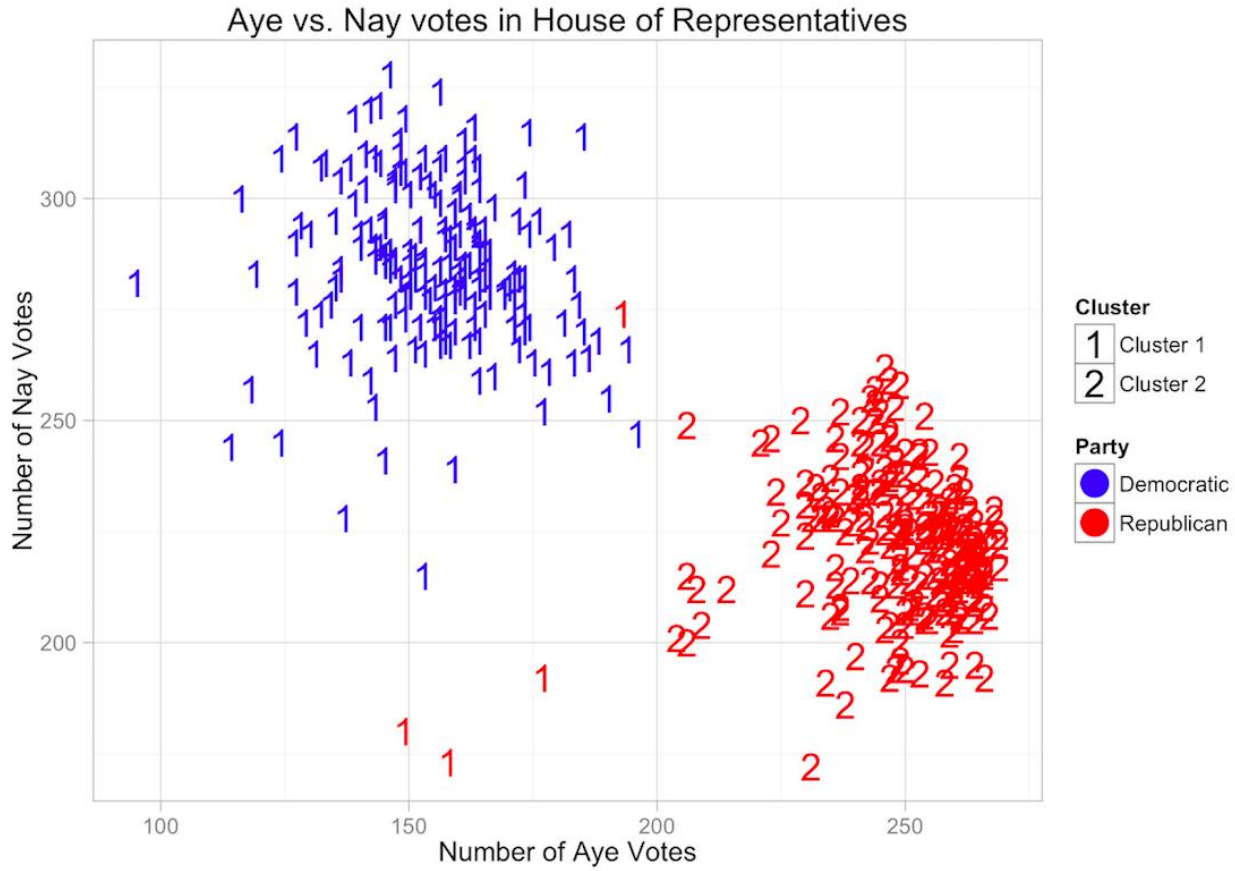
What is clustering?
What kinds of problems can clustering help solve?

Clustering

- Since the group categories themselves are unknown in advance, clustering is a type of **unsupervised** machine learning.

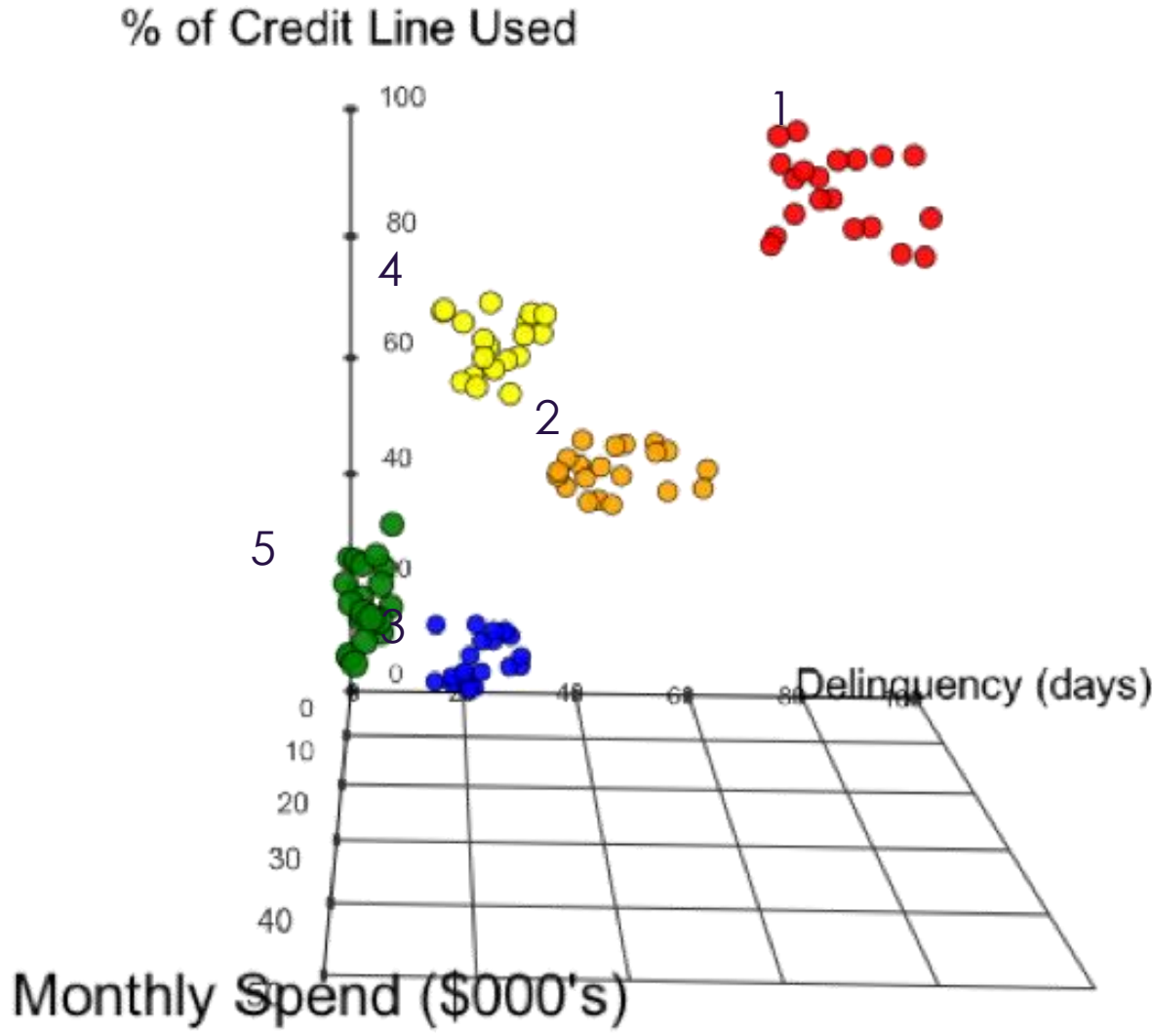
- However, you might find that the generated clusters reveal other interesting information when you apply extant labels to them.

- For instance, Cluster 1 to the right is not exclusively Democrat



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.



How can you use clustering?

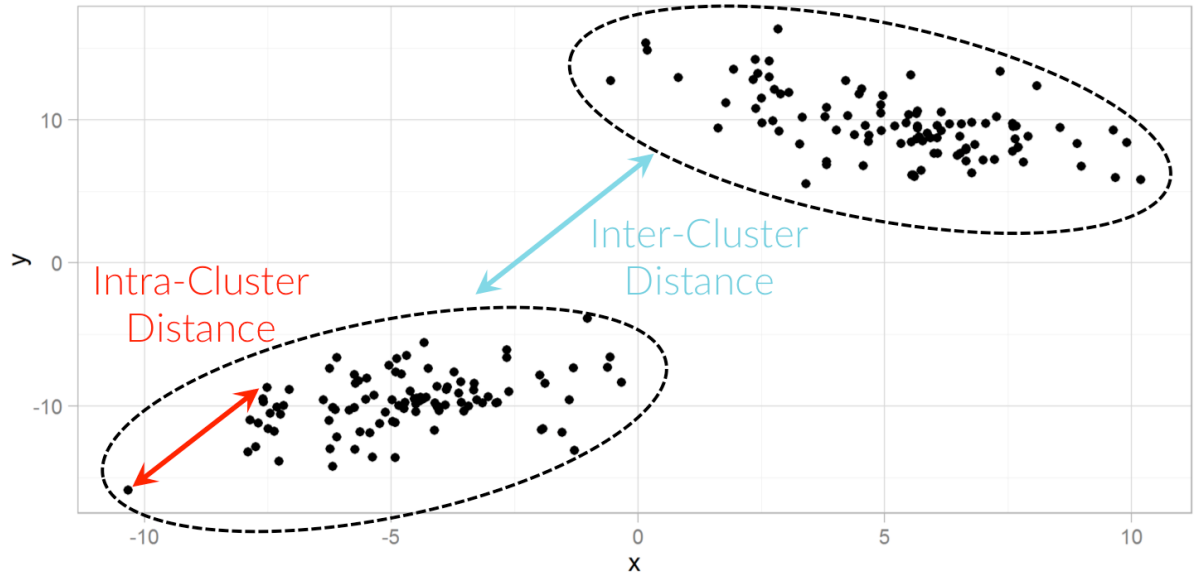
- Clustering answers the questions:
  1. Who/what is this person/object similar to?
  2. Is there a hidden pattern in the data that we can't see?
  3. Are there groups of data with similar attributes?

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

Evaluating the accuracy of the model

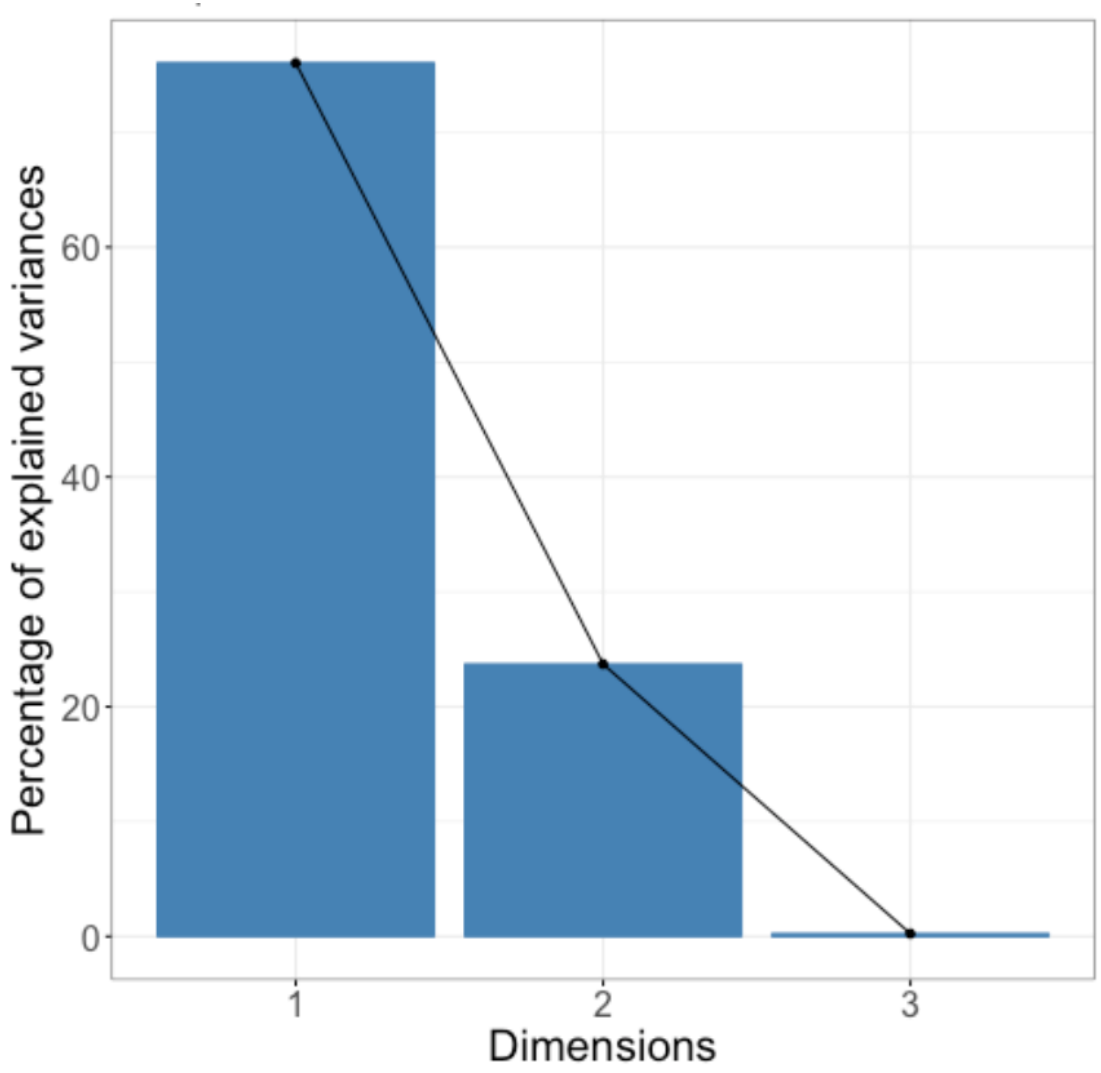
- Goal of clustering is to maximize the separation between clusters and minimize the distance within clusters
- The ratio of inter-cluster variance to total variance can help you assess the performance of algorithms, although this is dependent on the model you use

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



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

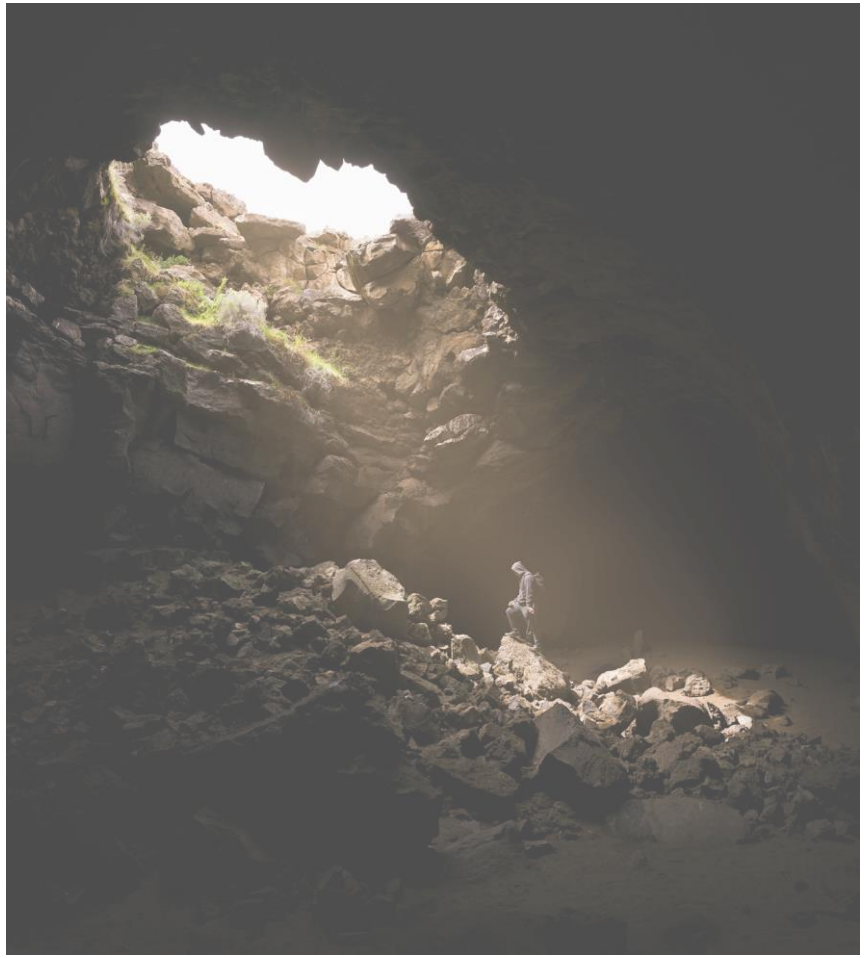
1. How was the distance measure identified?
2. Did you scale the data appropriately?
3. How many clusters do you expect or want? Why?
4. Does your algorithm scale to the size of the data?
5. What can we learn from the groups that the algorithm identified?

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 

Break

Agenda

- Foundational data science methods
- Clustering problems
- Classification problems
- Regression problems
- Working with text data
- Working with network data

What is clustering?
What kinds of problems can clustering help solve?

Chat question

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

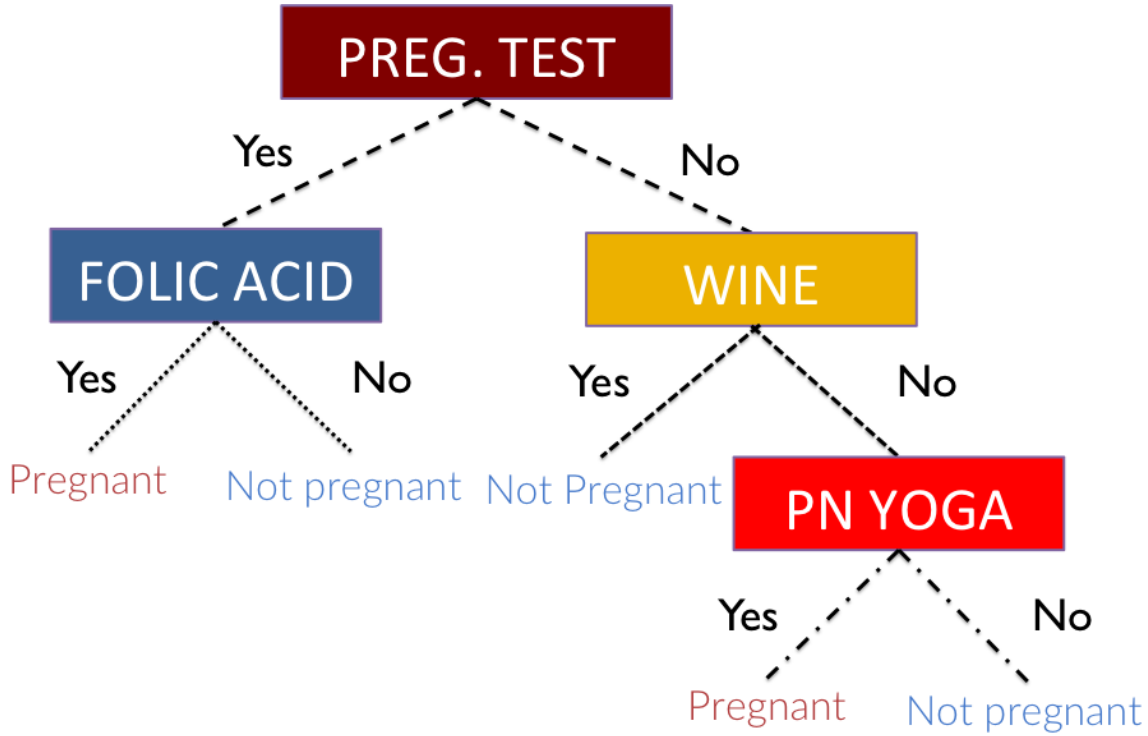
"Out of all the web traffic to our website, how can we tell which visitors are bots and which are humans?"

A classification problem

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

Classification

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

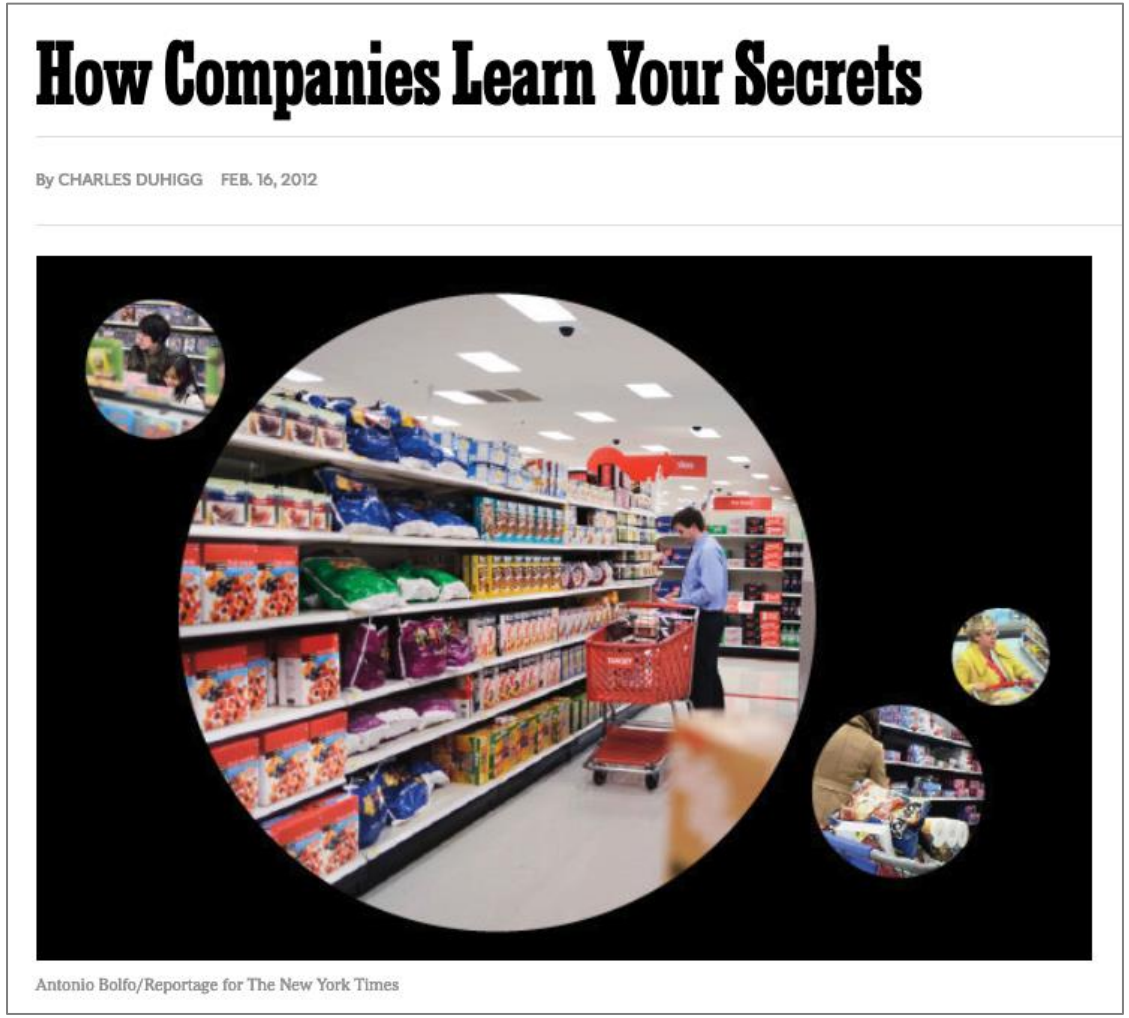


Example: predicting pregnancy

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

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

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



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

Chat question

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

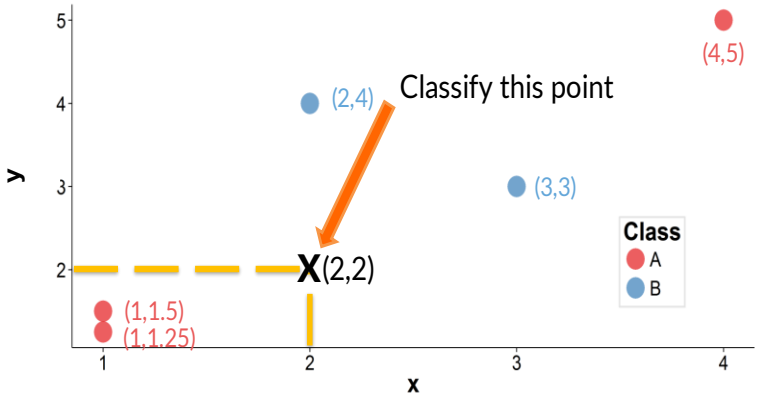
How can you use classification?

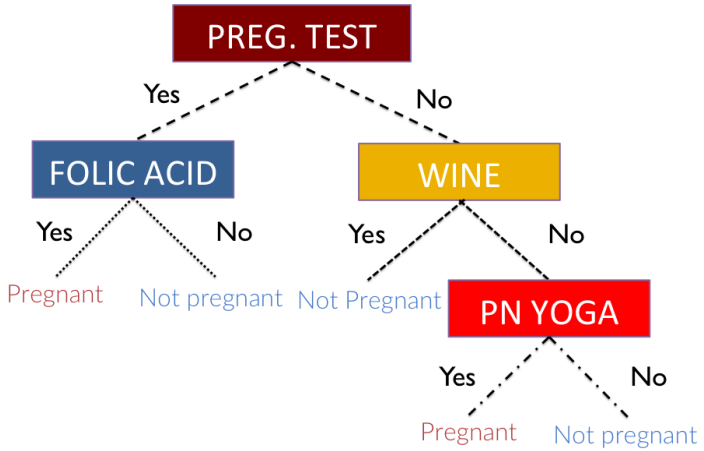
- Classification answers the questions:
  1. Which is the probability of an object / person being in a particular group?
  2. What category is this person / object in?
  3. What is this person / object most similar to?

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

Common classifiers

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





Evaluating the 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.

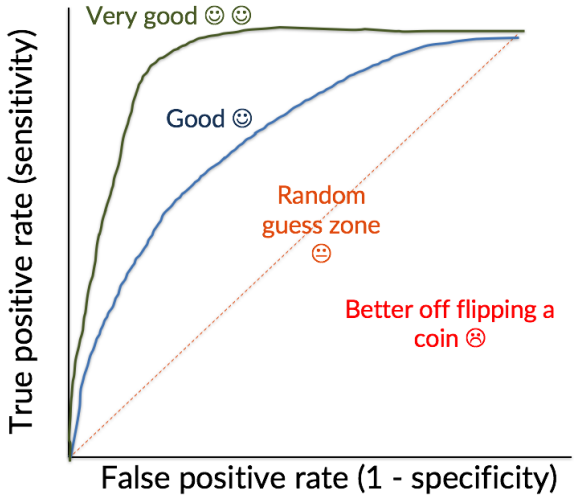
- Then, compare the outcomes that the model produced to the actual outcomes to determine how accurate your model is, and how well it generalizes to new data.

- 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

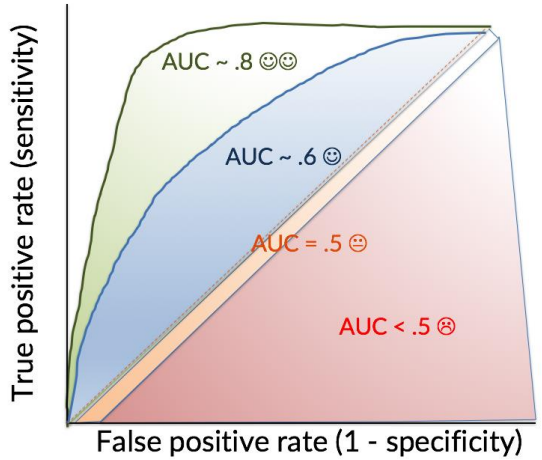
Evaluating accuracy, ctd.

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



- Finally, calculate the AUC (area under curve), to compare different models.

- An AUC above .5 is better than a random guess.



Questions to ask about classification

- How did you determine the threshold between categories?
- On what data did you train the model?
- How did you split the data into a training and test set?
- What thresholds did you use for ROC and AUC?

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 ☑

Agenda

- Foundational data science methods
- Clustering problems
- Classification problems
- Regression problems
- Working with text data
- Working with network data

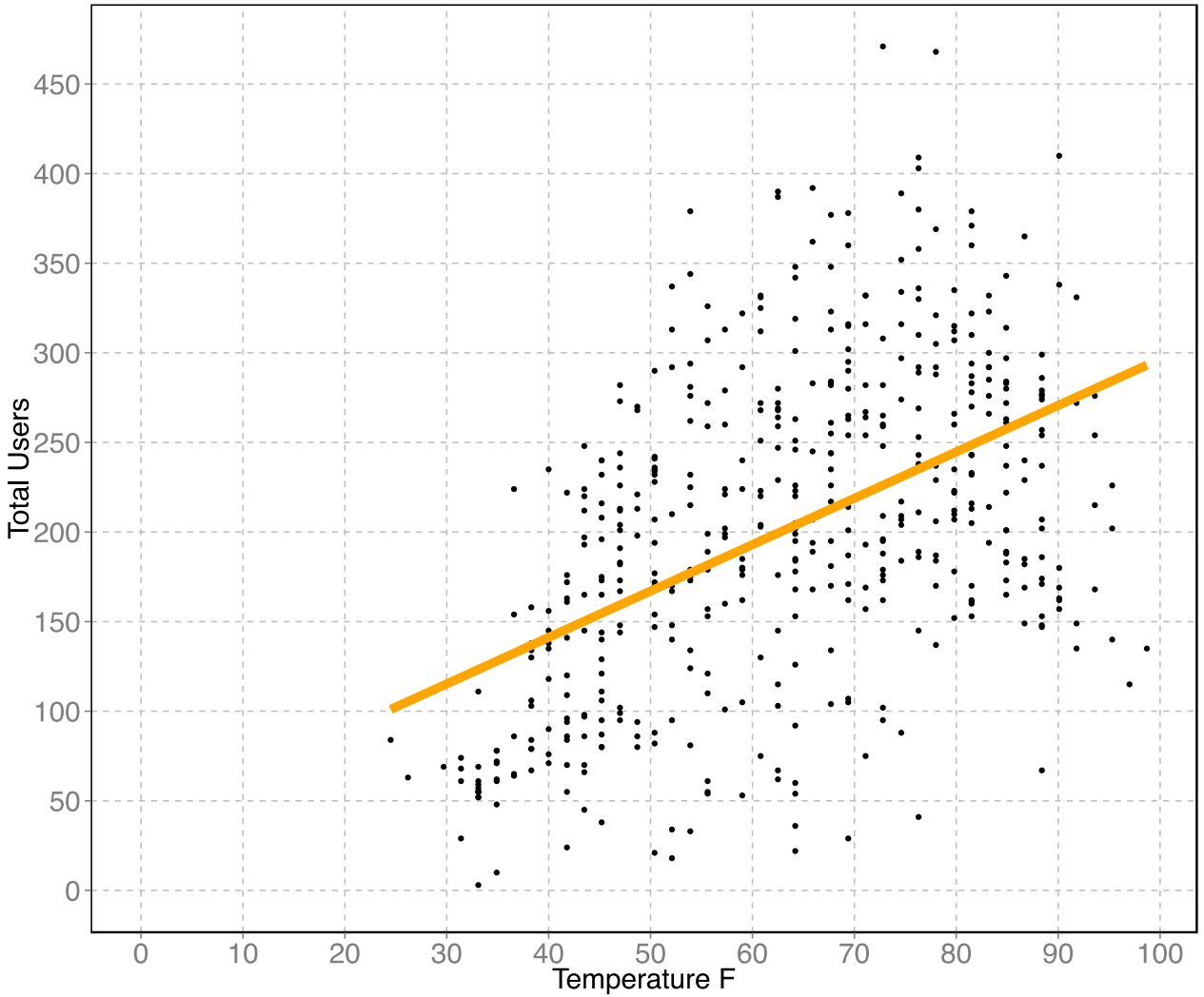
What is regression?
What kinds of problems can regression help solve?

A regression problem

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

Regression

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



Use case: predicting city movements

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

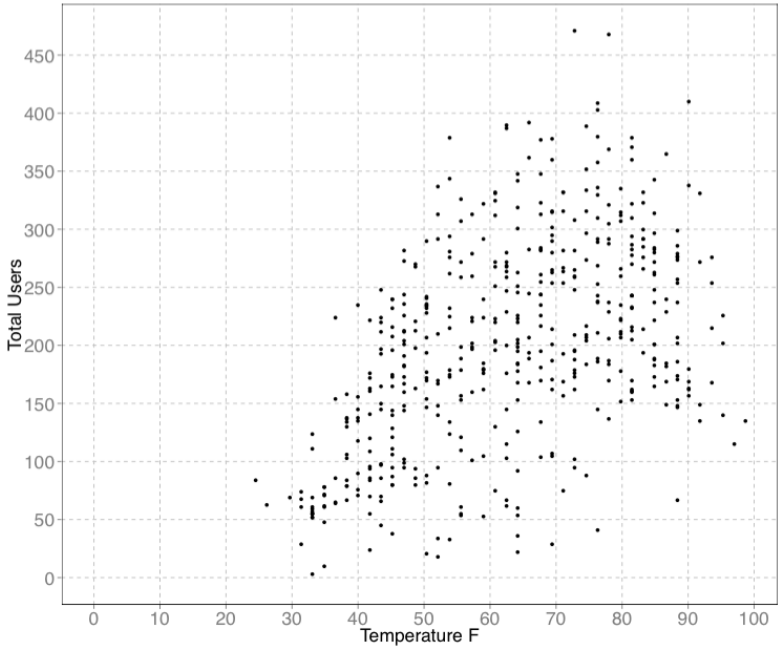
Chat question

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

Simple linear regression

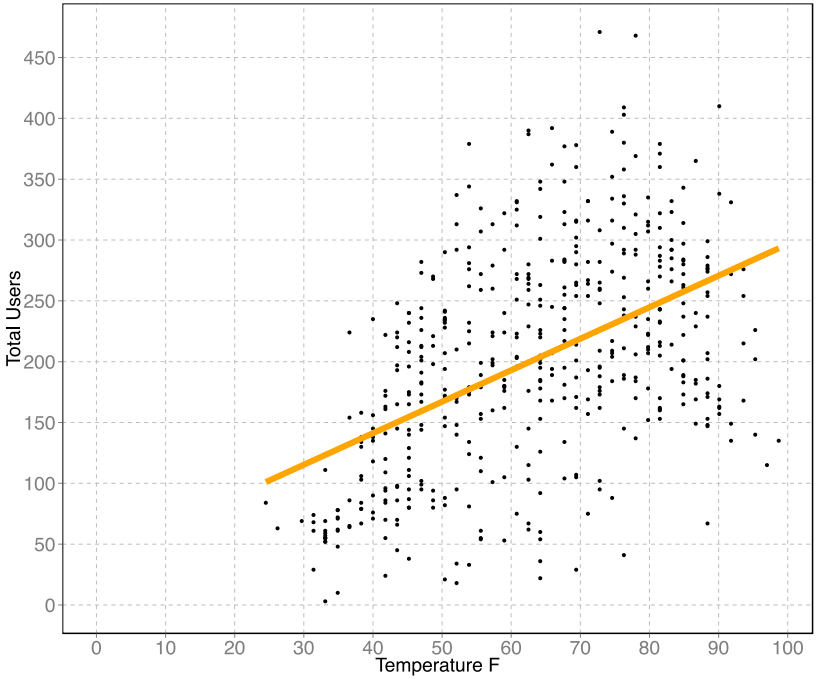
1. Gather data on variables in question
2. Plot the data
3. Draw the line to best fit the data
4. Evaluate model performance

- Measure error
- Deal with outliers
- Determine accuracy



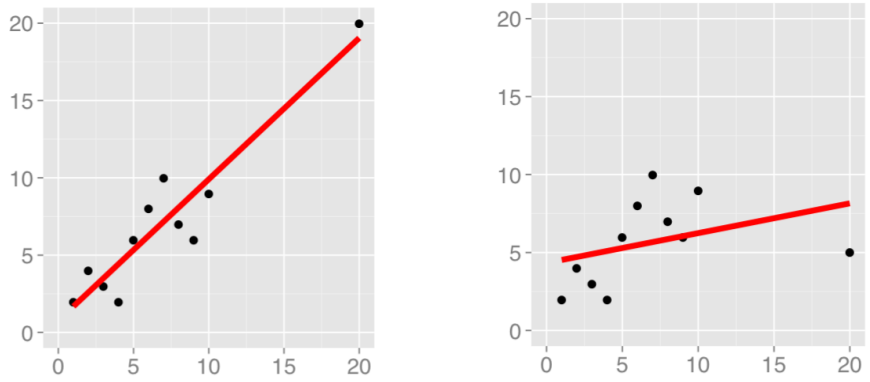
$y = mx + b$

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

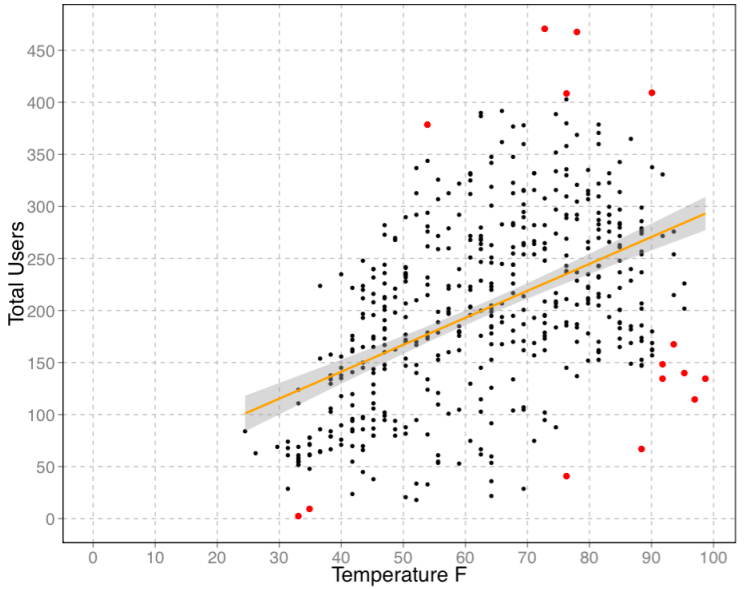


Deal with outliers

- 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"

How can you use regression?

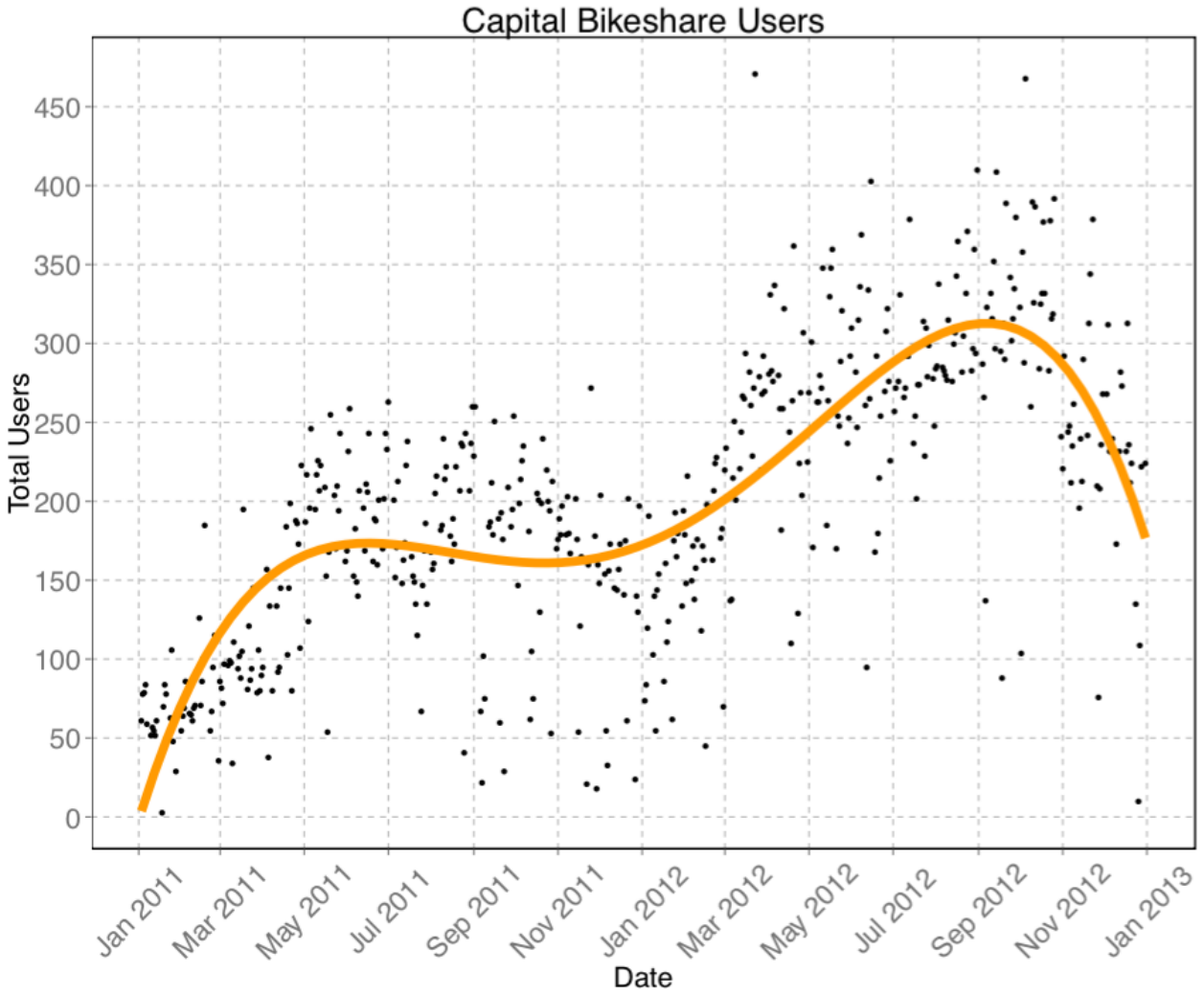
- Regression both explains predictors (inference) and enables forecasting (prediction) by answering the following questions:
  1. Which factors matter most?
  2. Which can we ignore?
  3. How do those factors interact with each other?
  4. How certain are we about all of these factors?
  5. What happens to the outcome if we change a factor?

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

Common regression techniques

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

- **Multiple linear regression** tracks several independent variables, but you must account for special kinds of interference.



Questions to ask about regression

- How well do we understand the underlying data distribution?
- Did you identify any outliers? Were they significant? Did you remove them?
- Are you sure each of the independent variables really is independent? Are we double counting anything?
- What was the R^2 metric?

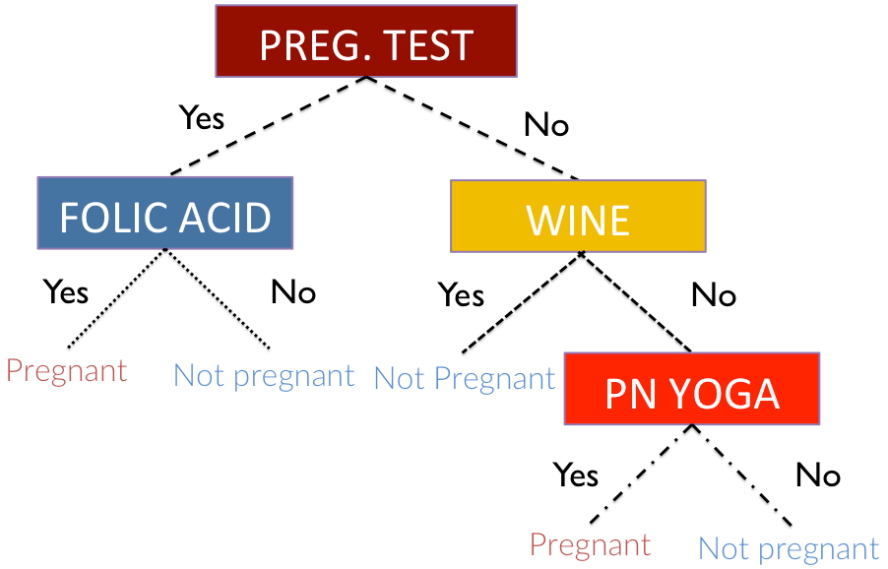
Recap: when should you use regression?

- 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?

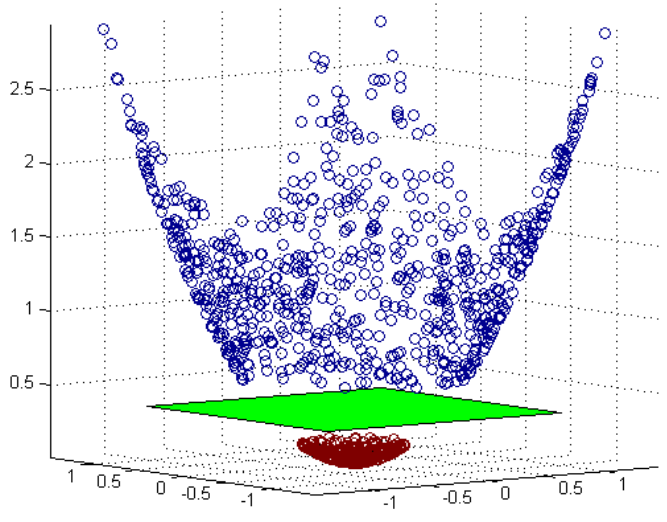


Polling question

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

Polling question

A support vector machine separates data points into classes by using an optimal hyperplane. Is this an example of clustering, classification, or regression?



Break

Agenda

- Foundational data science methods (con-t)
- Clustering problems
- Classification problems
- Regression problems
- Working with text data
- Working with network data

What does a data project look like in practice?
What kinds of considerations are there for working with complex data?

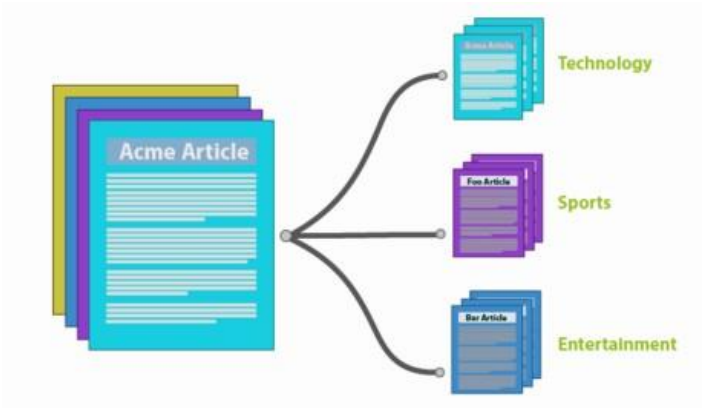
Limitations to foundational methods

- Clustering, classification, and regression are useful for working with a lot of different kinds of data.
- Even more advanced methods ultimately build on these foundations to address datasets with greater complexity or with special considerations.
- Certain types of data are more difficult to reduce to easily processed statistical inputs or might benefit from intense processing power for interpretation.

Advanced methods

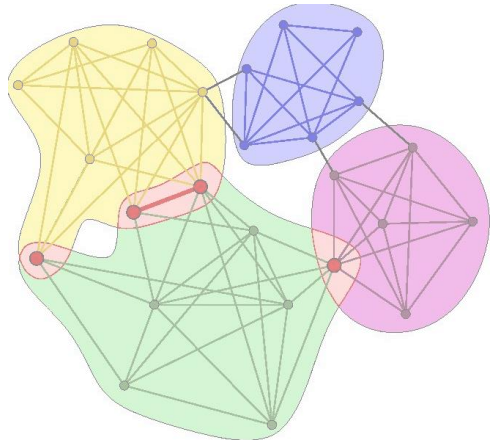
TEXT MINING

inference and prediction based on textual data



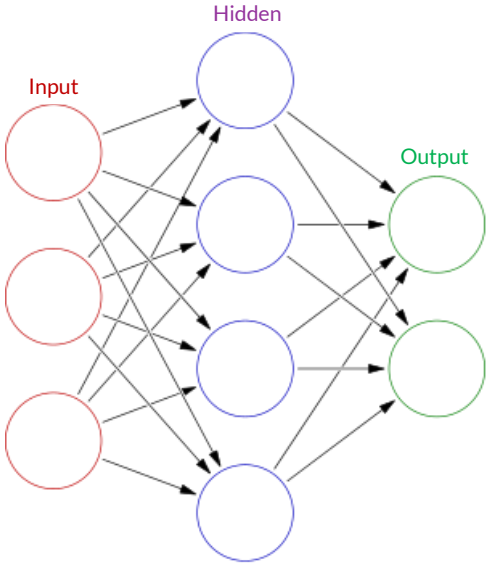
GRAPH ANALYSIS

inference and prediction based on network data



NEURAL NETWORKS

supervised prediction of complex data



Chat question

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

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

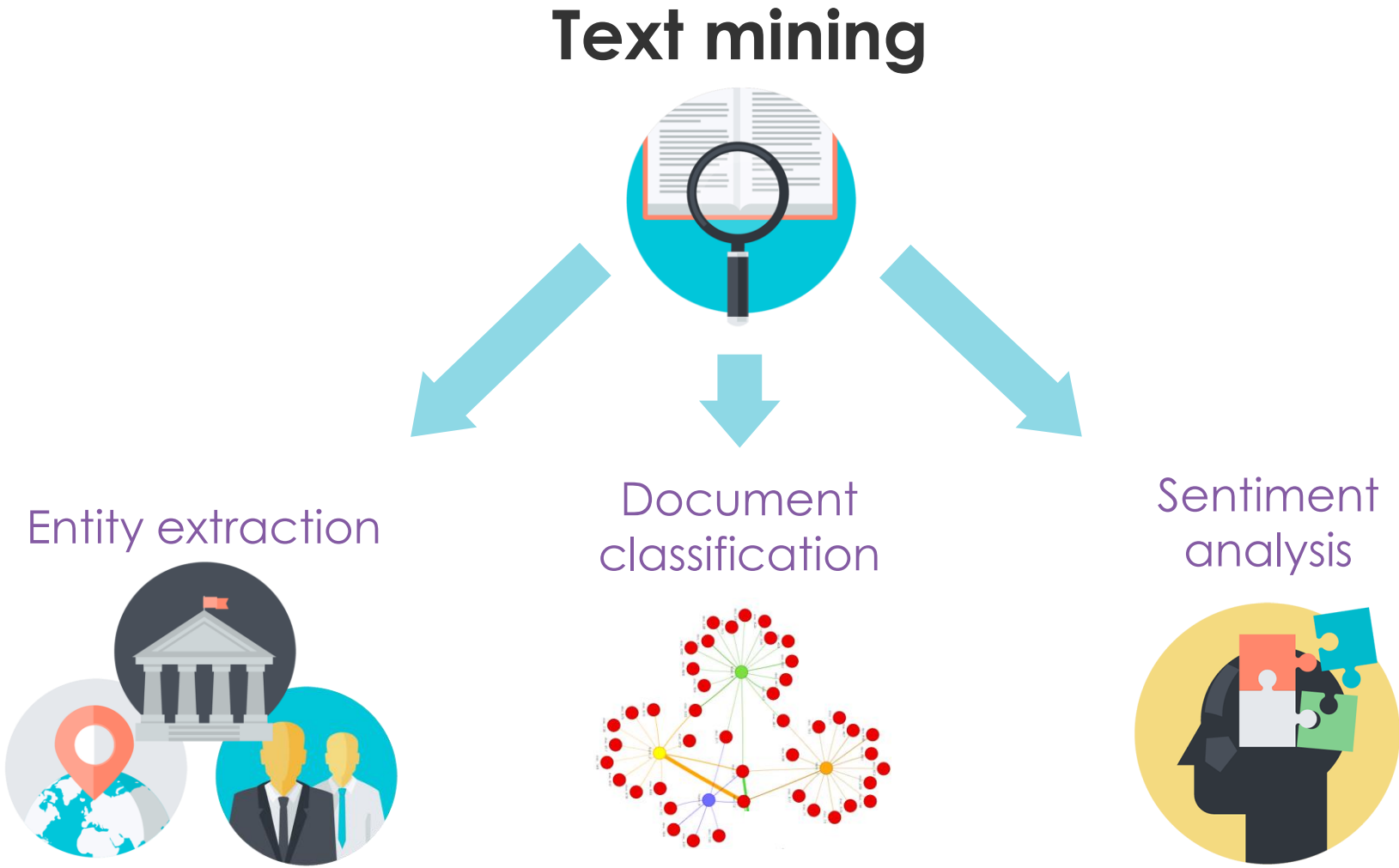
What is text mining?

- Text mining is the process of getting insightful and valuable information out of text data.

- It can answer questions such as:
  - What topics do these papers / articles have in common?
  - What is the sentiment of these social media posts?
  - How are people reacting to an event?

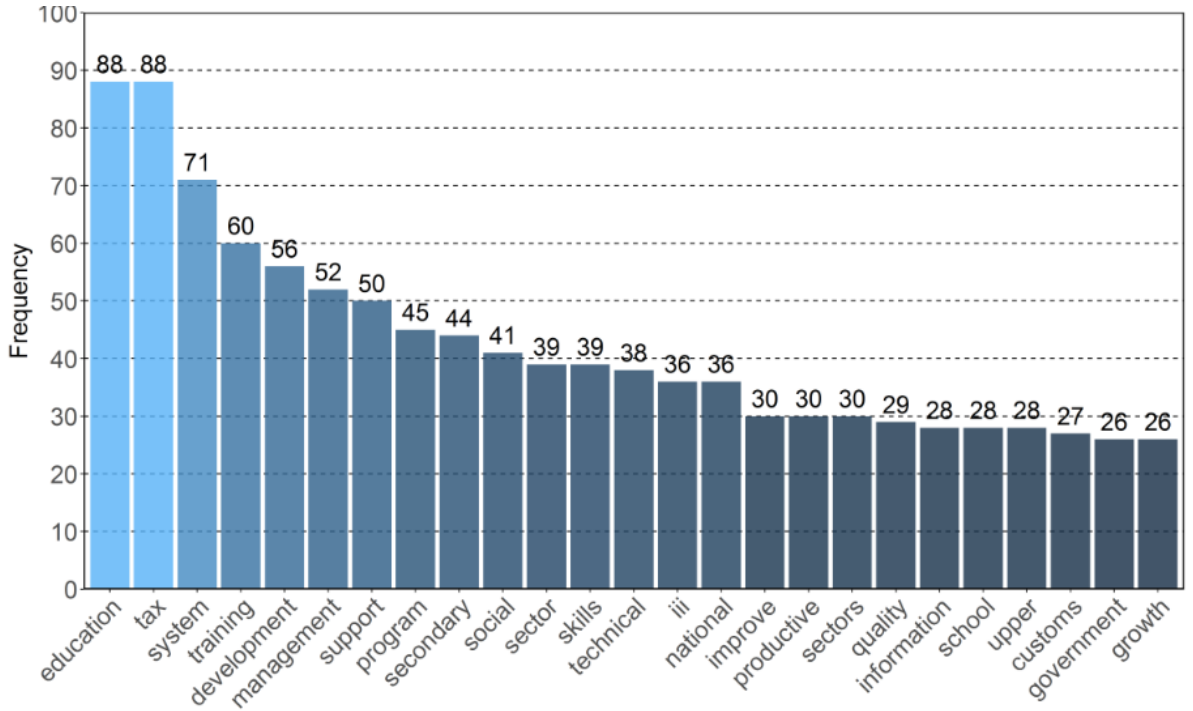
- It employs methods from various fields including mathematics, statistics, computational linguistics, and programming.

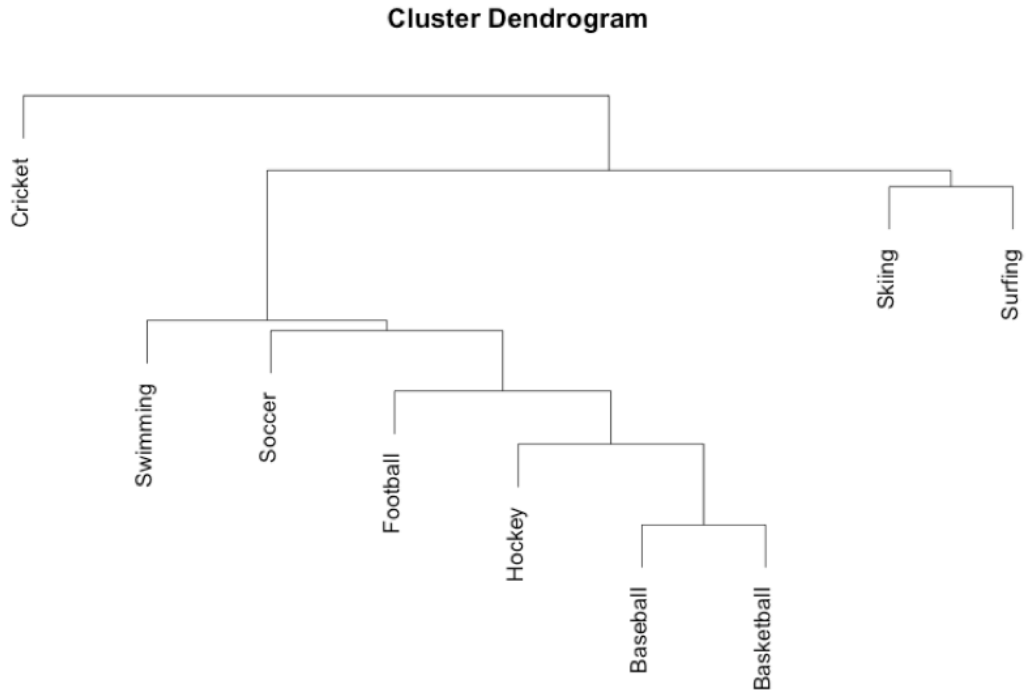
Text mining branches



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.

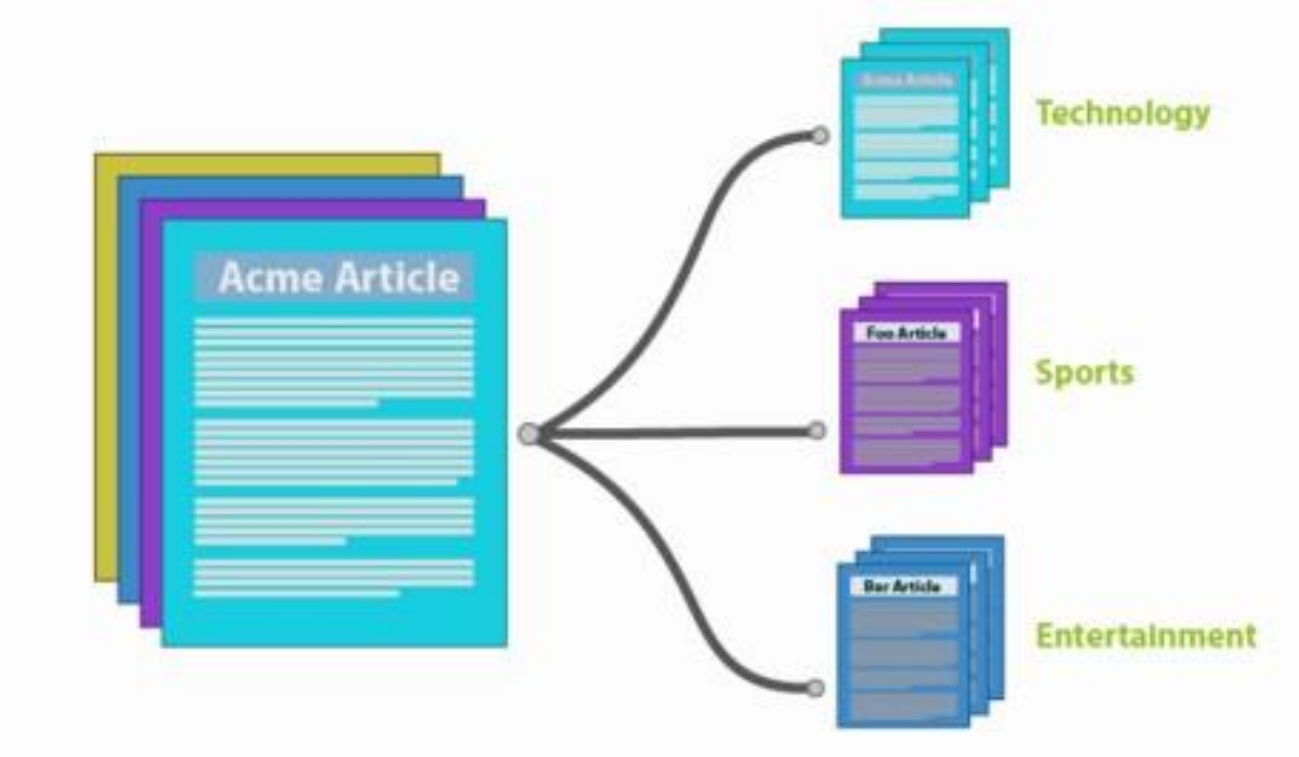




Document classification

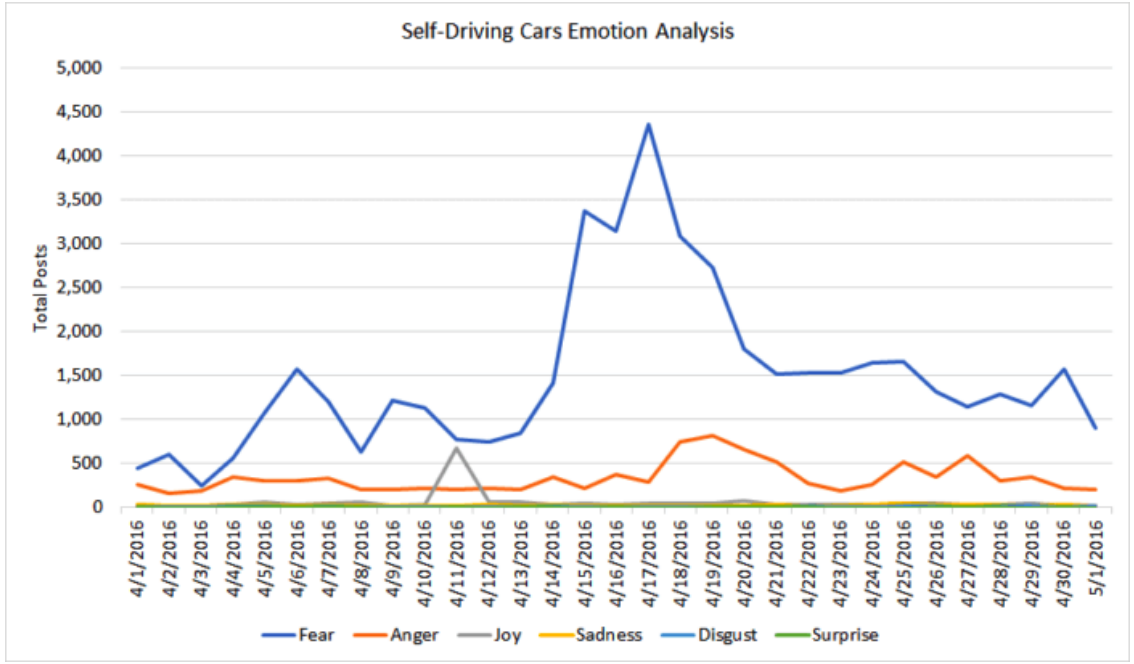
- Use document classification when you want to sort through documents and identify groups o similar articles.

- Based on similarity of topics / other metrics.



Sentiment analysis

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



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

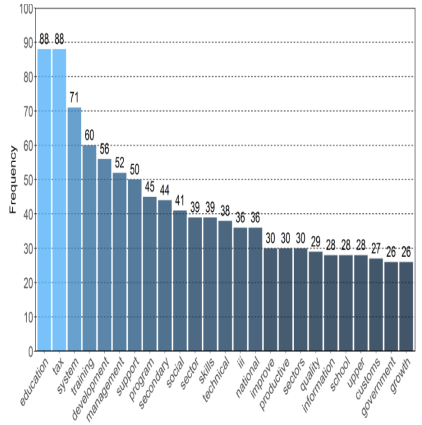
Text mining process

Scrape / collect

Clean & organize

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

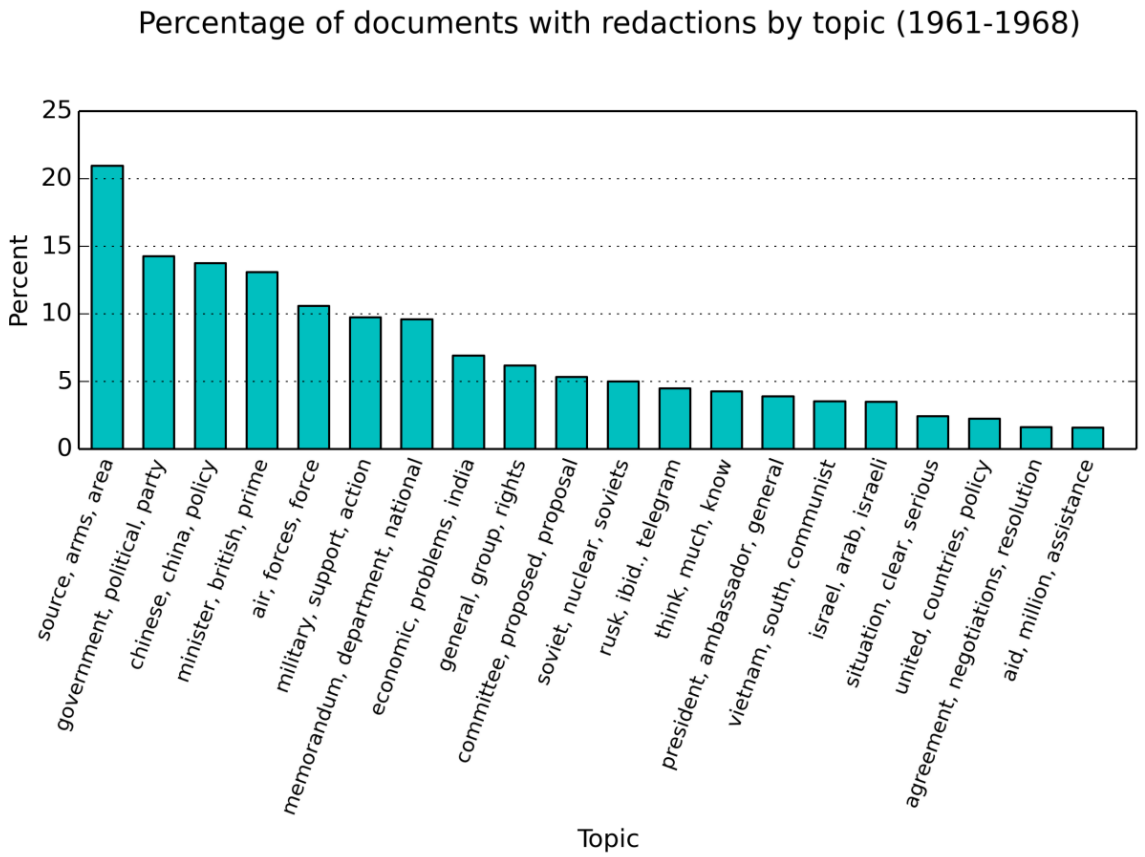
Visualize



Analyze

Example: Topic modeling redactions

- A research collective, History Lab, has digitized over 3 million FOIA documents to begin historical trends.
- One project used topic modeling to determine which subjects were likely to be redacted in a collection of State Department cables.
- They also created a classifier to predict whether a cable was likely to be classified as "secret."



Evaluating quality 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.

Questions you 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?

Agenda

- Foundational data science methods (con-t)
- Clustering problems
- Classification problems
- Regression problems
- Working with text data
- Working with network data

What does a data project look like in practice?
What kinds of considerations are there for working with complex data?

Chat question

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

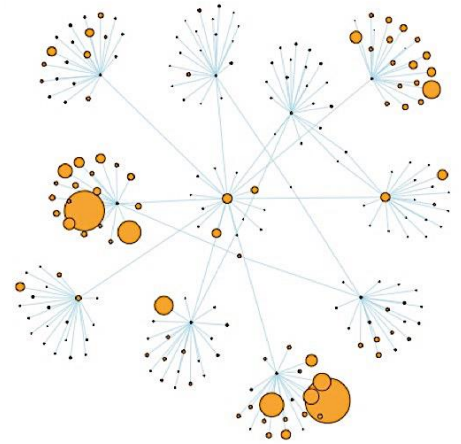
"In the event of a flood in the region, which roads are the likeliest evacuation routes?"

Graph analysis

- Graph analysis (also known as network analysis) seeks to find patterns within a network.

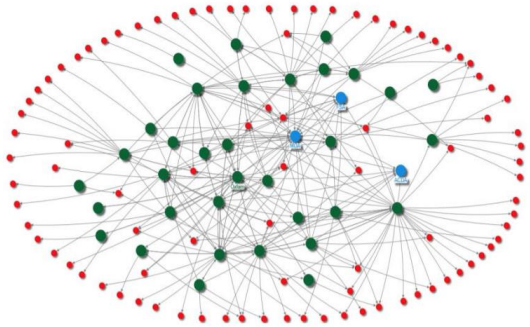
- Networks can represent organizational relationships; communications patterns; economic relationships; environmental relationships; connections based on interests, preferences and similarities; as well as geographic relationships.

- It can answer questions such as:
  - What communities exist within a target population?
  - How will a message / disease spread through a population?
  - Which individuals are most trusted in a community?



Example: Oxfam

- As part of a mid-program evaluation, Oxfam analyzed the formal and informal network structures of NGOs participating in a food security project in the South Caucasus.
- They found that informal information-sharing is much higher and more frequent than information-sharing through formal relationships – even with respect to Oxfam itself
- Analyzing informal networks helped them determine where to direct aid.

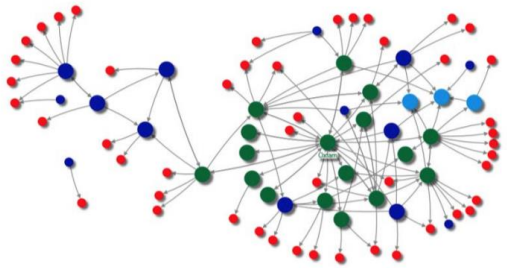


Georgia: Informal network sharing

DENSITY: 0.017
Organizations: 130 **Links:** 290
DEGREE CENTRALIZATION INDEX:
17.0% (IN), 22.5% (OUT)

DENSITY: 0.019
Organizations: 82 **Links:** 126
DEGREE CENTRALIZATION INDEX:
6.83% (IN), 21.83% (OUT)

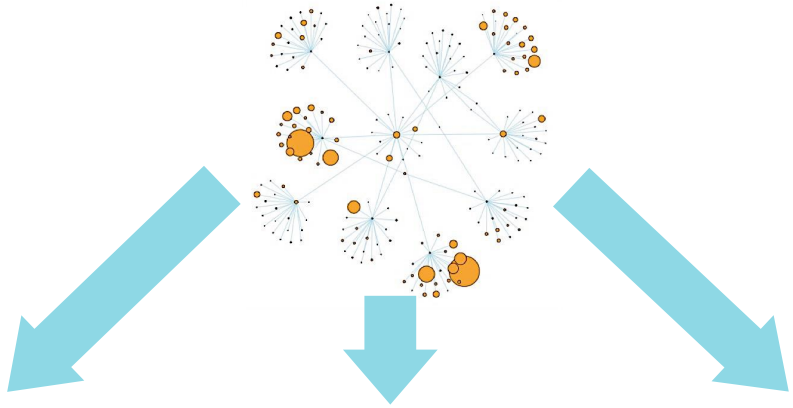
Source: https://oxfamilibrary.openrepository.com/bitstream/handle/10546/620117/cs-social-network-analysis-south-caucasus-241016-en.pdf?sequence=1&isAllowed=y



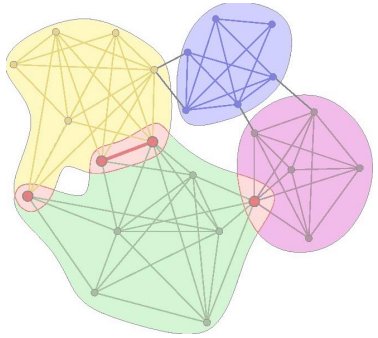
Georgia: Formal network sharing

Types of graph analysis

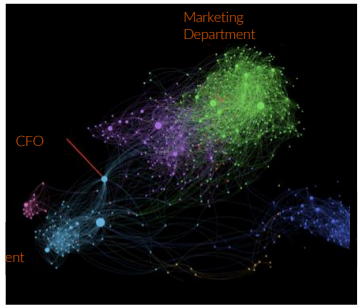
Graph analysis



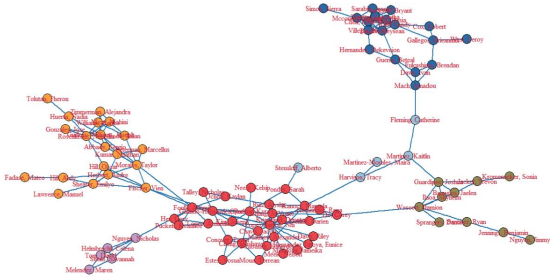
Community detection



Centrality metrics



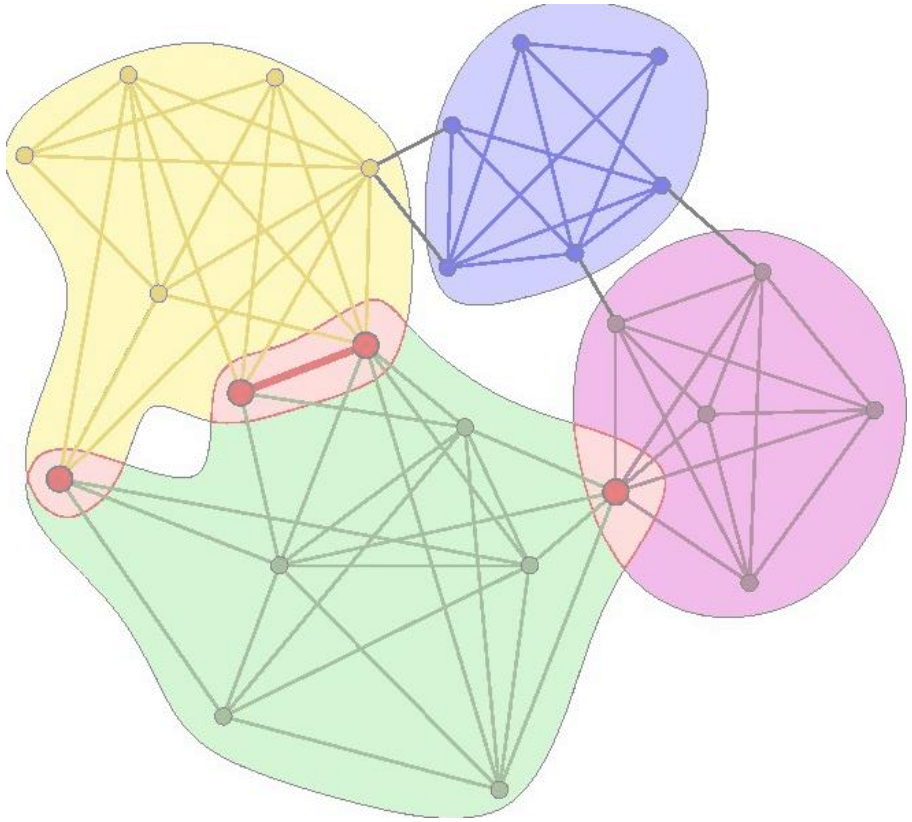
Social media



Community detection

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

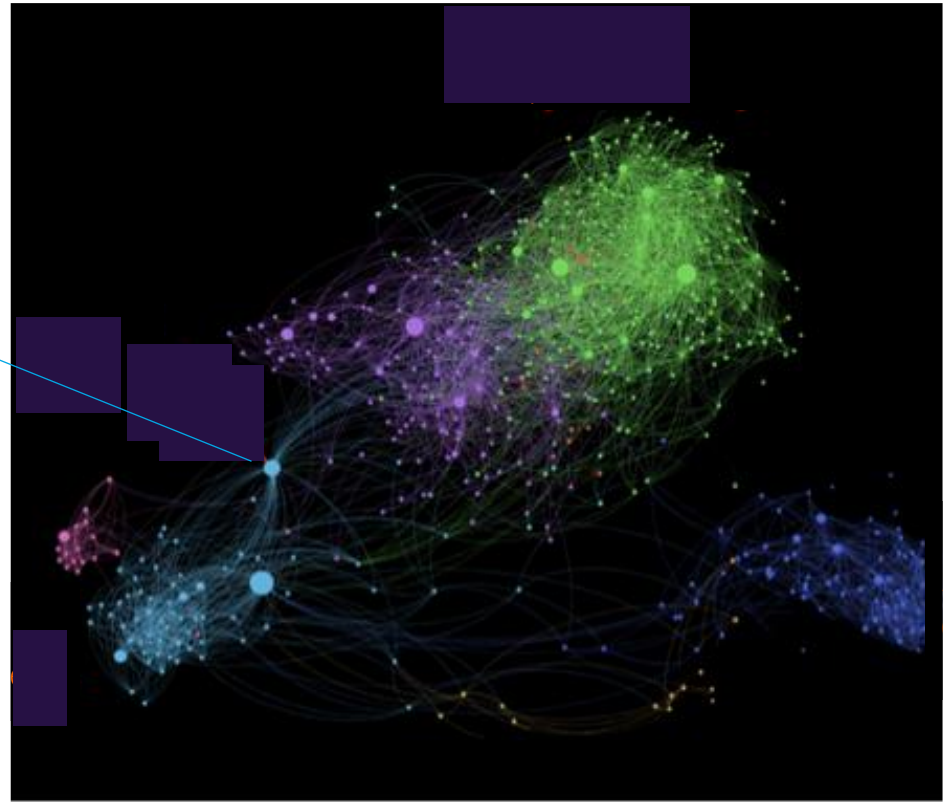
- Identifies groups of individuals / nodes that belong together; can detect latent connections and communities.



Centrality metrics

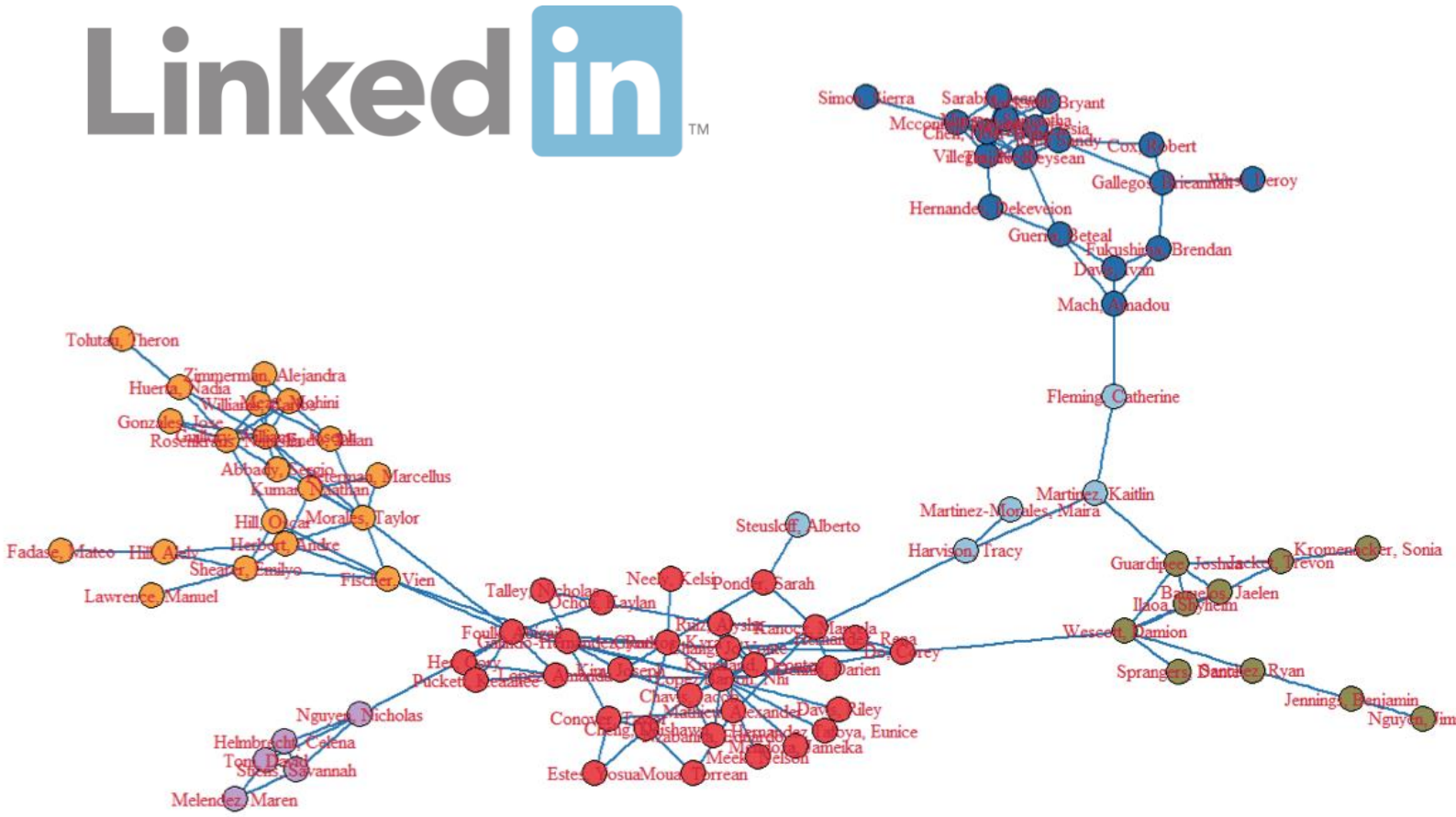
- Use centrality metrics when you want to look at an overview of a network and identify key nodes.
- Identifies the most important nodes, most central nodes, shortest paths, etc.

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.



Evaluating quality 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.

Flowminder in Nepal

- In April 2015, a magnitude 7.8 earthquake rocked Nepal.
- Non-profit Flowminder partnered with Ncell to track population movement through anonymized SIM card data from 12 million mobile phones.
- The maps they generated were implemented by the United Nations and other relief agencies within 2 weeks, accelerating the process of administering aid by a factor of months.

What concerns do you think this project raised?

Questions you should ask

1. 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)?

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

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

End of Day 3

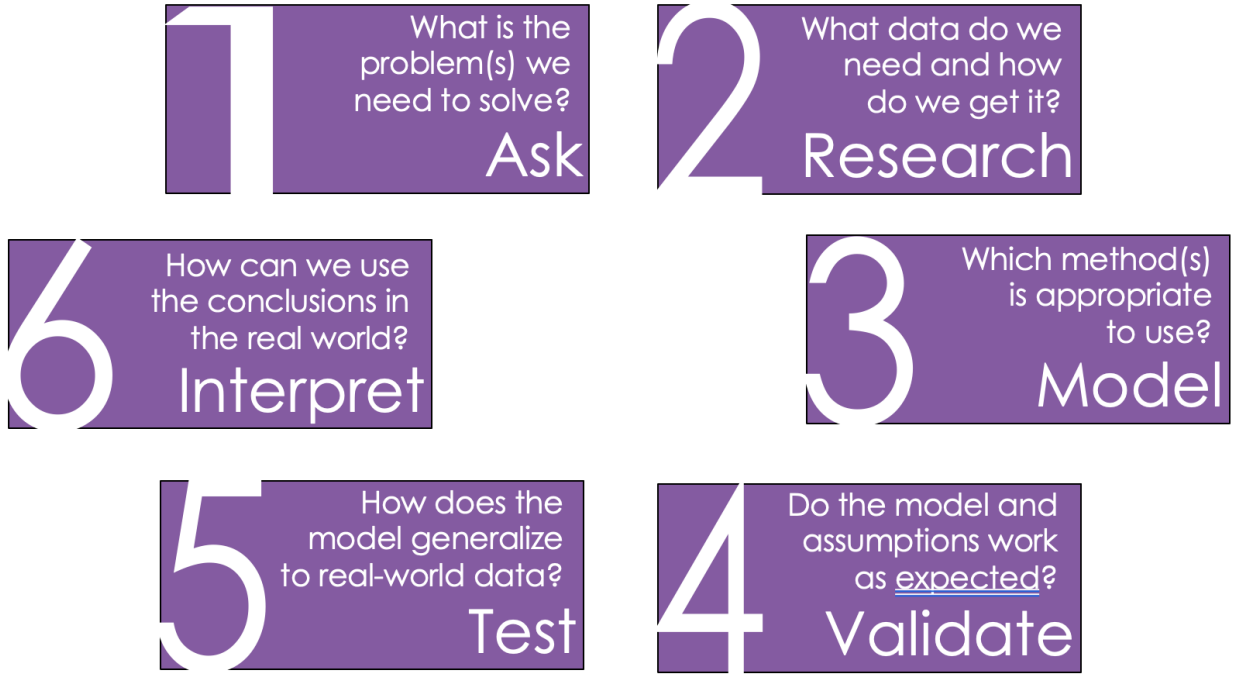
DATASOCIETY:

Fundamentals of Data Literacy

Day 4

Recap

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



Agenda

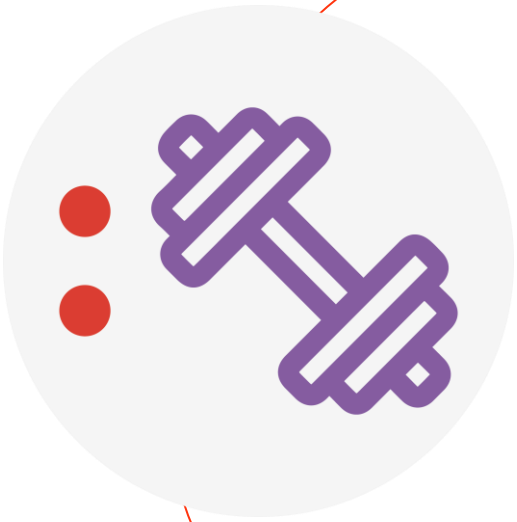
- Neural networks
- Refining your data project
- Intro to data visualization
- Best practices in data viz

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

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.



What are neural networks?

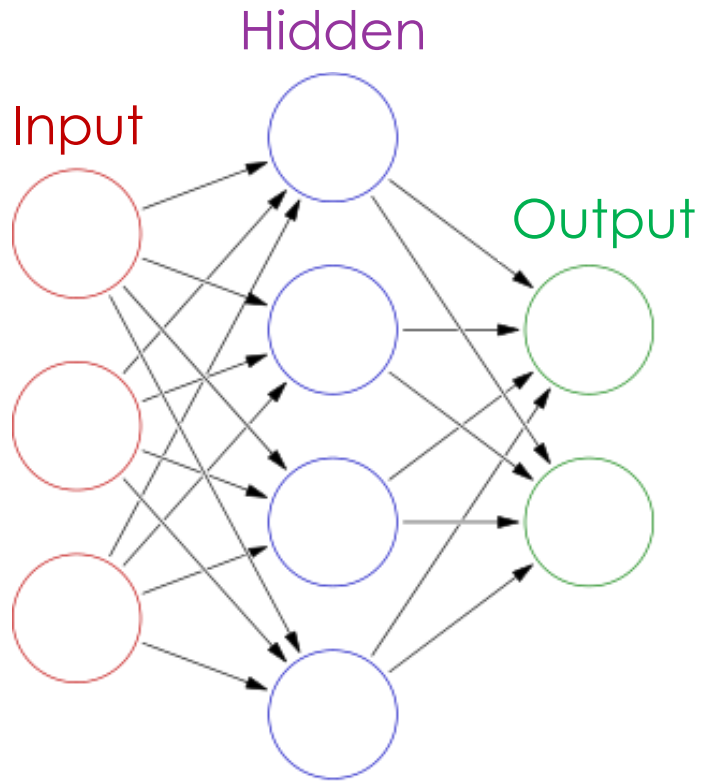
- A neural network is born ignorant and builds on itself to get smarter and smarter.
- It starts out with a guess, and then tries to make better guesses as it learns from its mistakes.
- Neural networks cover the same topics that we've reviewed previously. In theory, you can apply them to almost any 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.

How machines learn

Chat question

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

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

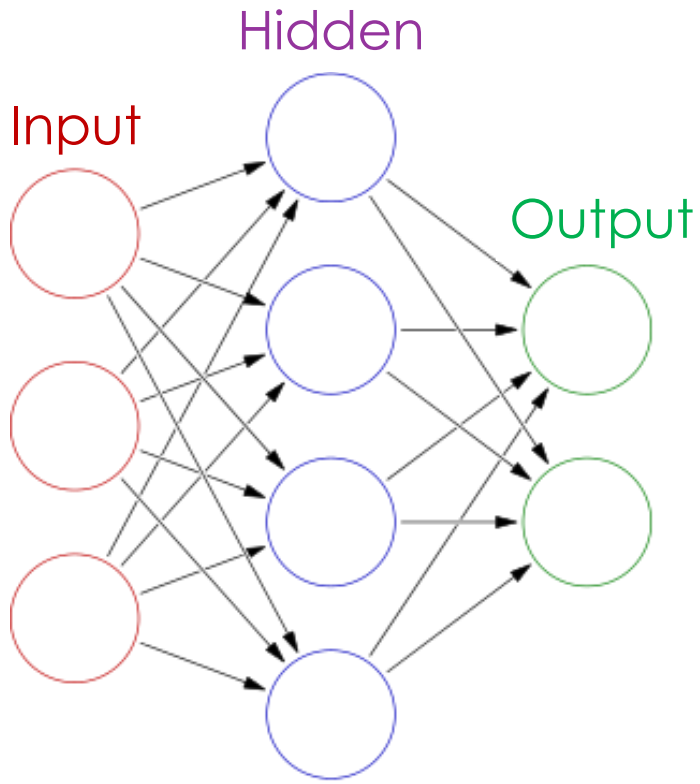
Does that mean you're a good (or bad) artist?

Potential pitfalls

- Machine learning projects are not immune to failure.
- Numerous high-profile cases involving the replication of biases, failure to keep up with real-world circumstances, poorly chosen or badly cleaned input data, and excessive parameter tuning to insert human influence.
- Taking on a data project means being vigilant about where things can go wrong, and taking the steps necessary to resolve conflicts and questions at each point in the data science cycle.

The black box problem

- A data scientist approaches a government agency with an algorithm with very strong predictive power.
- It was generated using a neural network with multiple hidden layers and a nonlinear structure.
- It could replace an algorithm with less predictive power, but whose workings can be explained step by step.



What concerns do you think this project raised?

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.

Agenda

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

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

Chat discussion: data project methods

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

CLUSTERING	CLASSIFICATION	REGRESSION
TEXT MINING	GRAPH ANALYSIS	NEURAL NETWORKS

- Are any of these methods useful to your data project? How might it be modified or enhanced to take advantage of one of these methods?

How to detect a Twitter bot

Chat question

What seems especially difficult about working with Twitter data?

What other questions about Twitter bots would you want to answer?

Break

Agenda

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

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

What is data visualization?

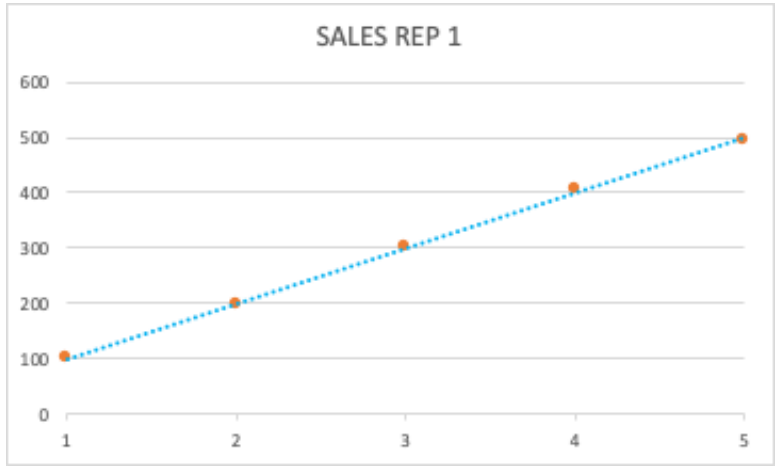
- Data visualization is any attempt to make data more easily digestible by rendering it in a visual context.

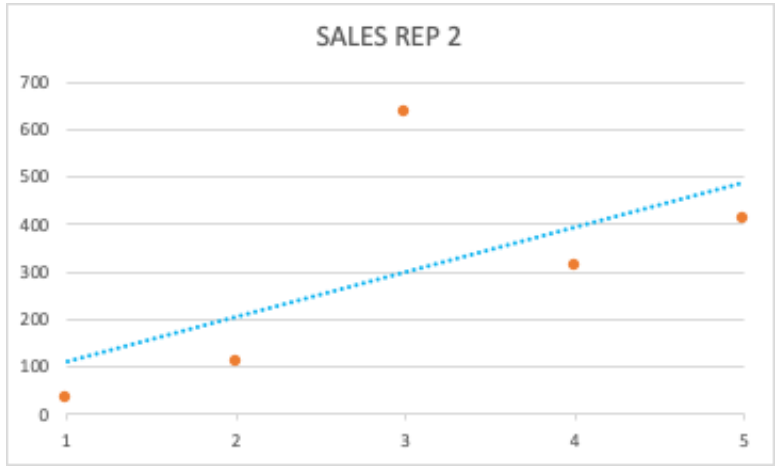
- Common data visualizations include tables, charts, graphs, and dashboards.

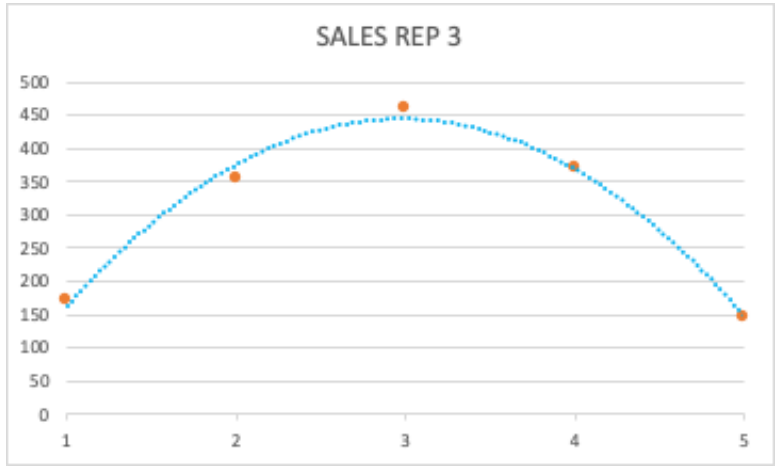
Explore or explain

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

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







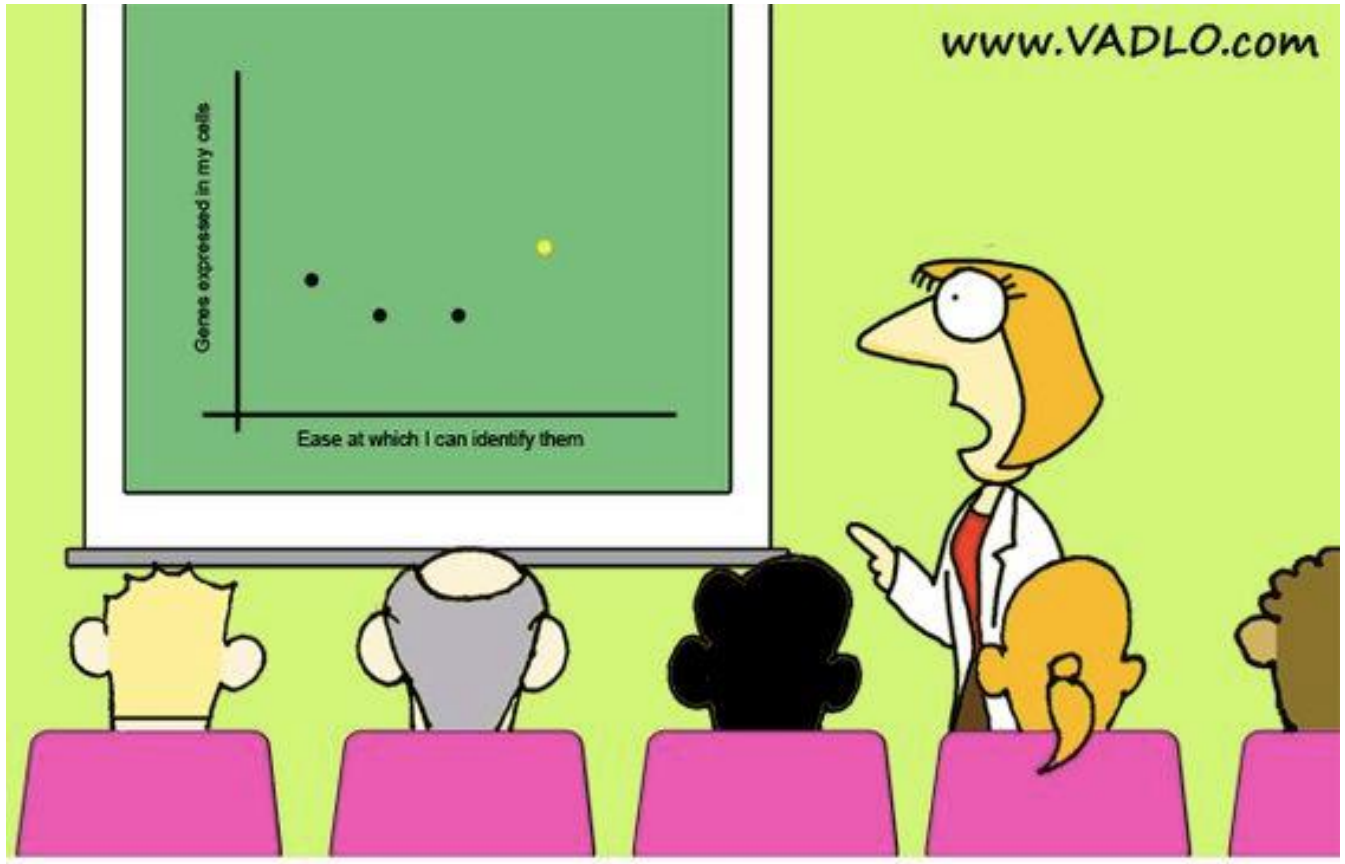
Chat question

Is data visualization an important part of your job?

What types of data visualization does your organization produce?

Getting it right

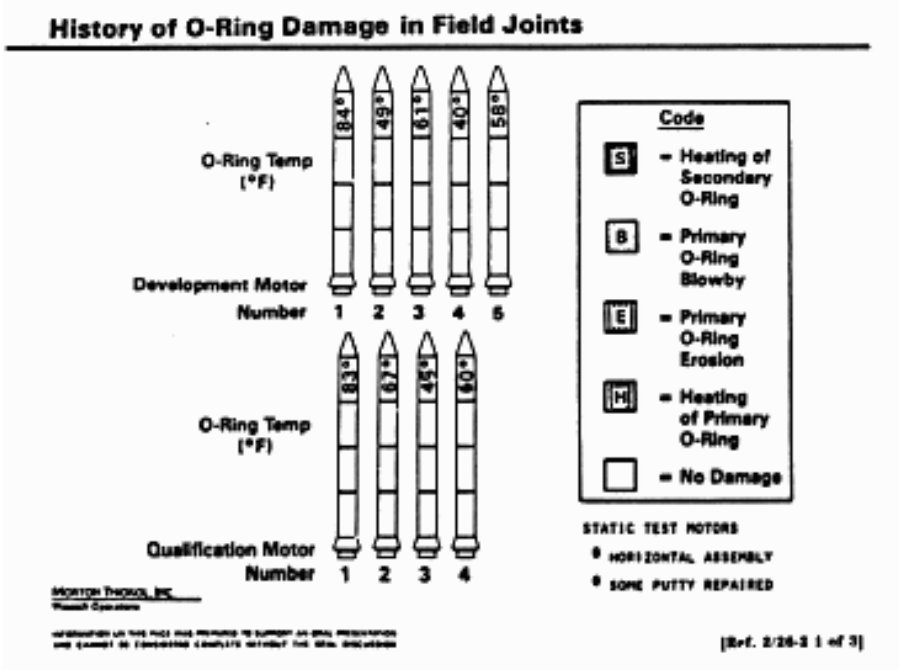
- Using visualizations incorrectly can cause you to lose your audience, lose the value in your data, and ultimately lead to poor decision making.



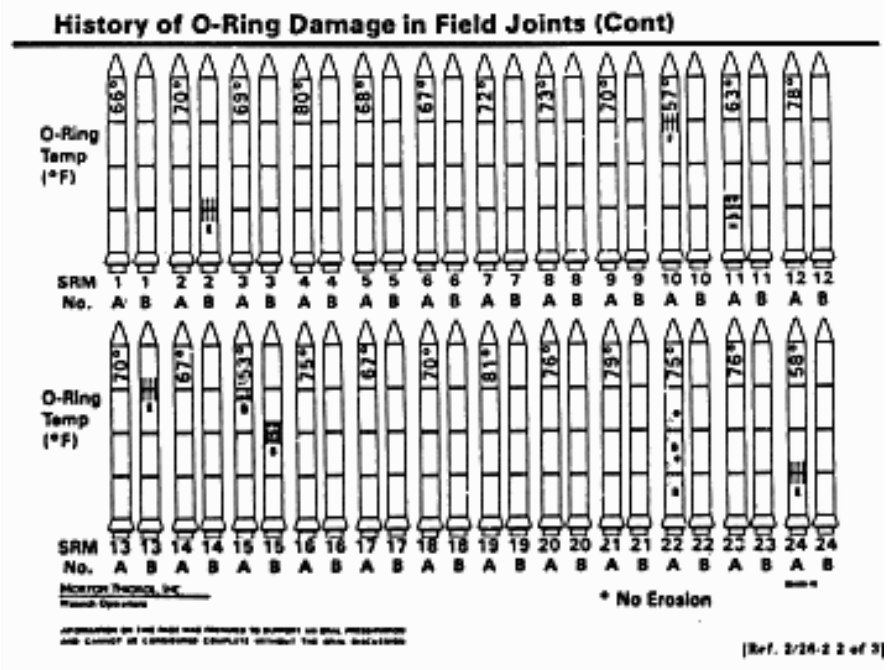
"Same graph as last year,
but now I have an additional dot."

Example: The Challenger

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



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



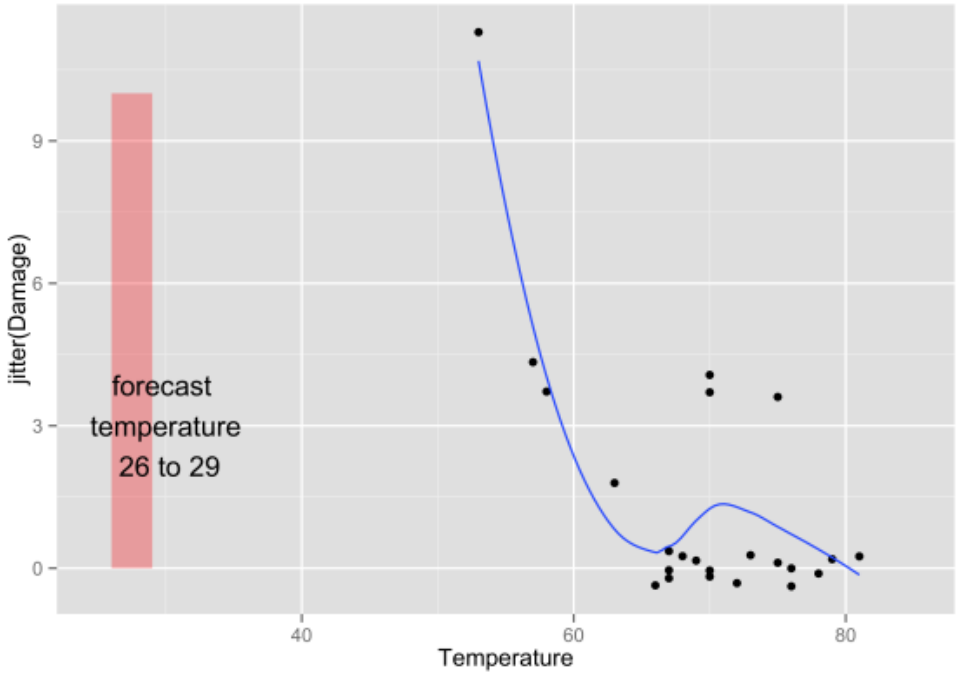
Example: The Challenger

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

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

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

Is it easier to see the issue?



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

Speed-to-insight

- Data visualizations can dramatically change how long it takes the reader or user to process statistical information.

- The extract to the right is from an AFSA article on supposed diplomatic budgetary bloat over the past decade.

How long did it take you to identify the key statistical takeaway?

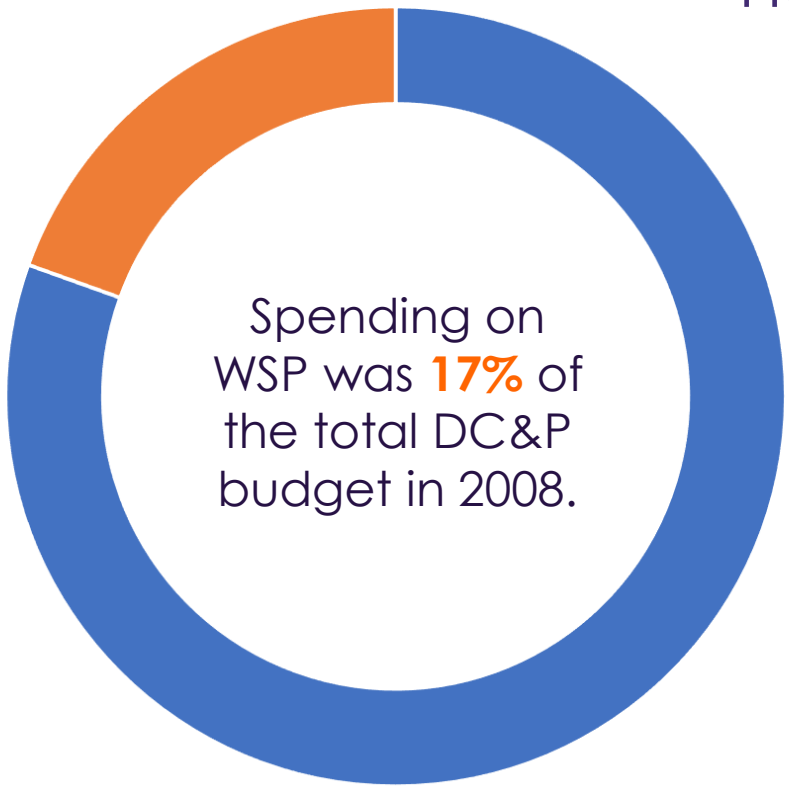
"Yes, the overall budget has increased, with the growth in security costs a major factor. Spending on Worldwide Security Protection was 17 percent of the total 'Diplomatic and Consular Programs' budget in 2008. As the 2018 CBJ shows, by 2016 WSP had grown to 41 percent of the total D&CP budget, while the share for core diplomacy was squeezed to 59 percent.

"The proposed budget for 2018 continues this trend, with WSP growing to 45 percent of total D&CP spending while core diplomacy declines further, to 55 percent."

http://www.afsa.org/does-america-spend-too-much-diplomacy

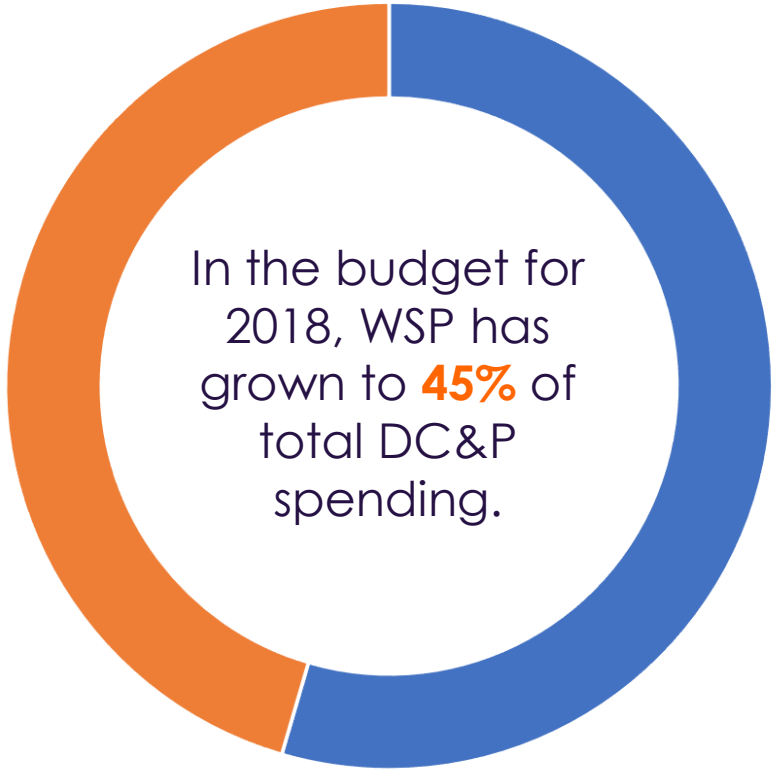
A speedier insight

2008



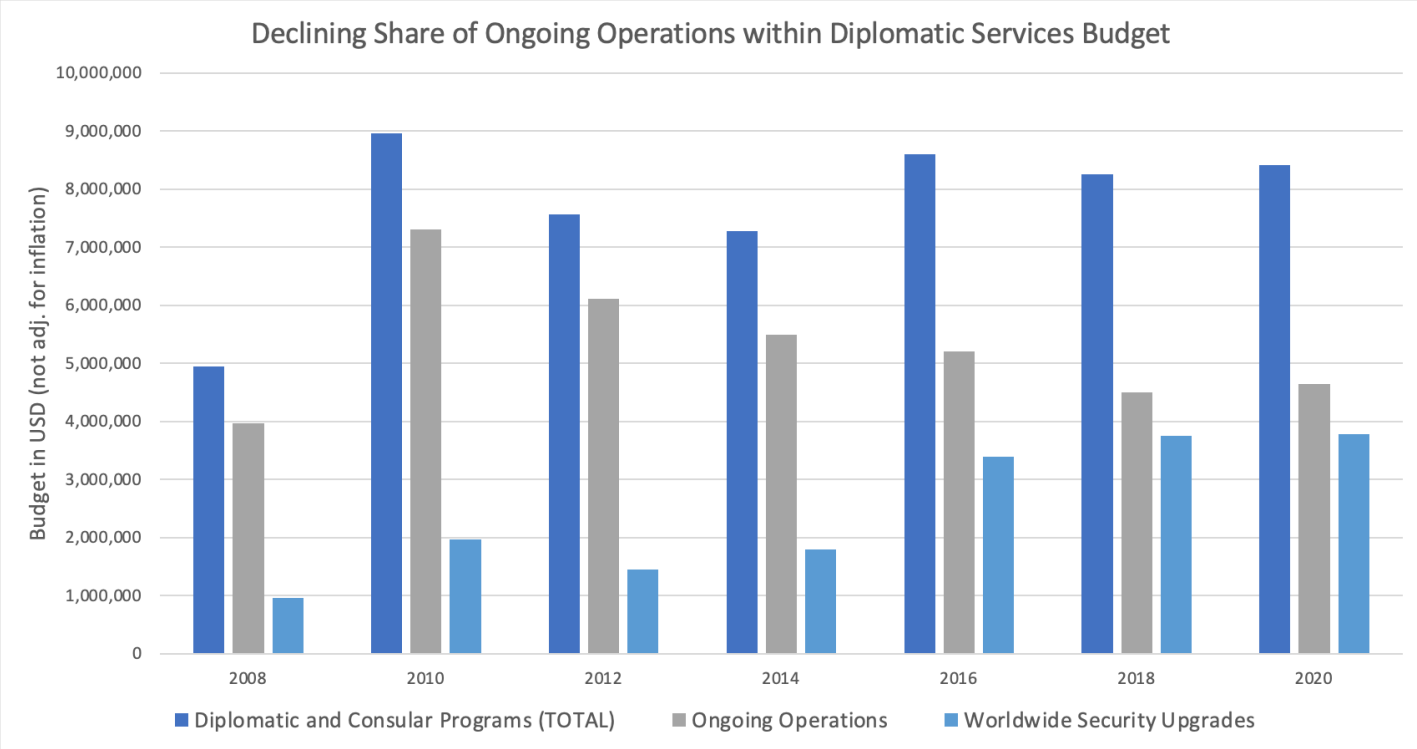
How about now?

2018



Not every visualization is speedy

- Visualization is both **art** and **science**.
- Bogging down readers with unnecessarily detailed visuals can inhibit clarity.

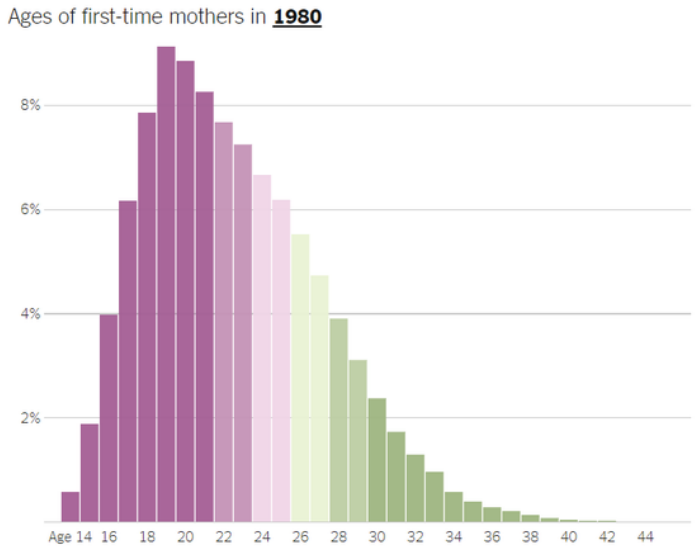


Representing distribution

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

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

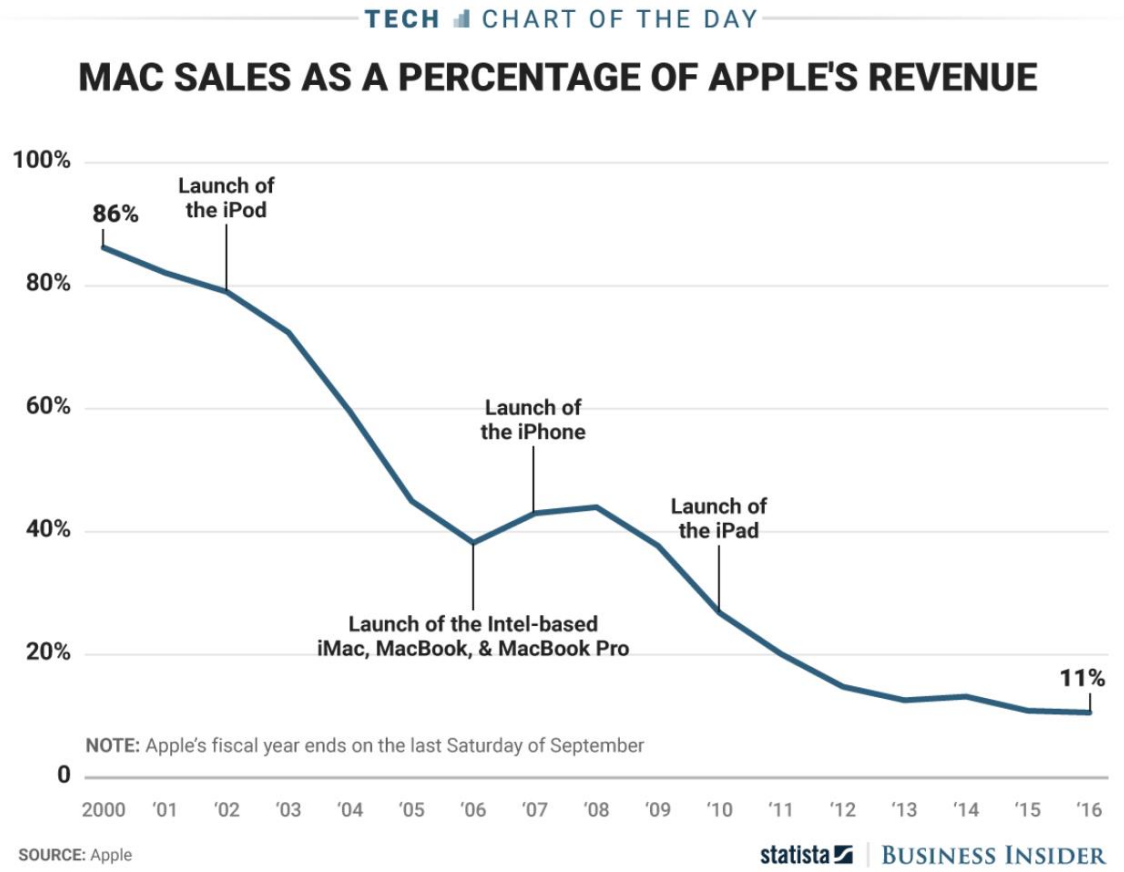
- A boxplot displays median, higher/lower quartiles and maximum/minimum.
- A histogram groups numeric data into bins, displaying the bins as segmented columns.



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

Representing time series

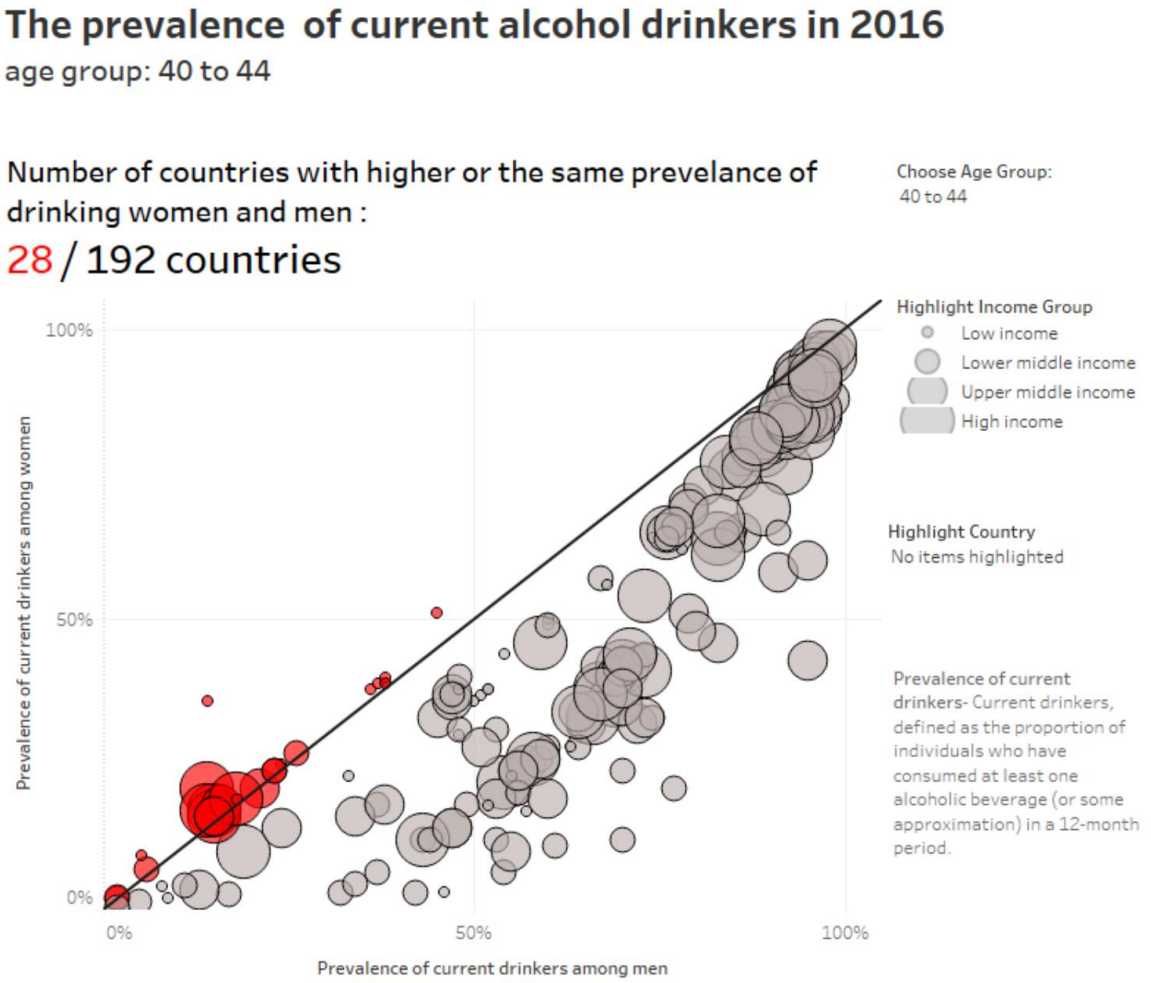
- A line chart displays information as a series of data points called markers.
- The markers are connected by straight line segments to show trend.
- An area chart is a line chart with the area below the lined filled with colors or textures.



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

Representing association

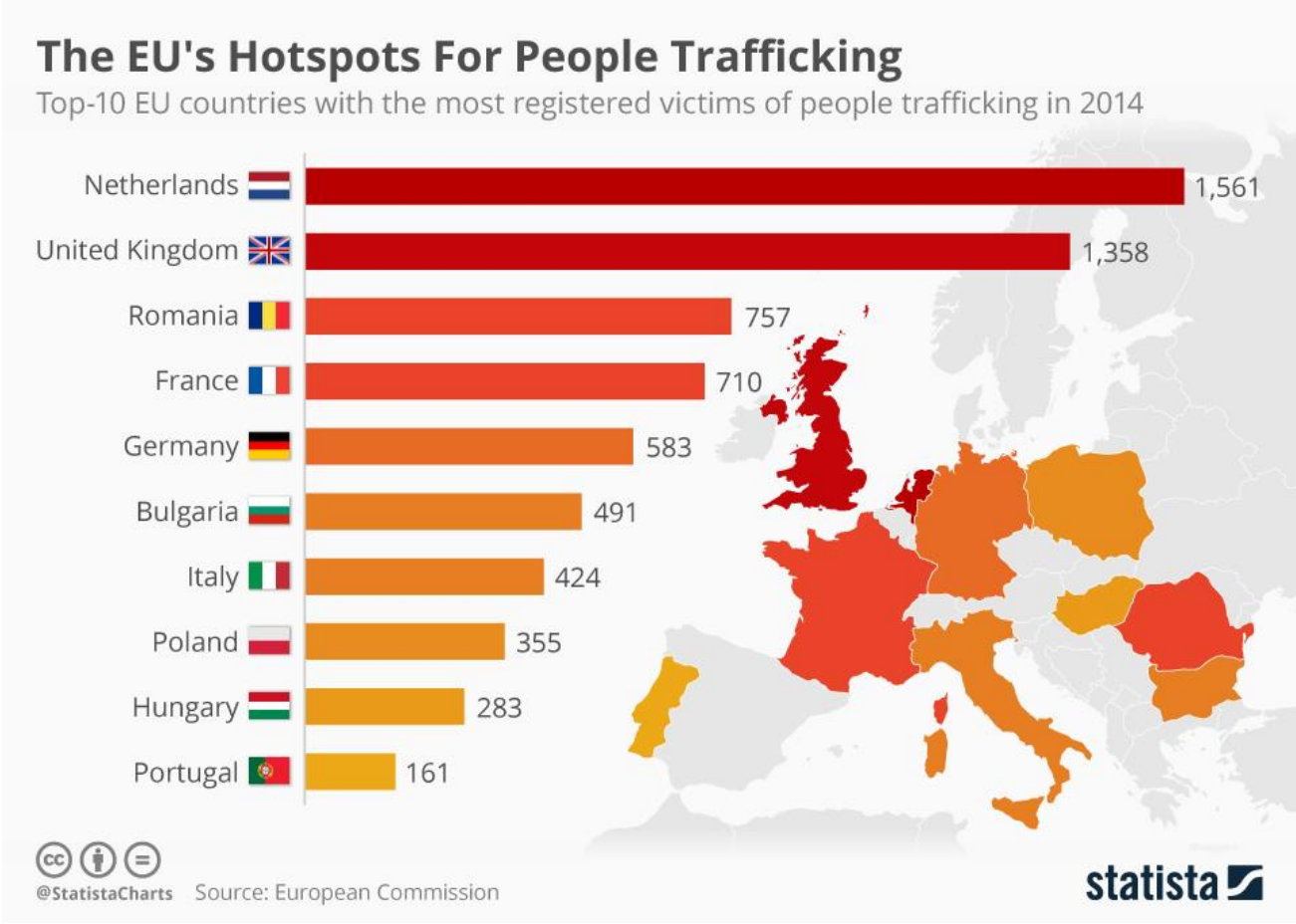
- A scatterplot is a type of diagram that uses coordinates to display values for two variables for a set of data.
- Additional data can be encoded by varying the color or shape of each marker.
- Bubble charts encode another variable through the size of the markers.



https://public.tableau.com/app/profile/anna.h.dzikowska/viz/SWDChallengescatterplot2018/Scatterplot

Comparing categories

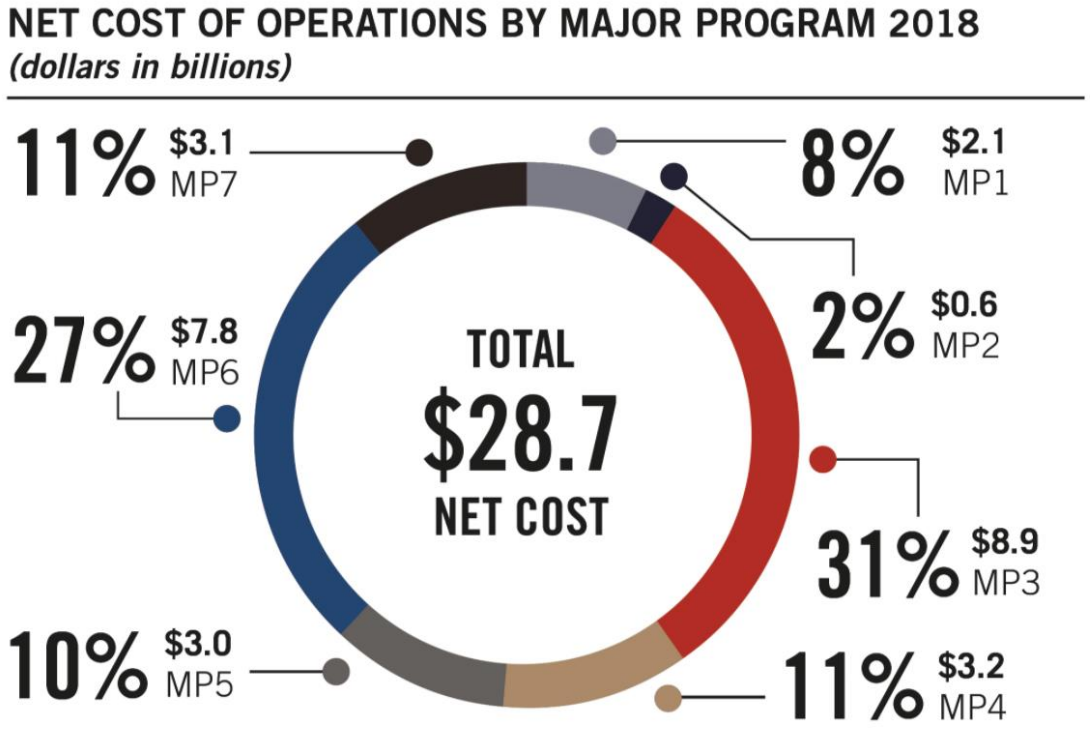
- A bar chart is a chart with rectangular bars with lengths proportional to their values.
- One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value.
- Vertical bar charts are also called column charts.



https://www.statista.com/chart/4947/the-eus-hotspots-for-people-trafficking/

Comparing part to whole

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



Cases per 100k people

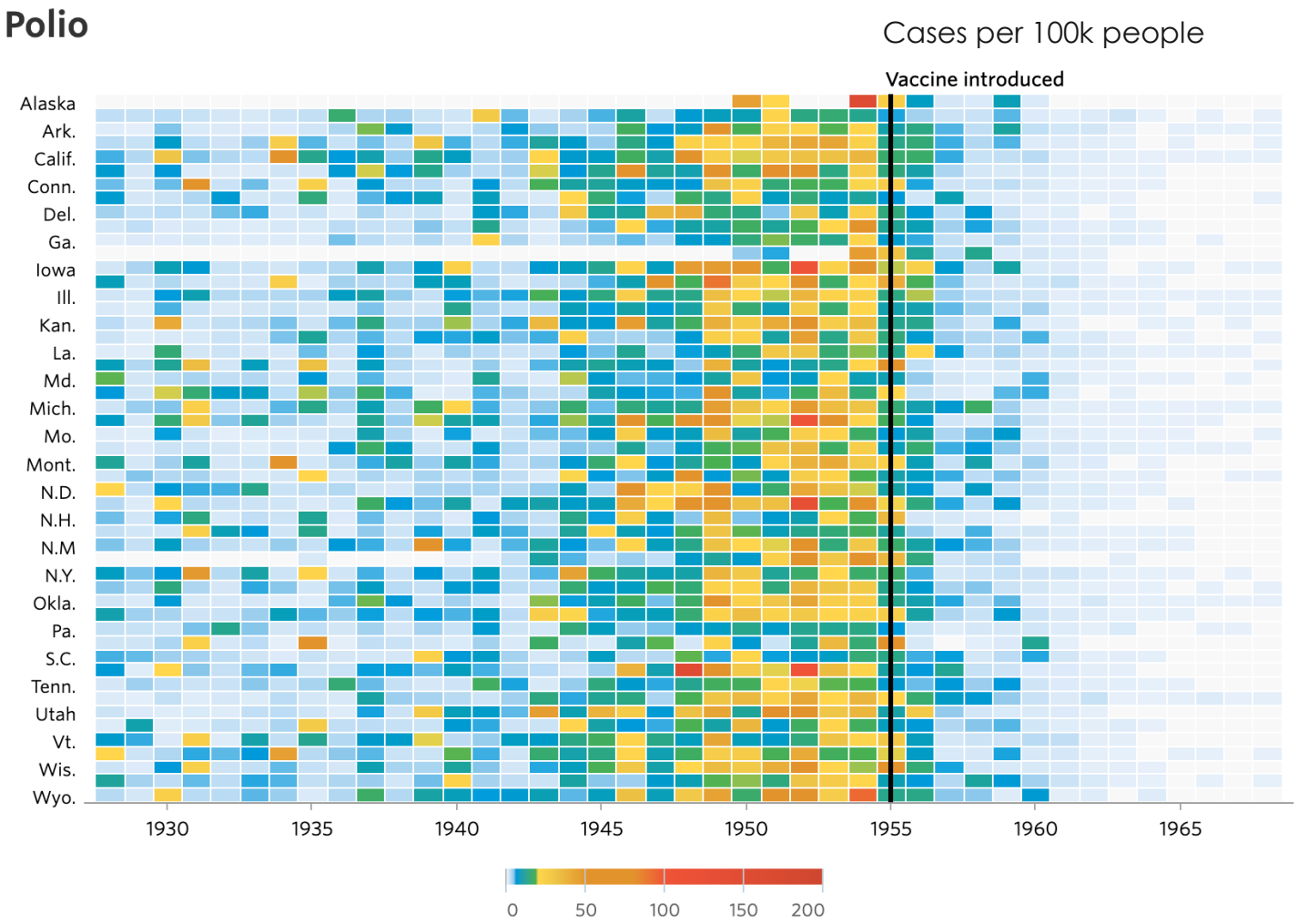
- A doughnut chart is a pie chart with a blank center to display data.

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

Representing density

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

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

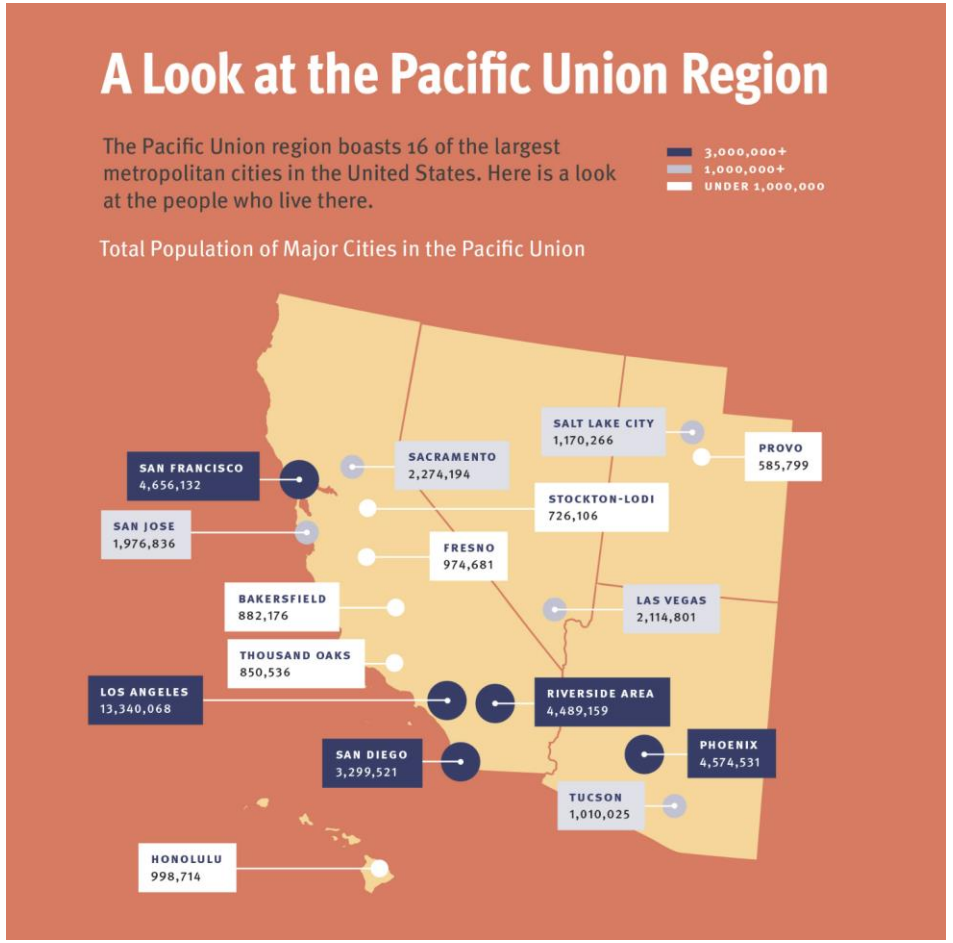


http://graphics.wsj.com/infectious-diseases-and-vaccines/

Representing territory

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

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



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

Just a few numbers

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

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

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

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

Unique data

- Don't overcomplicate!

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

Valid Passports in Circulation (1989-2020)

Enter Text To Filter Table Below

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

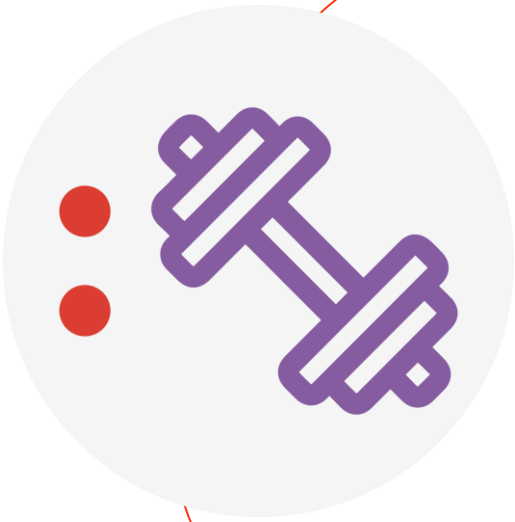
https://travel.state.gov/content/travel/en/about-us/reports-and-statistics.html

Activity: choosing data visualizations

- Turn to page 17 of your participant guide to find the **what would you viz?** activity.

- Read the description of each dataset. Then choose which of the available visualizations you would use to best represent the data.

- Check how you did using the answer key on page 19



Agenda

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

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

To get started with data viz

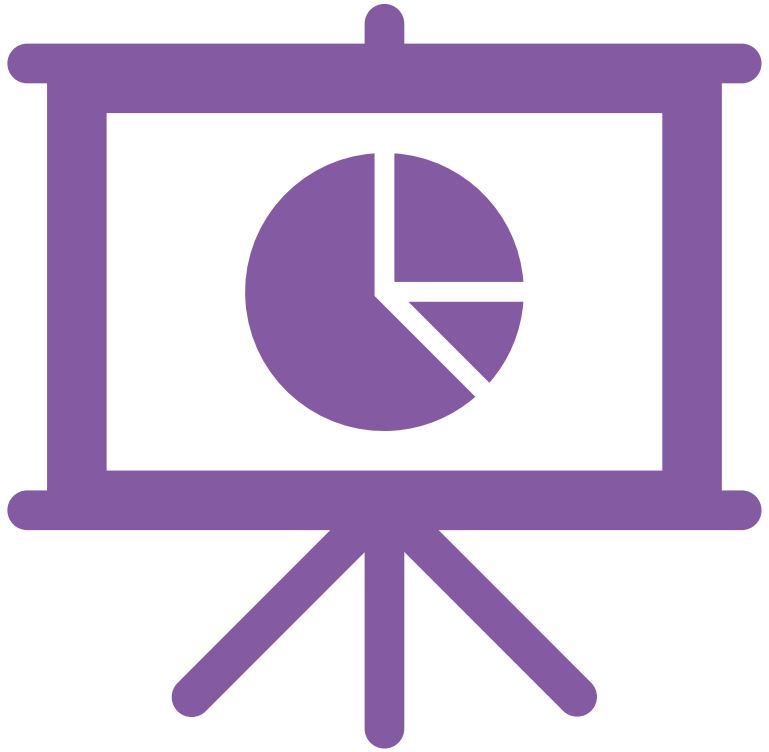
1. Know your audience and understand how it processes visual information. (Who)

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)

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.



Designing compelling visuals

- Our eyes "load" information while the brain "processes" it.

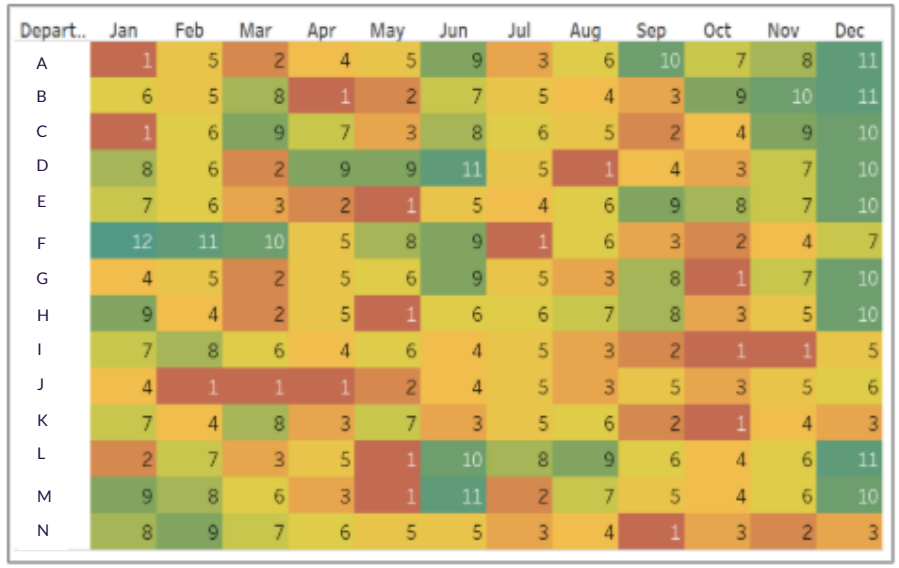
- We give the most attention to what looks good and struggle when our working memory is overwhelmed
  - Use visual clues to make data visualizations easier for the audience.

- For information to be effective, it should not provide more data than what the human brain can process.
  - Reduce visual clutter to lower the cognitive load and help transmission of the message.

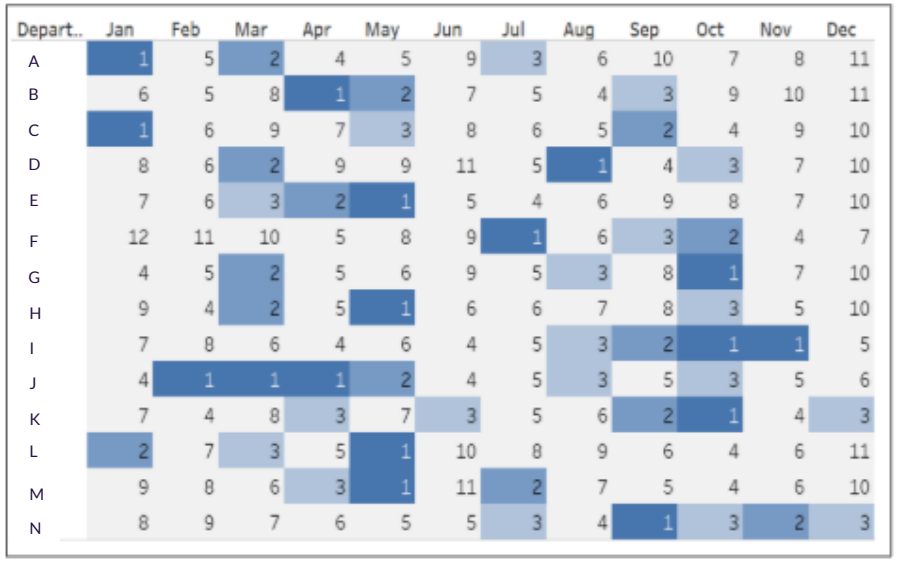
Use visual clues

- Visual clues can include elements such as color, positioning, and labeling.
- Look to visual design principles such as the Gestalt Principles or take a data viz course!

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

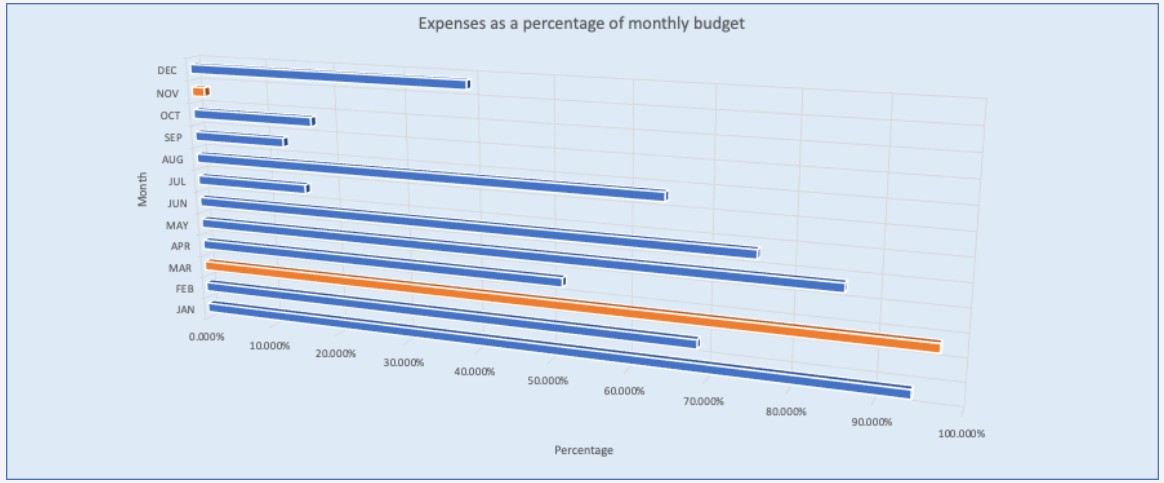


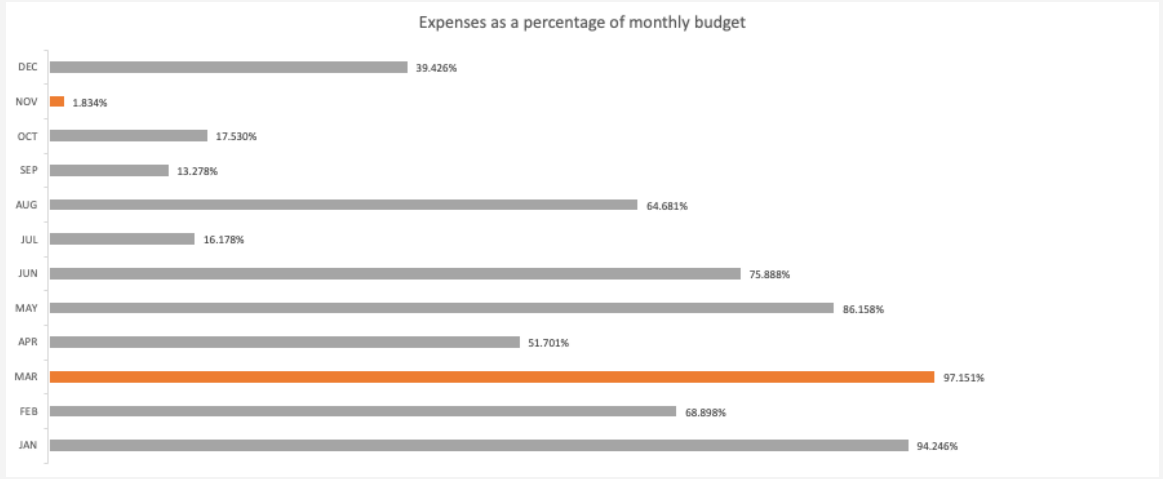
Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	1	5	2	4	5	9	3	6	10	7	8	11
B	6	5	8	1	2	7	5	4	3	9	10	11
C	1	6	9	7	3	8	6	5	2	4	9	10
D	8	6	2	9	9	11	5	1	4	3	7	10
E	7	6	3	2	1	5	4	6	9	8	7	10
F	12	11	10	5	8	9	1	6	3	2	4	7
G	4	5	2	5	6	9	5	3	8	1	7	10
H	9	4	2	5	1	6	6	7	8	3	5	10
I	7	8	6	4	6	4	5	3	2	1	1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

Reduce chart clutter

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

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

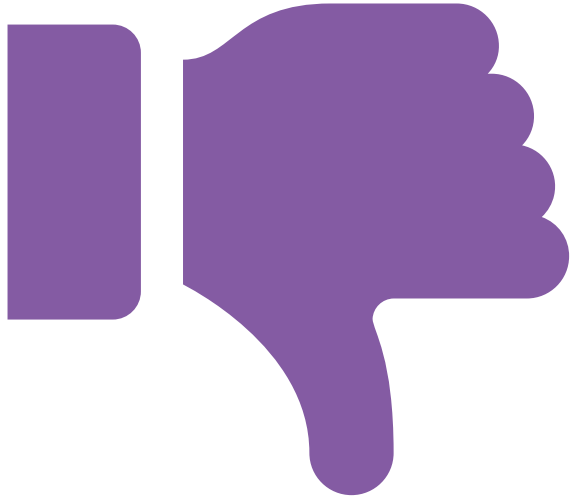




Misleading stats & visual distortions

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

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



Misleading statistics

- "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.

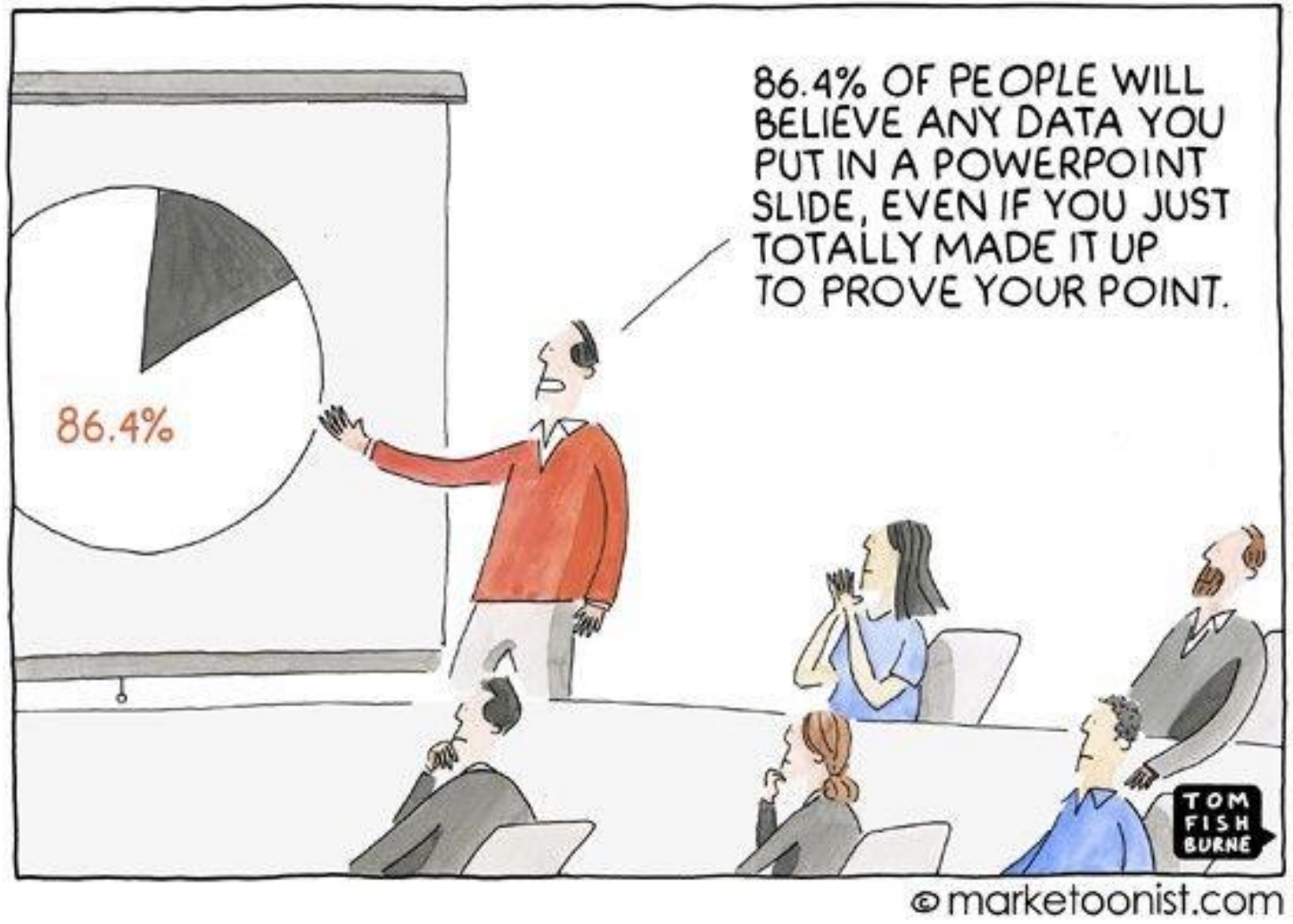
What's wrong with these statements?

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

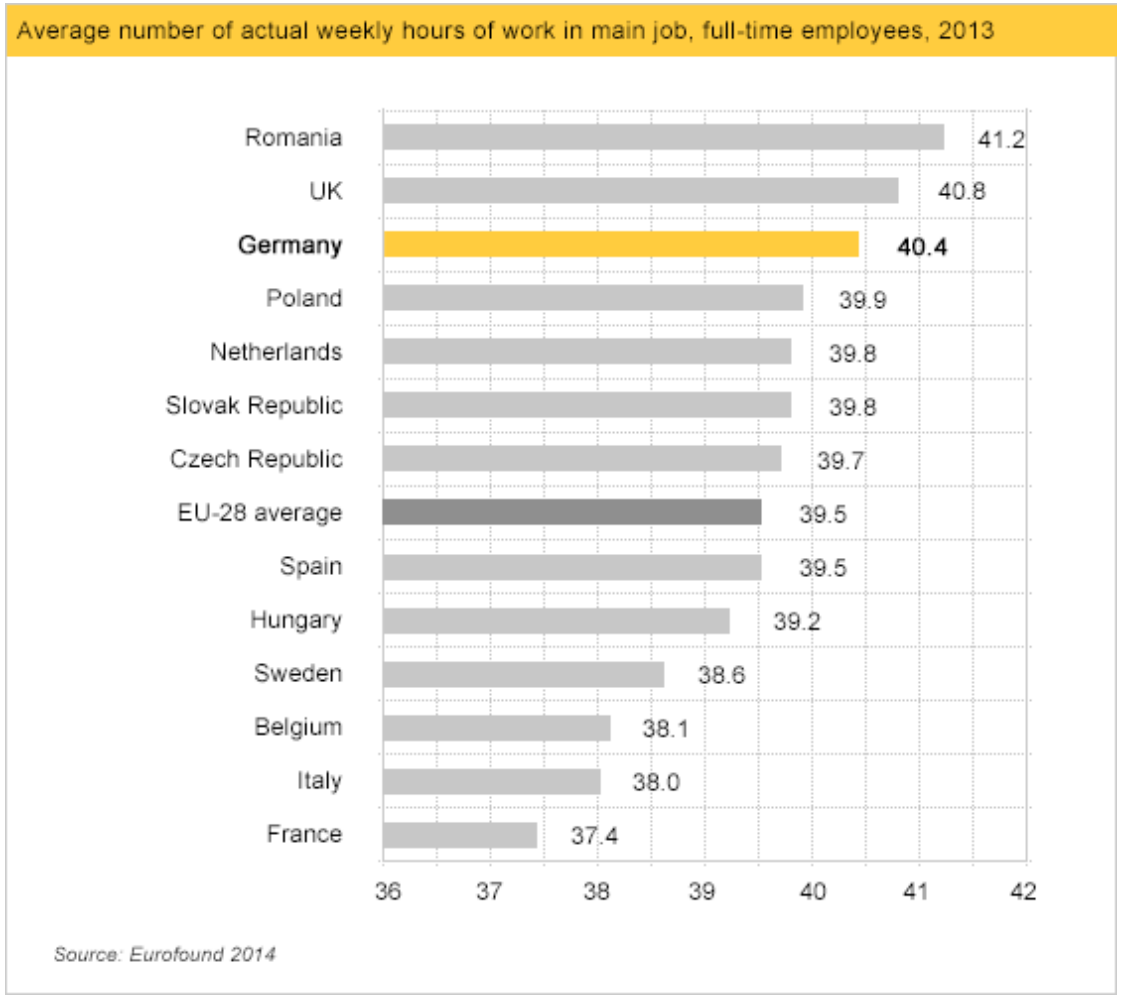
How to avoid being misled?

- 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

At first glance, what is the ratio of weekly hours of work in Germany versus in France?

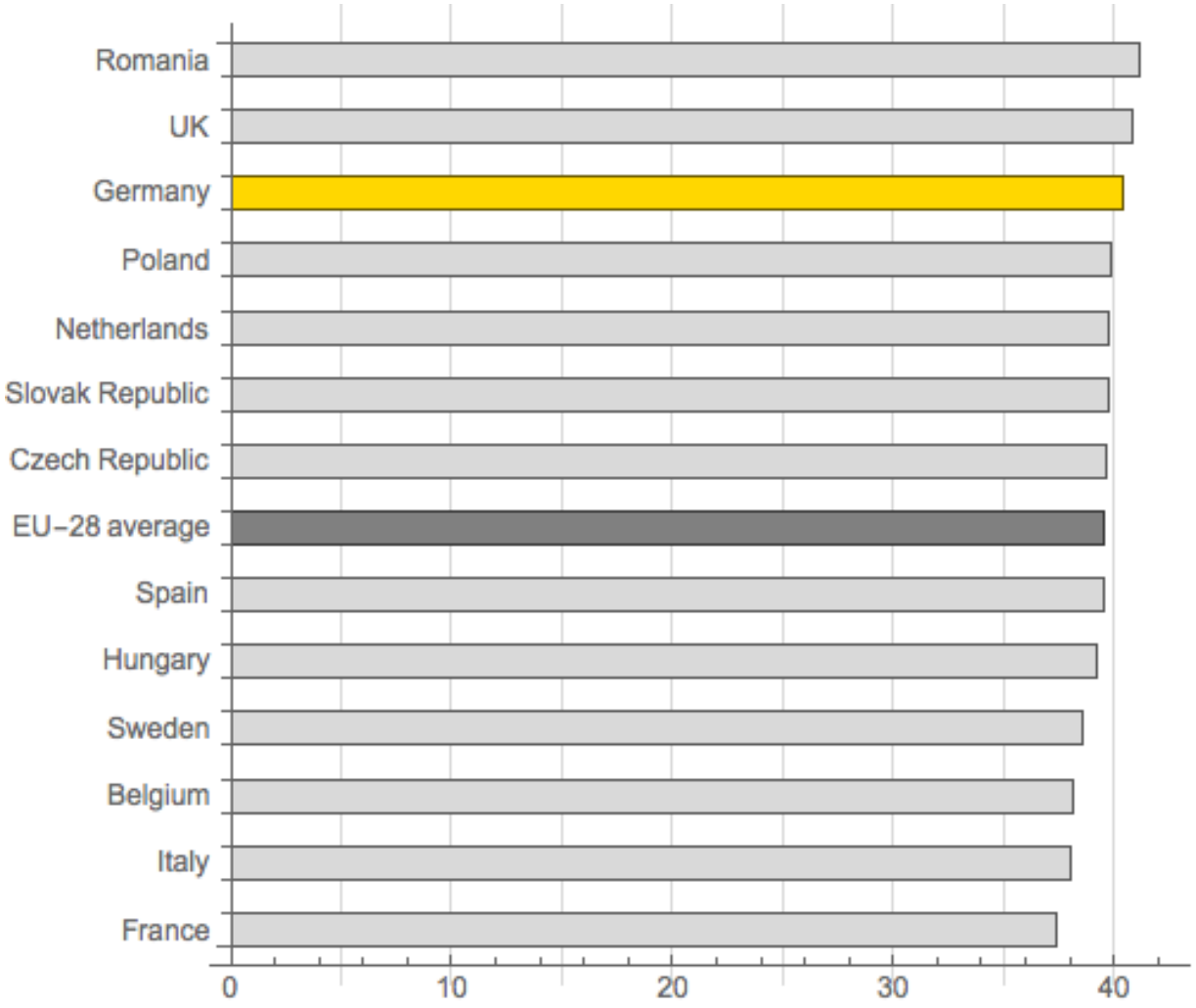
What is the actual difference between weekly hours of work in Germany versus in France?



Truncated graphs

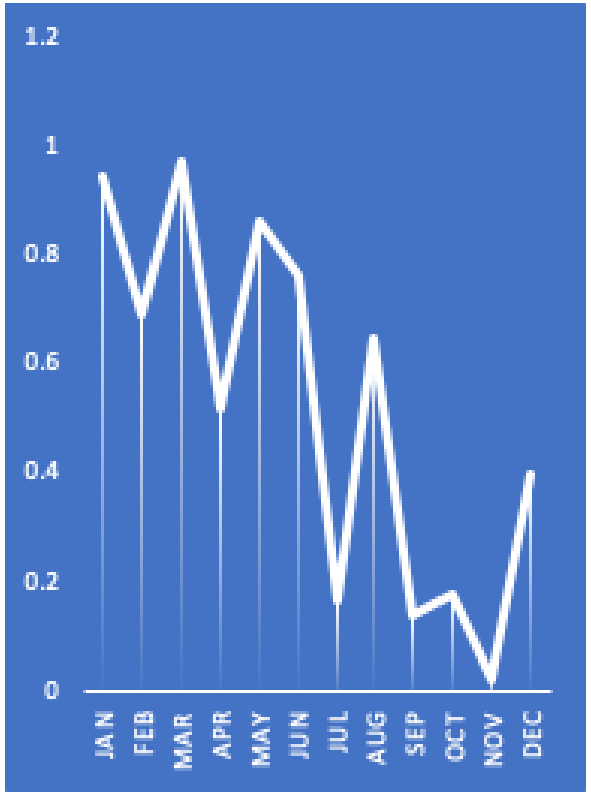
- One of the most common manipulations is omitting baselines or beginning the y-axis of a graph at an arbitrary number instead of 0.

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



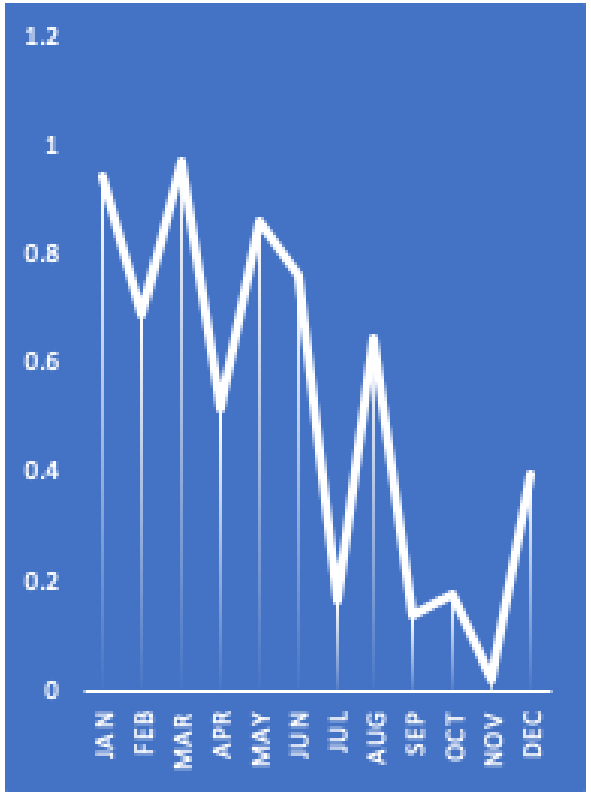
Visual distortions

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



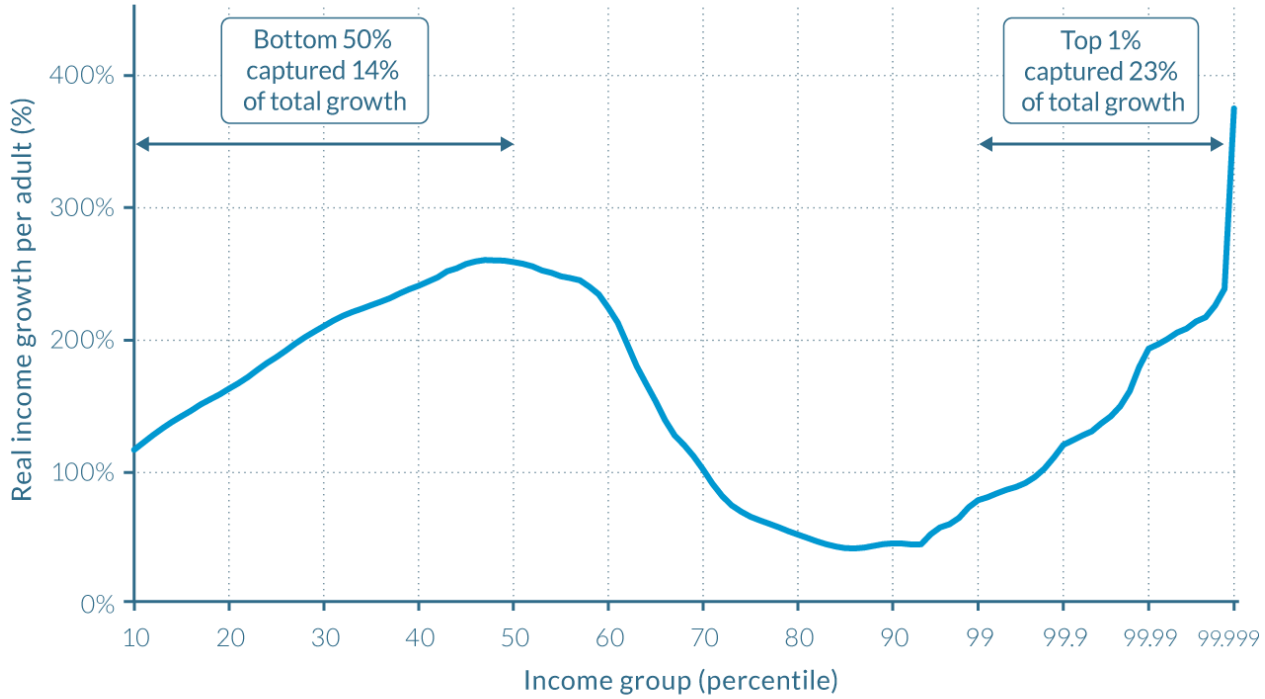
Exaggerated scaling

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



Visual distortion

How might this chart be misleading?



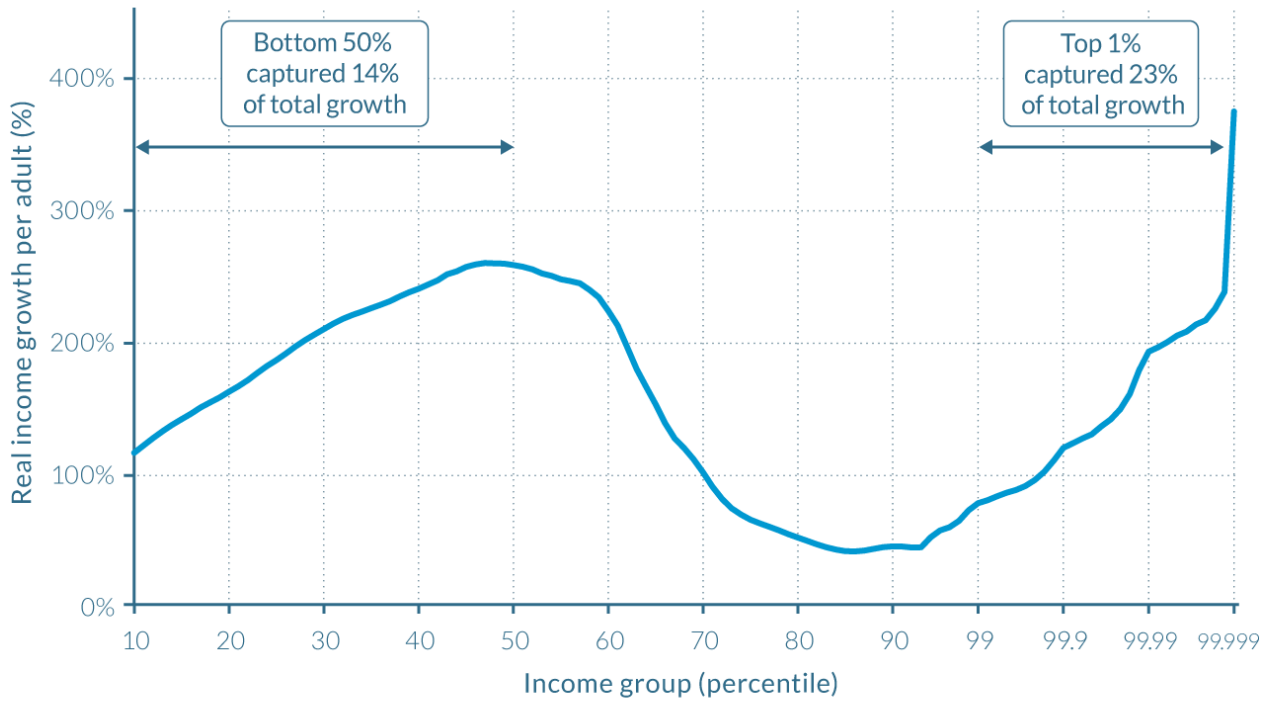
Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

Ignoring convention

- Deviating from convention (like changing the scale of an axis) can create confusion and misinterpretation of the facts.

- If this example followed a consistent scale, we would expect to see more modest gains across 90% of the graph followed by a sharp spike.

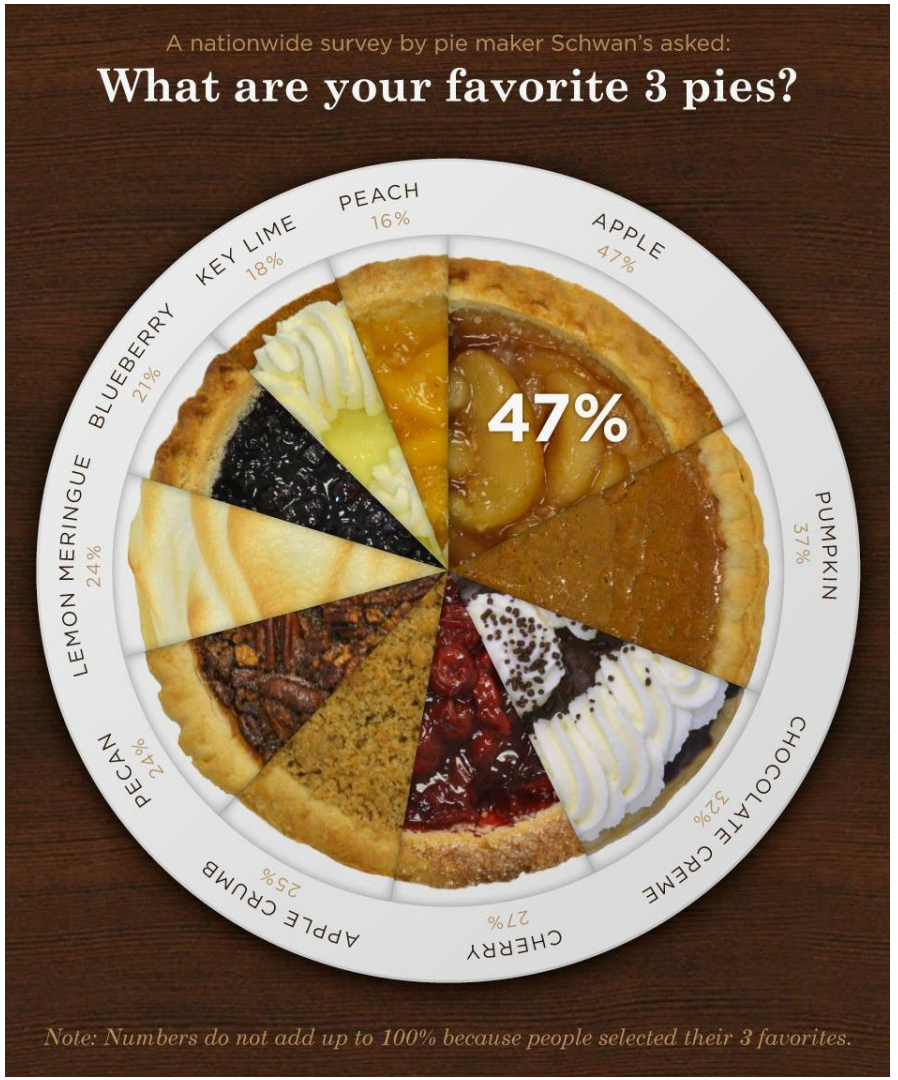


Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

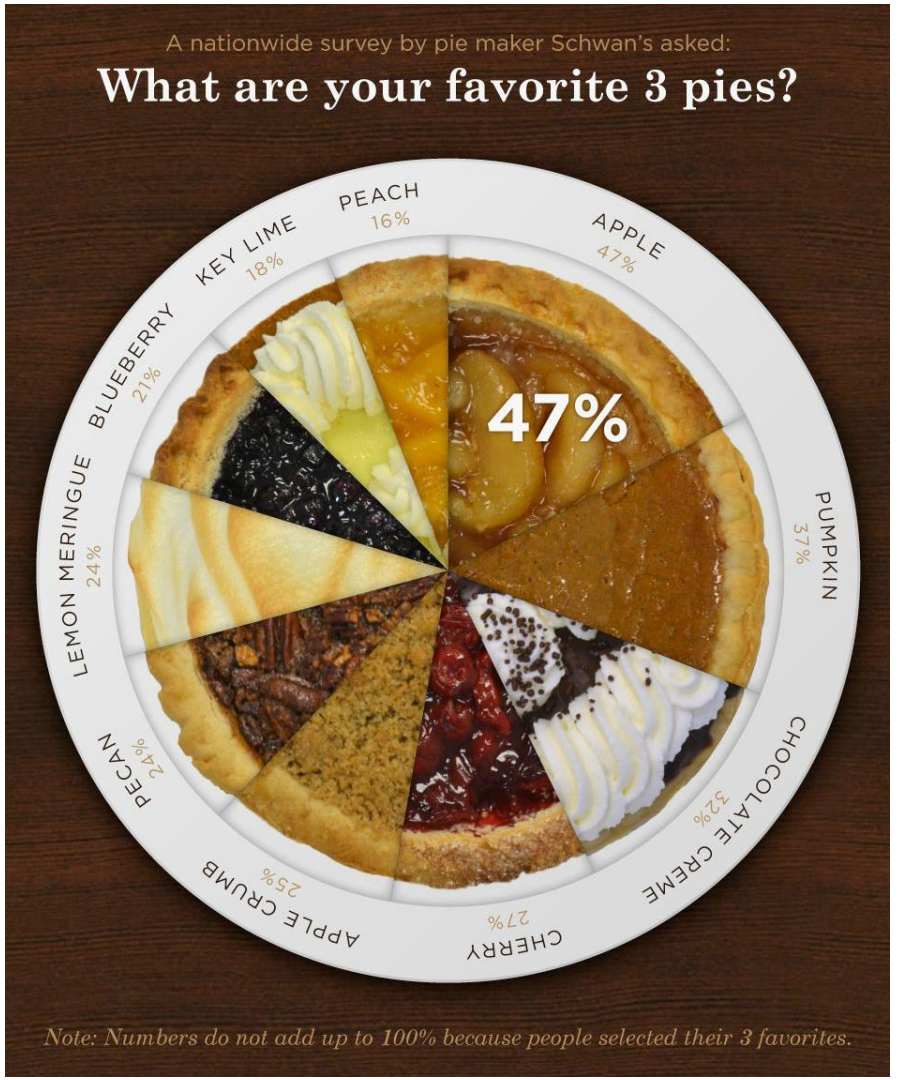
Visual distortion

What do you notice about this pie chart?



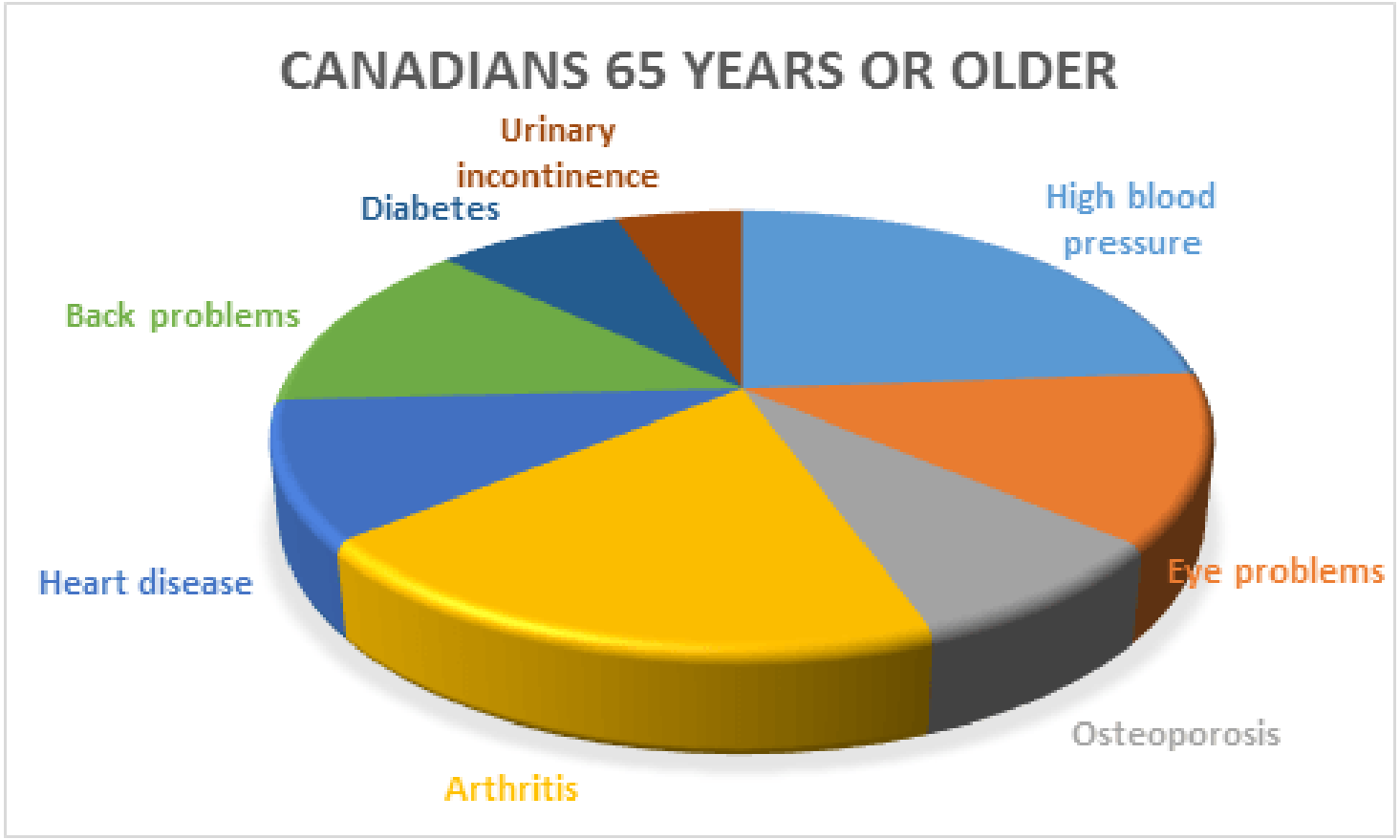
Numbers don’t add up

- 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.



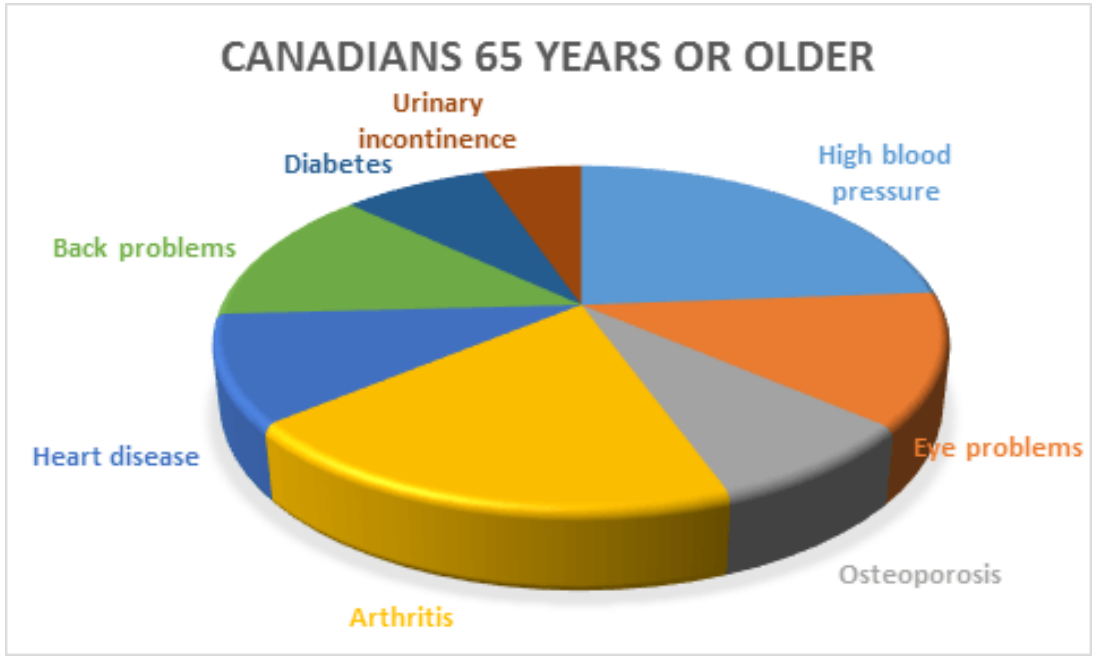
Visual distortion

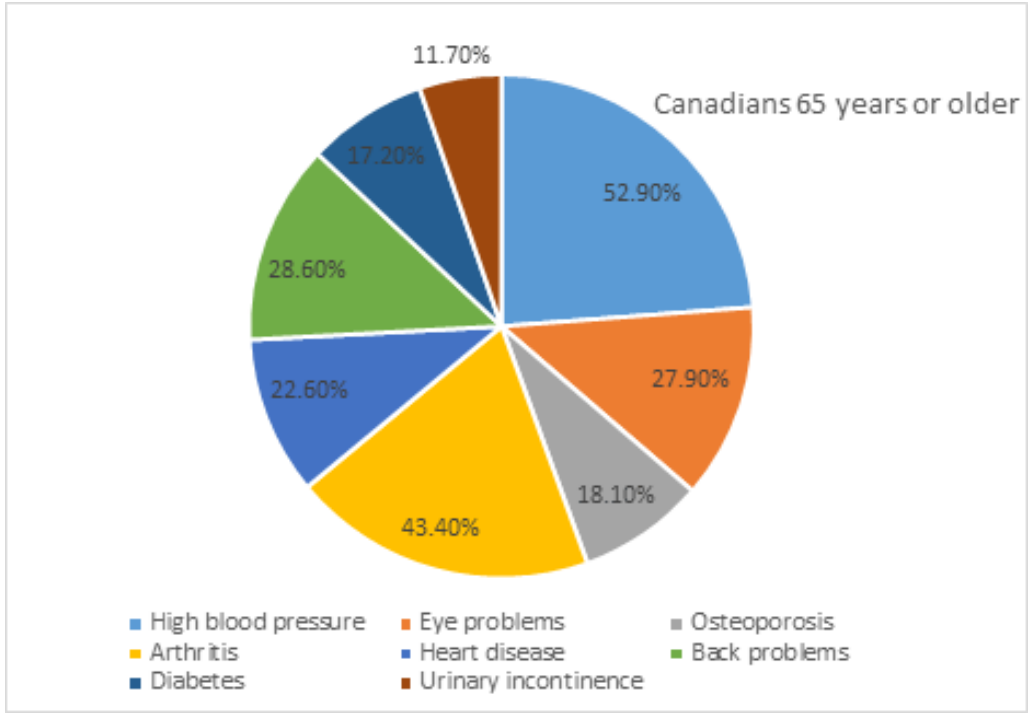
Do more Canadians over 65 suffer from arthritis or from high blood pressure?



3D distortion

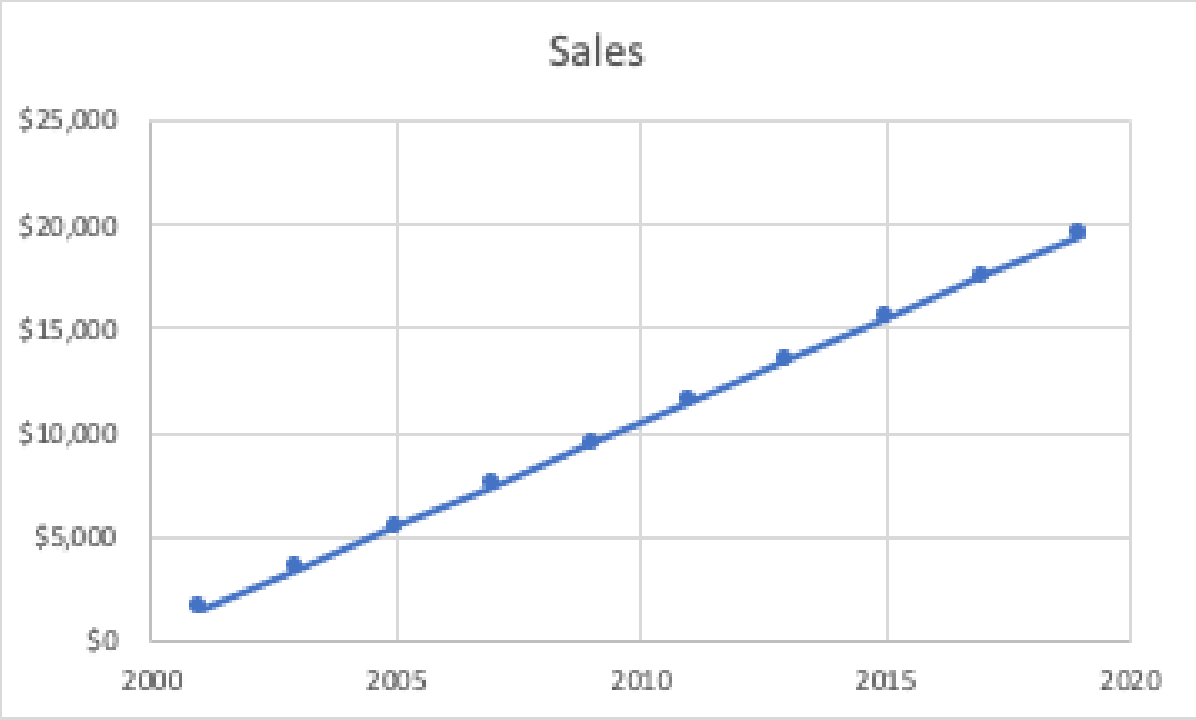
- 3D pie charts can be used to distort and cause a misinterpretation of the data.
- The same data is represented in both charts below.
- Note that in neither case do the numbers add up to 100%.

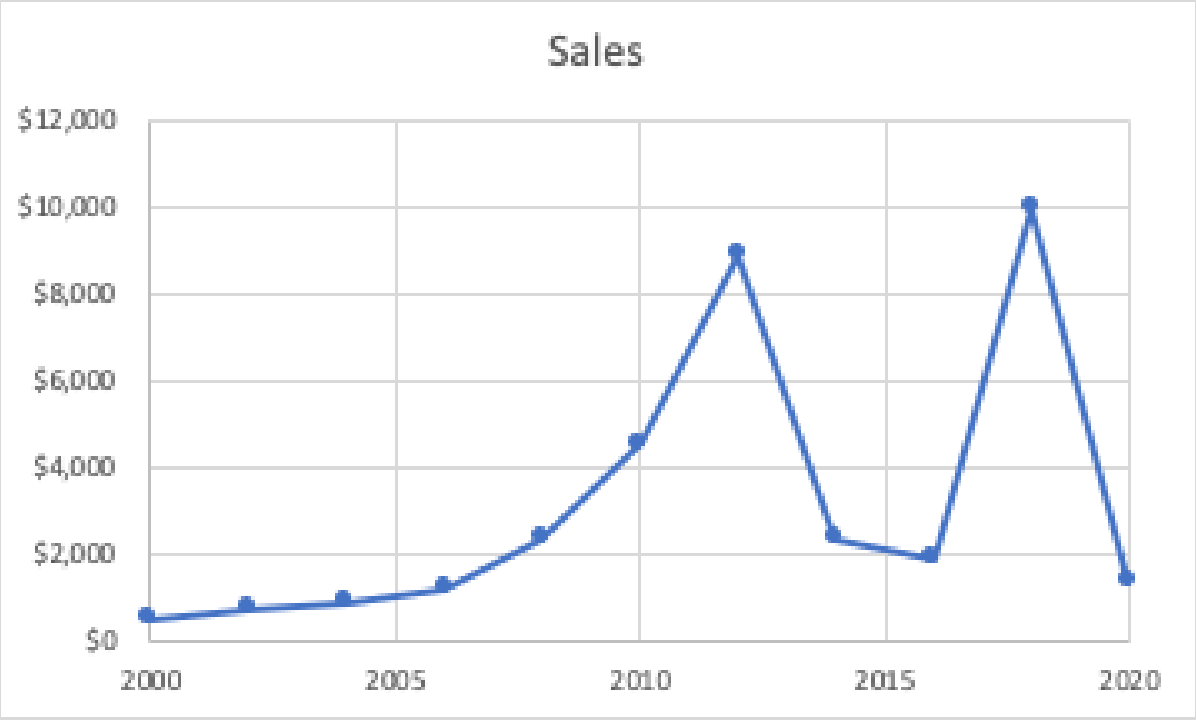




Visual distortion

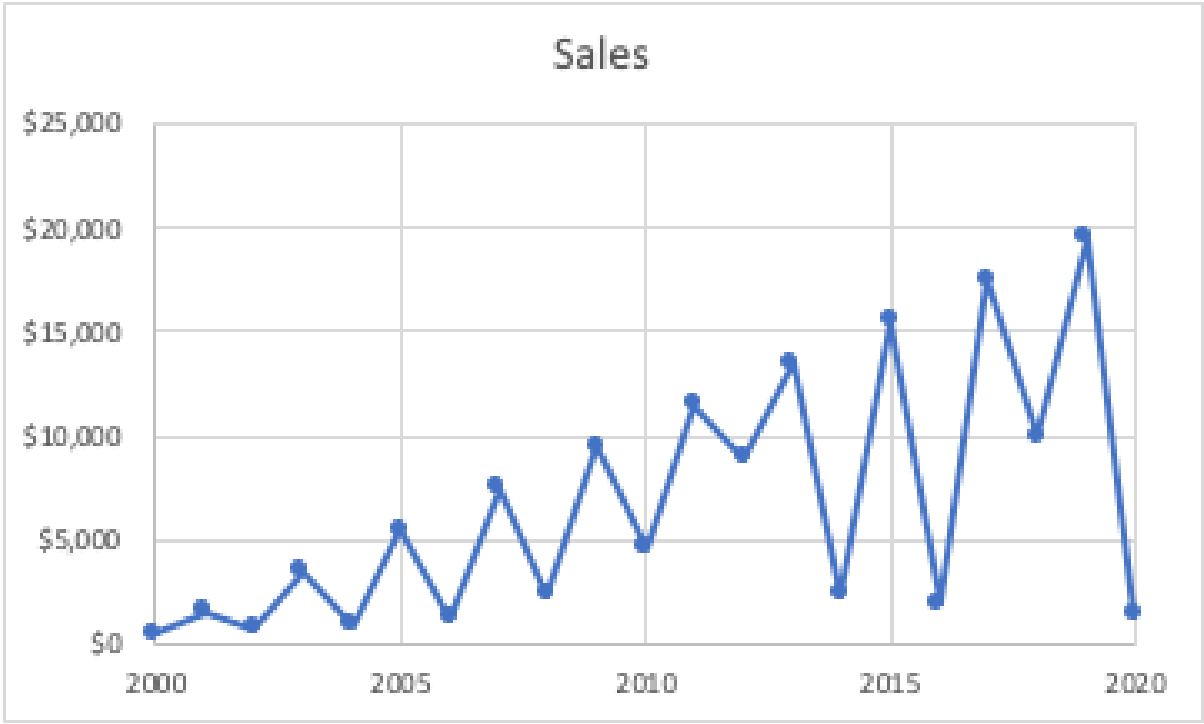
Which company has a better sales trajectory?



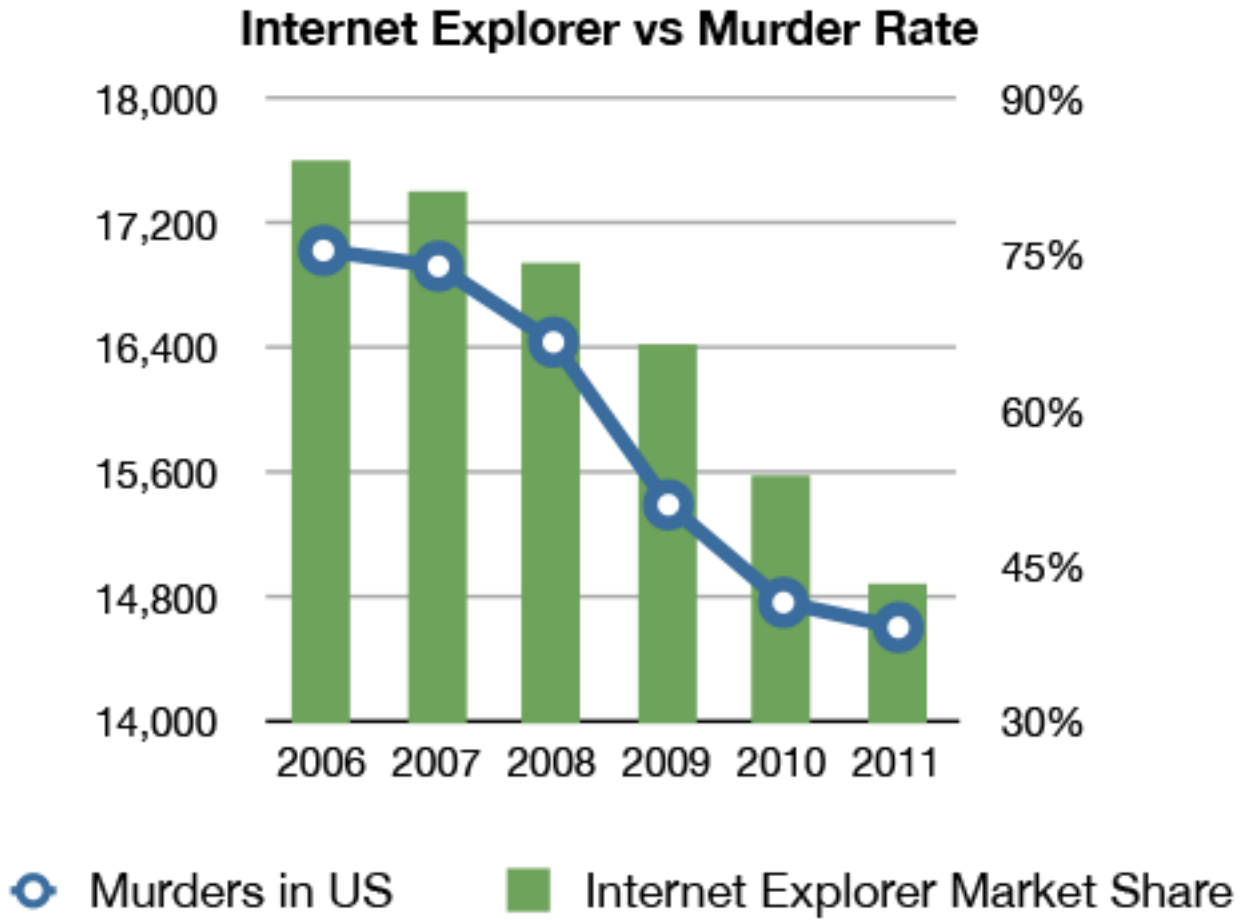


Improper extraction

- 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.



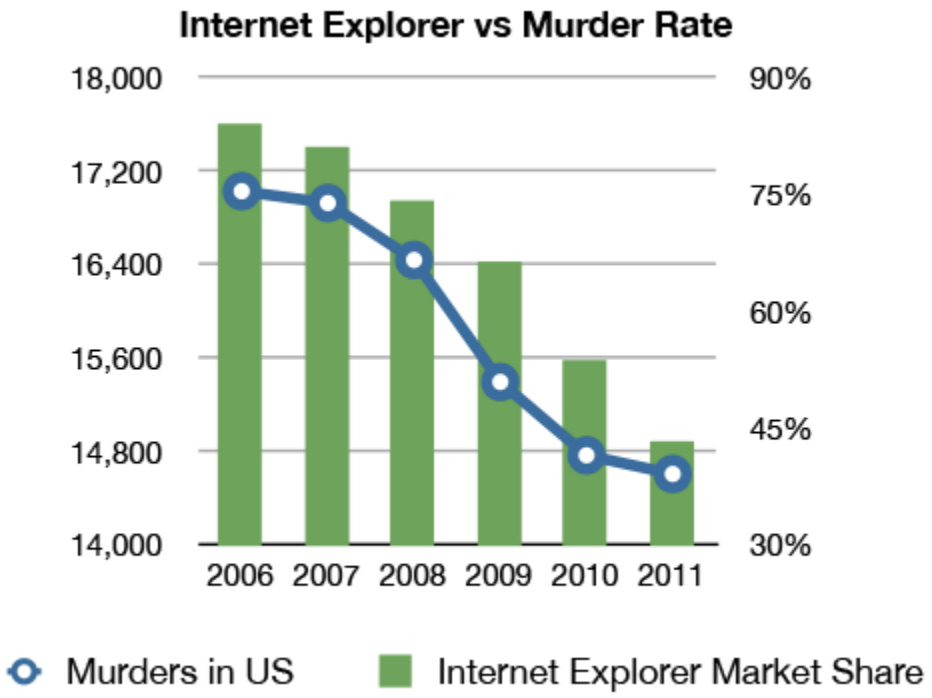
Visual distortion



What story does this visualization tell?

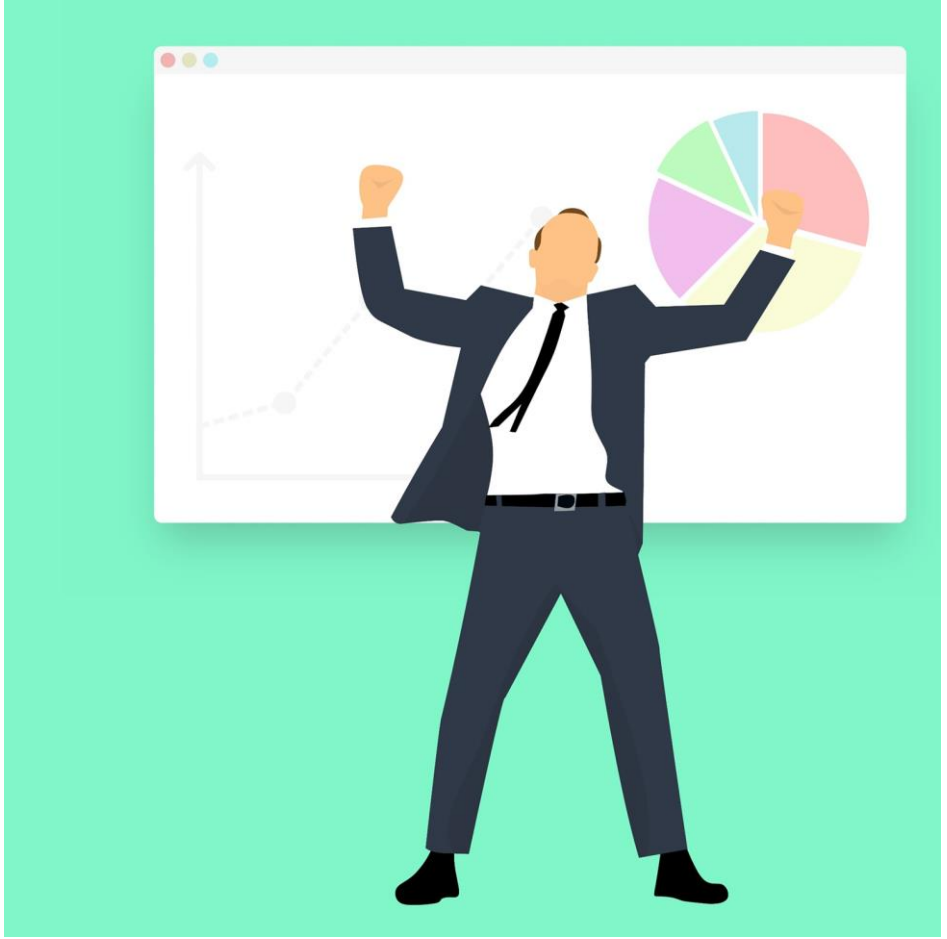
Correlating causation

- Data visualizations can imply causal links through how data is presented to the viewer.
- However, *correlation does not equal causation*.



Recap

- To avoid being misled, look for:
  - misleading statistics
  - truncated graphs
  - exaggerated scaling
  - ignored conventions
  - numbers that don't add up
  - 3D distortion
  - improper extraction
  - correlating causation



Thank you!

hello@datasociety.com

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

(202)600-9635

datasociety.com