DATA SOCIETY®

"If you can't measure it, you can't manage it"

- Peter Drucker, management consultant, educator, and author

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

Welcome

- Instructor introduction

- Schedule:
  - 4 sessions
  - 11 am – 2 pm
  - 1 or 2 short breaks each session
  - Q & A at the end of each session

Best practices for virtual classes

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

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

3. Participate in activities and ask questions in Q and A tab

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

Class materials

- You should have received access to the following materials:
  - Slides
  - Participant guide
    - Needed during class
    - Contains activities, a data science glossary, information about popular data science tools, and more!

Poll Question

What would you rate your current data literacy level on a scale of 1-10?

0
No knowledge or awareness; may read articles or documents that contain percentages

5
Has combined data from multiple sources; has made basic charts/graphs; understands data limitations

10
Works with Big Data; has deployed machine learning

Agenda

Day 1

- Data and its uses
- Data analytics overview
- Data governance
- Data tools
- Data teams

Day 2

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

Day 3

- Foundational data science methods
- Advanced data science methods

Day 4

- Data visualization
- Data storytelling

Agenda

Day 1

- Data and its uses
- 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

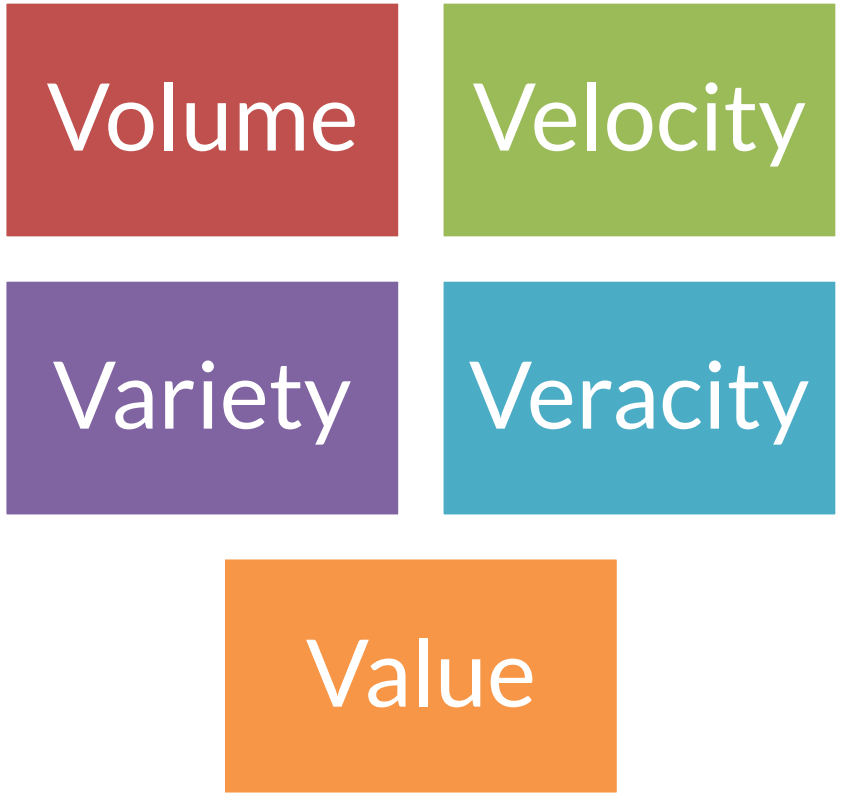
3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Merriam Webster

What is big data?

- Refers to a large volume of data that can be mined for information and used in machine learning projects and other analytics applications
- Typically, the size of big data is described in terabytes, petabytes, even exabytes!
- "Big data" **is not** analytics. You can't "do" big data, but you can use it for data analytics.

Five V's of big data



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

- Images
- Video
- Audio
- Raw Sensor Data
  Text Data

Sources of data

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

External sources of data

- Publicly-accessible APIs
  - e.g., api.data.gov
- Other open data sources
  - e.g., data.worldbank.org
- Large businesses (e.g., Wal-Mart, Best Buy, Trip Advisor, Expedia, Google, and Spotify) are increasingly giving people access to their data
- Data is sometimes available for purchase (e.g. weather data)

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

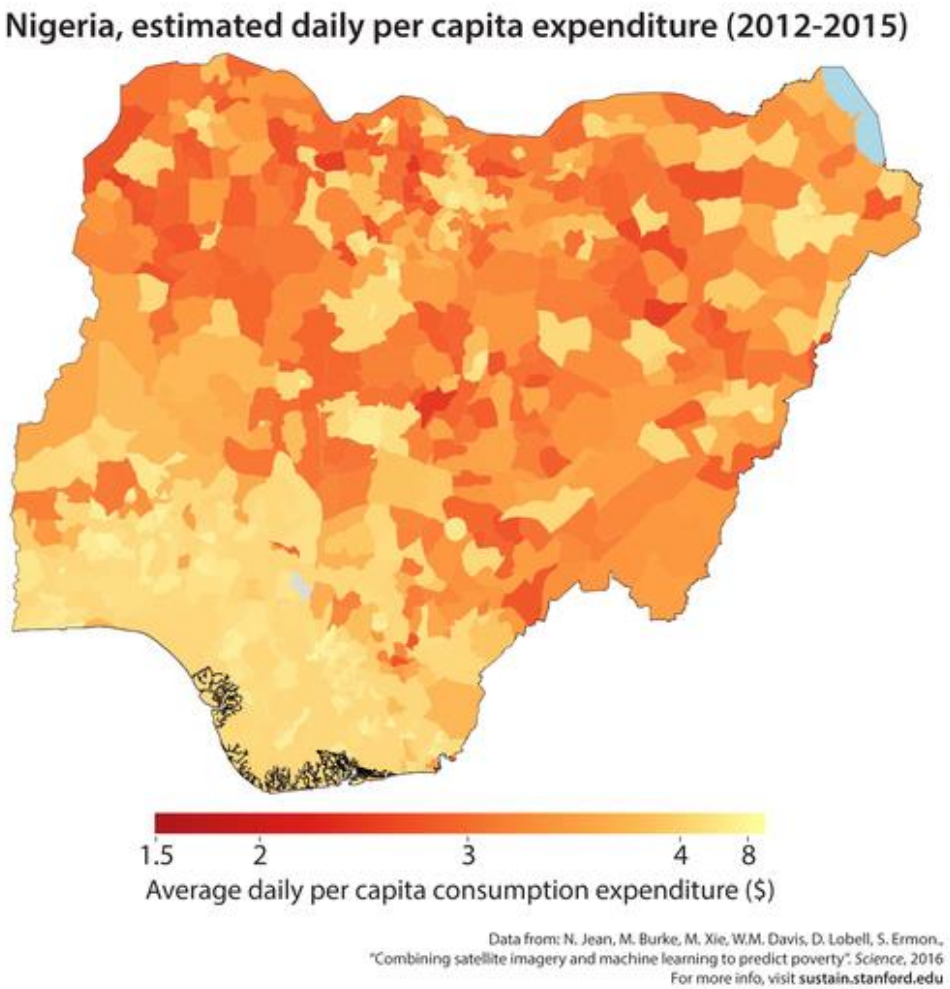
5. Optimize resources

6. Enable experiments

Find a needle in haystack

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

- 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

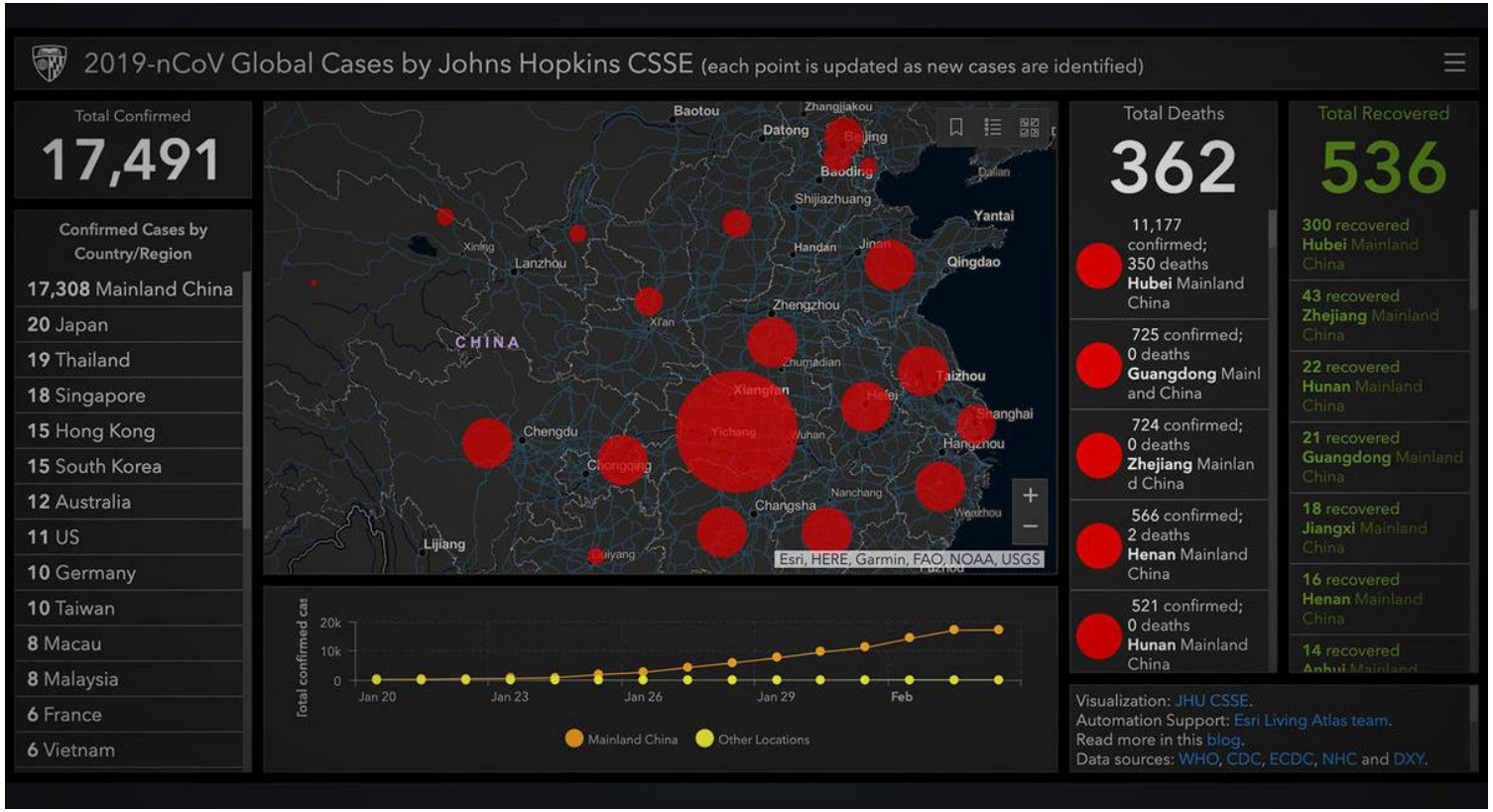
- Consultants in Philadelphia developed a model for prioritizing building inspections based on a location's:
  - Distance to nearby vacant properties
  - Distance to certain crimes
  - Distance to infestation reports

- 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

- The Center for Systems Science and Engineering at Johns Hopkins University launched an online dashboard to track the spread of coronavirus across the globe in real time.
- The live dashboard pulls data from the World Health Organization and the centers for disease control in the US, China, and Europe.



https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

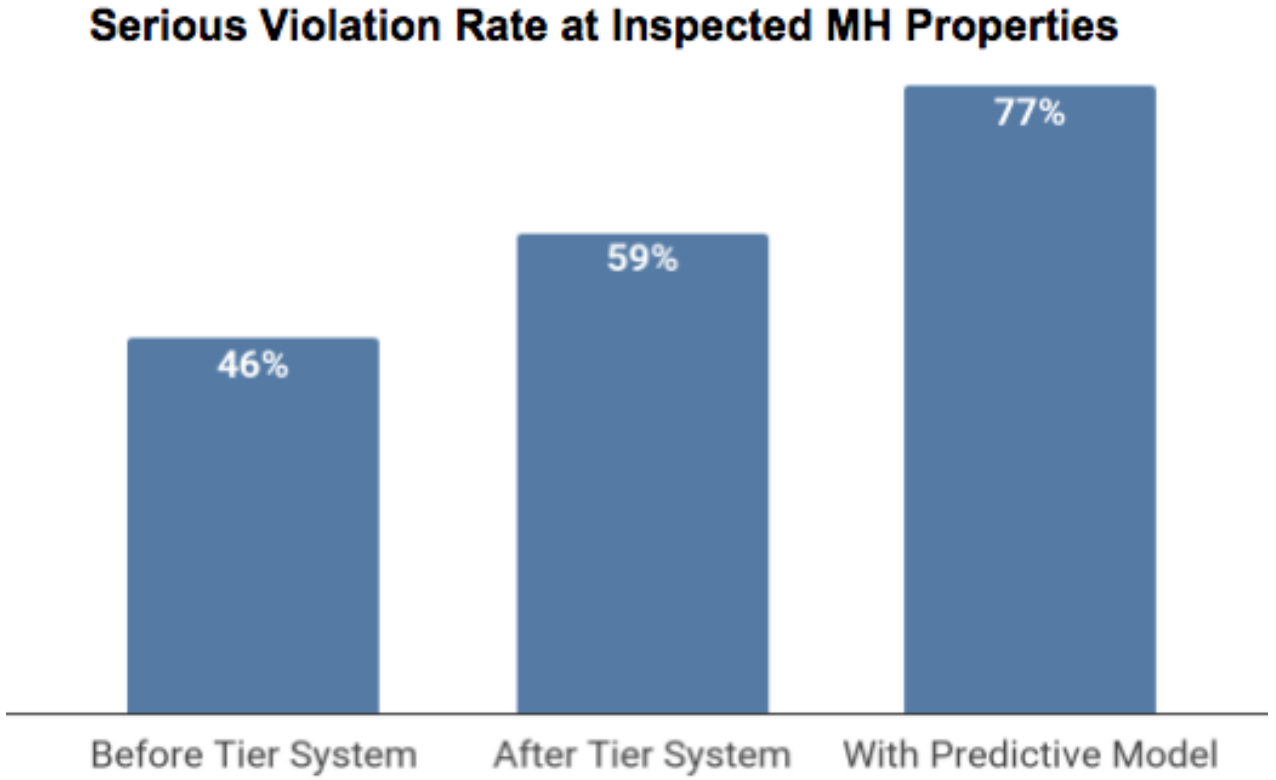
Speed up decisions

- Before 2011, the Office of Justice Program's (OJP) public safety grant review process depended heavily on the individual knowledge of grant managers for pre- and post-award decisions.
- OJP pulled disparate data systems together and automated its review processes to increase the accuracy and consistency of its decisions.
- The time needed for grant managers to capture grantee data in OJP's database went from 30 minutes to almost zero.

https://www2.deloitte.com/insights/us/en/industry/public-sector/data-driven-decision-making-in-government.html

Optimize resources

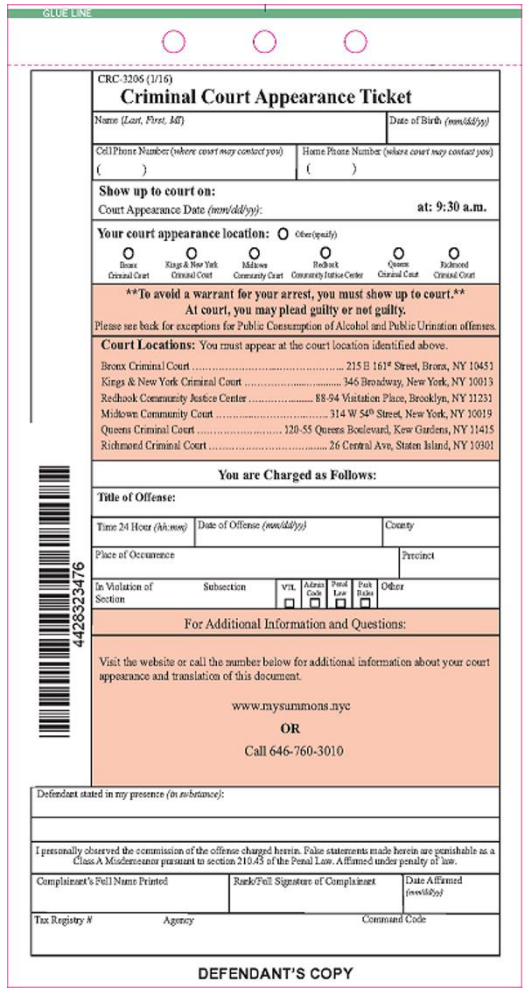
- In June 2015, a balcony collapsed in Berkeley, killing six people and injuring seven more
- San Jose's Multiple Housing Team implemented data analytics to prioritize high-risk cases
- By identifying the most important factors, they built a model to help prevent future building collapses



https://dssg.uchicago.edu/2017/07/14/data-driven-inspections-for-safer-housing-in-san-jose-california/

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



CRC-3206 (1/16)

Criminal Court Appearance Ticket

Name (*Last, First, MI*) | Date of Birth (*mm/dd/yy*)

Cell Phone Number (*where court may contact you*) | Home Phone Number (*where court may contact you*)

Show up to court on:

Court Appearance Date (*mm/dd/yy*): **at: 9:30 a.m.**

Your court appearance location: Other (specify)

Bronx Criminal Court | Kings & New York Criminal Court | Midtown Community Court | Redhook Community Justice Center | Queens Criminal Court | Richmond Criminal Court

****To avoid a warrant for your arrest, you must show up to court.****
At court, you may plead guilty or not guilty.
Please see back for exceptions for Public Consumption of Alcohol and Public Urination offenses.

Court Locations: You must appear at the court location identified above.

Bronx Criminal Court ..... 215 E 161st Street, Bronx, NY 10451
Kings & New York Criminal Court ..... 346 Broadway, New York, NY 10013
Redhook Community Justice Center ..... 88-94 Visitation Place, Brooklyn, NY 11231
Midtown Community Court ..... 314 W 54th Street, New York, NY 10019
Queens Criminal Court ..... 120-55 Queens Boulevard, Kew Gardens, NY 11415
Richmond Criminal Court ..... 26 Central Ave, Staten Island, NY 10301

4428323476

You are Charged as Follows:

Title of Offense:

Time 24 Hour (*hh:mm*) | Date of Offense (*mm/dd/yy*) | County

Place of Occurrence | Precinct

In Violation of Section | Subsection | VTL | Admin Code | Penal Law | Park Rules | Other

For Additional Information and Questions:

Visit the website or call the number below for additional information about your court appearance and translation of this document.

www.mysummons.nyc
OR
Call 646-760-3010

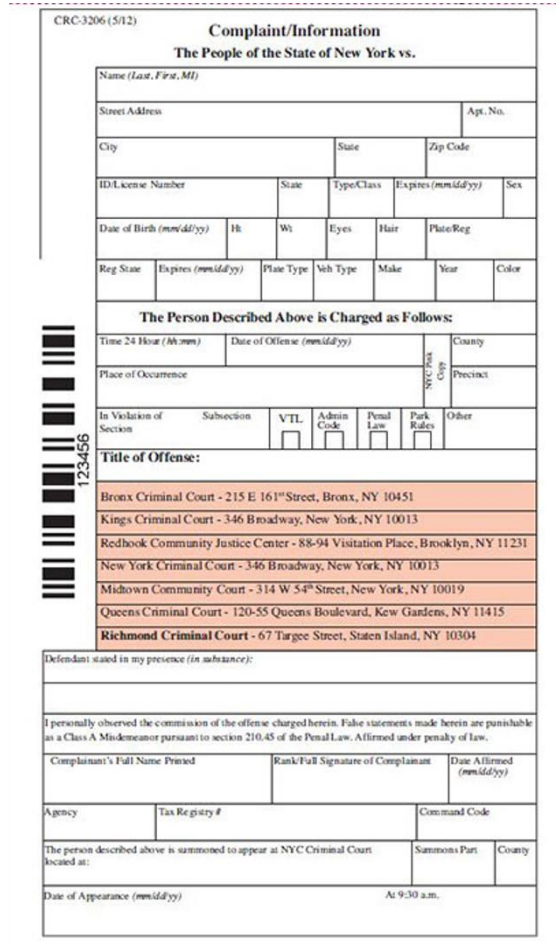
Defendant stated in my presence (*in substance*):

I personally observed the commission of the offense charged herein. False statements made herein are punishable as a Class A Misdemeanor pursuant to section 210.45 of the Penal Law. Affirmed under penalty of law.

Complainant's Full Name Printed | Rank/Full Signature of Complainant | Date Affirmed (*mm/dd/yy*)

Tax Registry # | Agency | Command Code

DEFENDANT'S COPY



CRC-3206 (5/12)

Complaint/Information

The People of the State of New York vs.

Name (*Last, First, MI*)

Street Address | Apt. No.

City | State | Zip Code

ID/License Number | State | Type/Class | Expires (*mm/dd/yy*) | Sex

Date of Birth (*mm/dd/yy*) | Ht | Wt | Eyes | Hair | Plate/Reg

Reg State | Expires (*mm/dd/yy*) | Plate Type | Veh Type | Make | Year | Color

The Person Described Above is Charged as Follows:

Time 24 Hour (*hh:mm*) | Date of Offense (*mm/dd/yy*) | NYC Clean Copy | County

Place of Occurrence | Precinct

In Violation of Section | Subsection | VTL | Admin Code | Penal Law | Park Rules | Other

123456

Title of Offense:

Bronx Criminal Court - 215 E 161st Street, Bronx, NY 10451
Kings Criminal Court - 346 Broadway, New York, NY 10013
Redhook Community Justice Center - 88-94 Visitation Place, Brooklyn, NY 11231
New York Criminal Court - 346 Broadway, New York, NY 10013
Midtown Community Court - 314 W 54th Street, New York, NY 10019
Queens Criminal Court - 120-55 Queens Boulevard, Kew Gardens, NY 11415
Richmond Criminal Court - 67 Targee Street, Staten Island, NY 10304

Defendant stated in my presence (*in substance*):

I personally observed the commission of the offense charged herein. False statements made herein are punishable as a Class A Misdemeanor pursuant to section 210.45 of the Penal Law. Affirmed under penalty of law.

Complainant's Full Name Printed | Rank/Full Signature of Complainant | Date Affirmed (*mm/dd/yy*)

Agency | Tax Registry # | Command Code

The person described above is summoned to appear at NYC Criminal Court located at: | Summons Part | County

Date of Appearance (*mm/dd/yy*) | At 9:30 a.m.

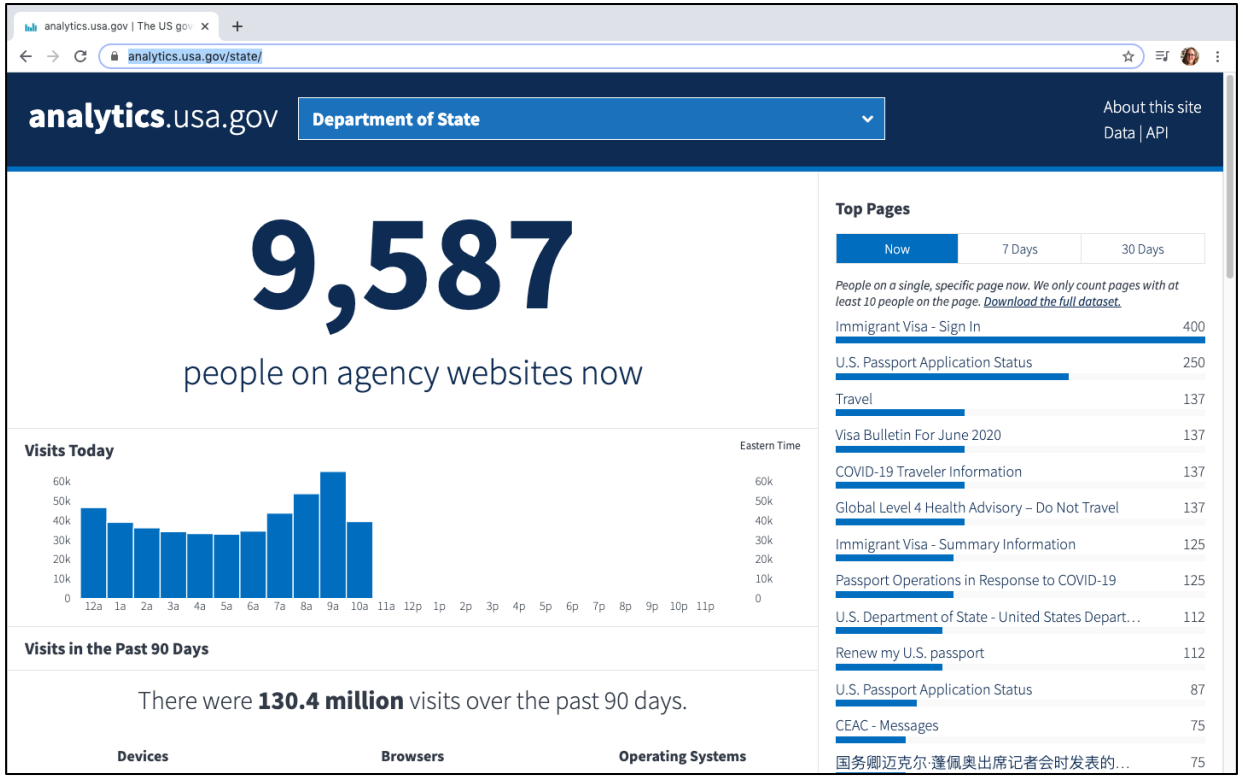
http://www.ideas42.org/wp-content/uploads/2018/03/Using-Behavioral-Science-to-Improve-Criminal-Justice-Outcomes.pdf

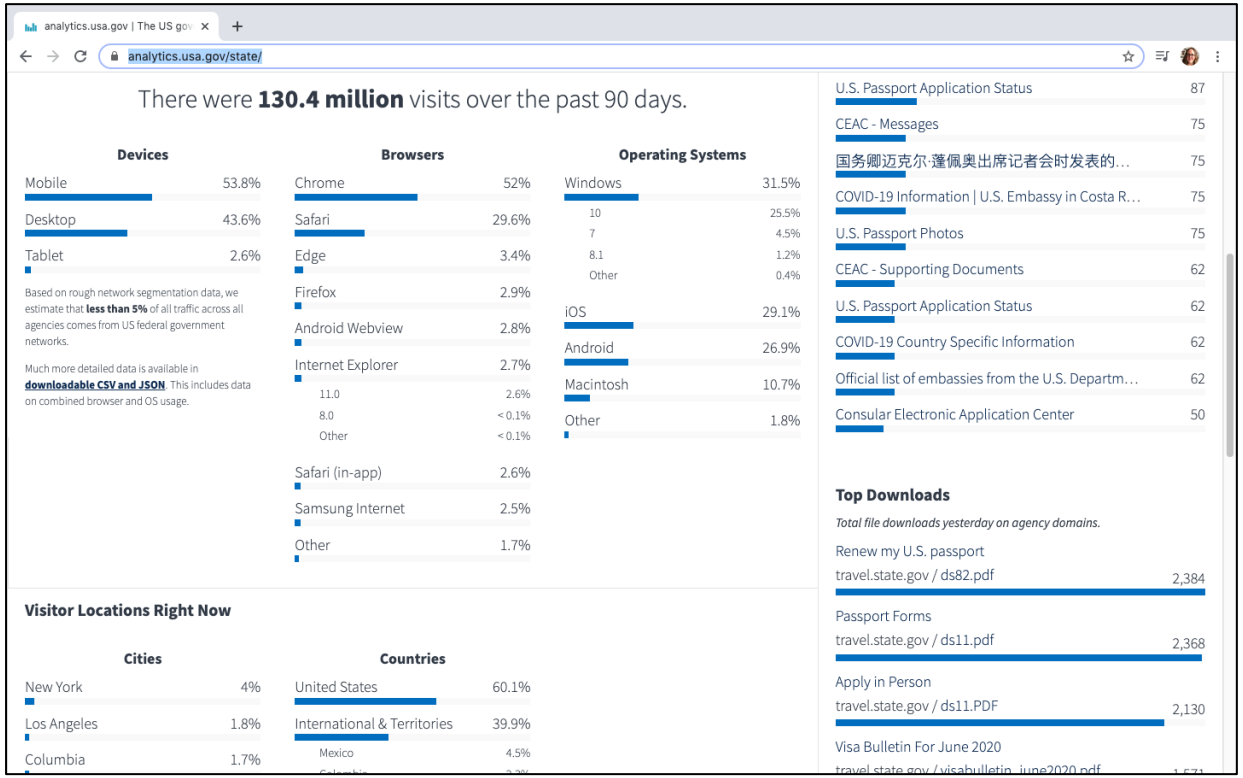
Poll question

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

analytics.usa.gov





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

Agenda

Day 1

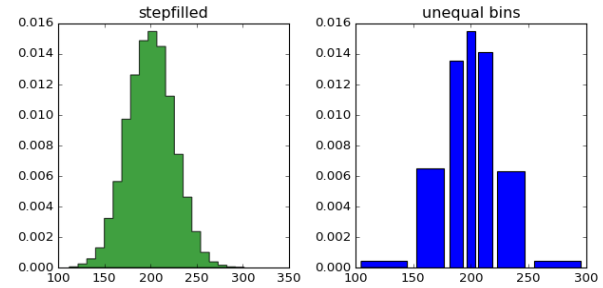
- Data and its uses
- Data analytics overview
- Data governance
- Data tools
- Data teams

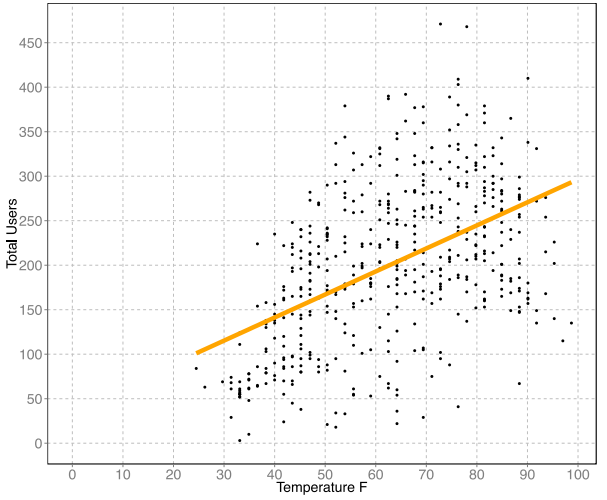
- What is data analytics and how can it be used?
- What are the principles of data science?

What is data analytics?

- Data analytics focuses on processing and performing 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.





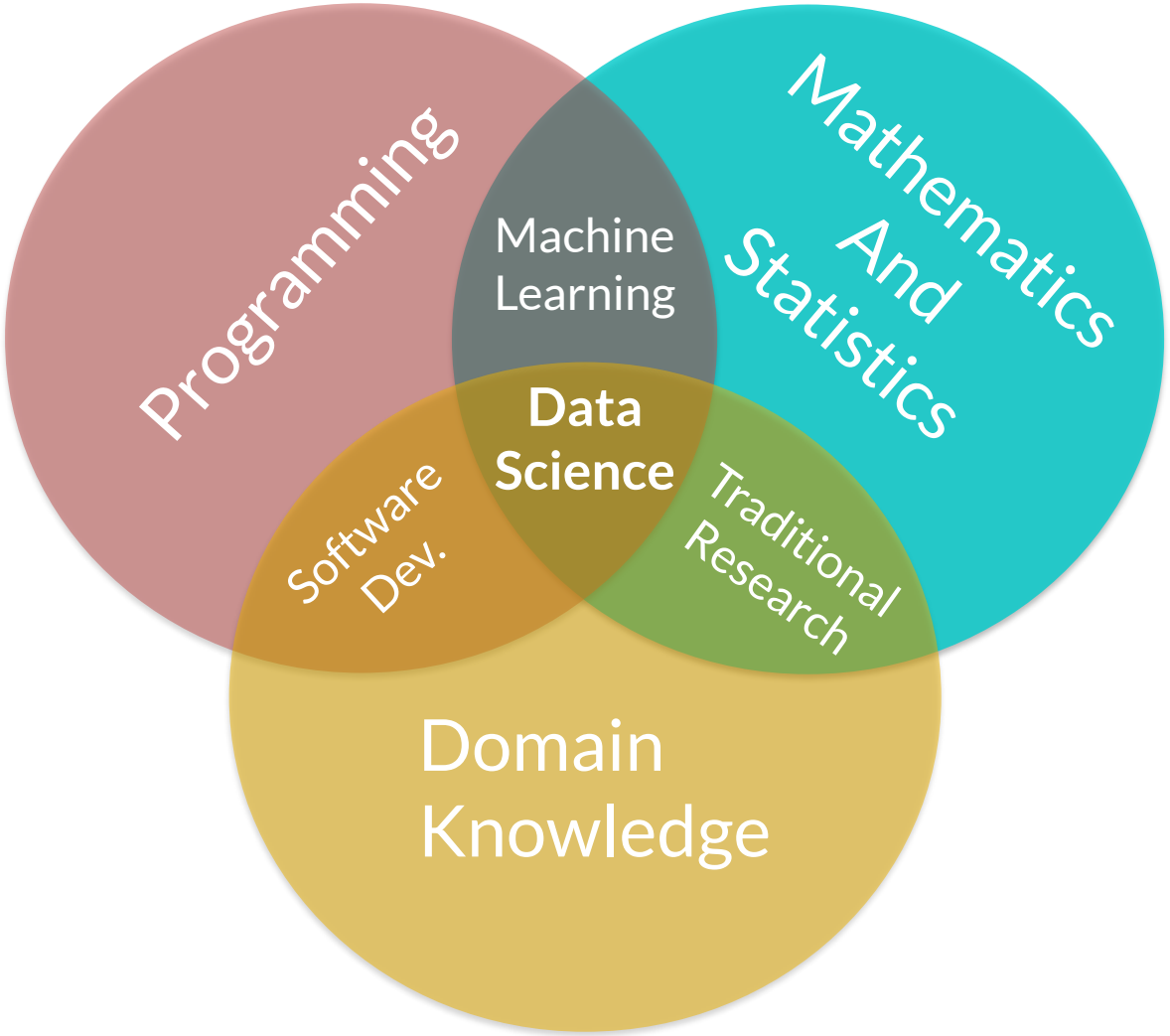
Poll question

Do you use data analytics within your organization currently?

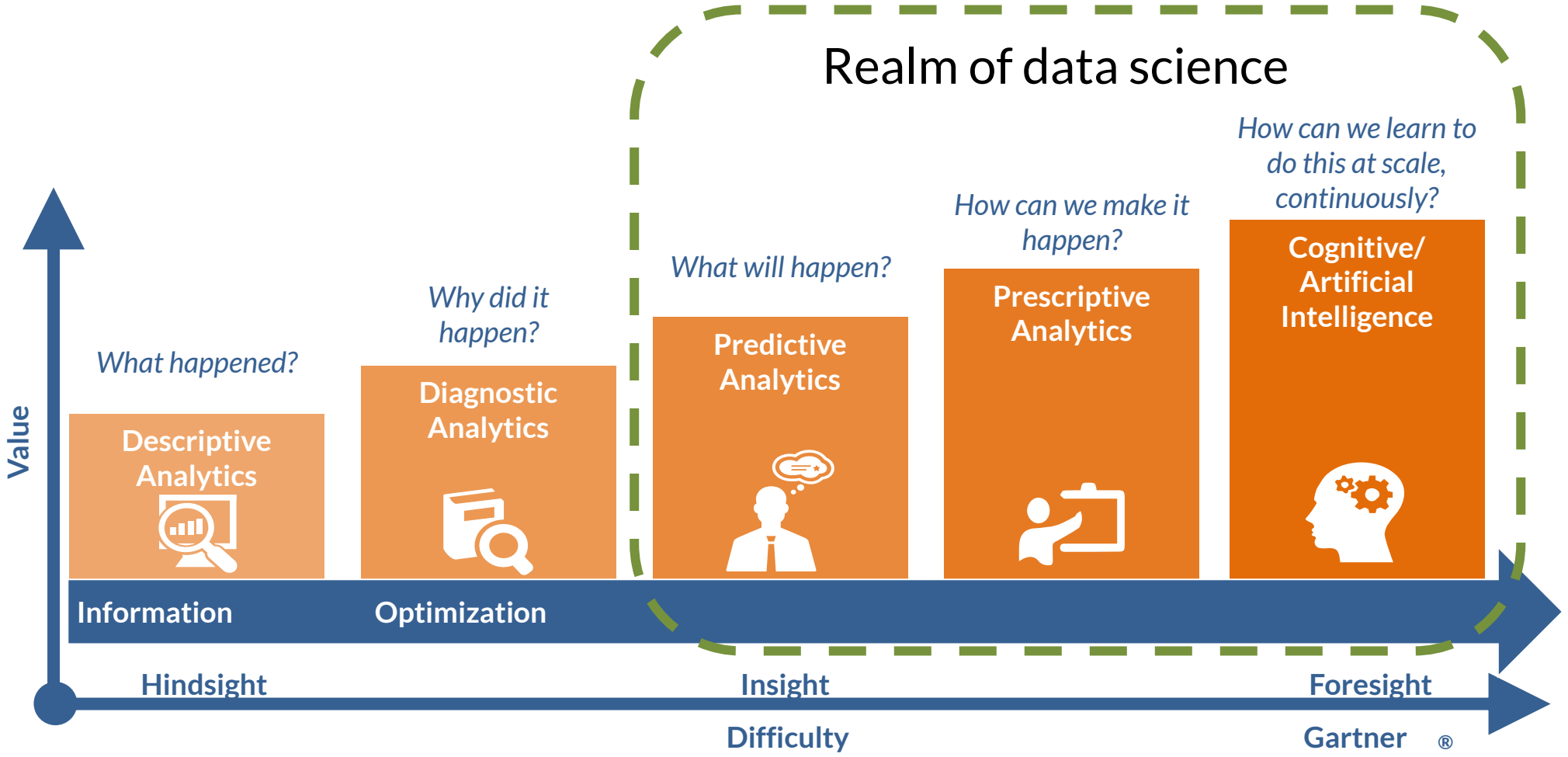
How do you use data analytics?

- Improve Business Processes
- Improve Customer Experience
- Reduce Cost
- Capture More Business

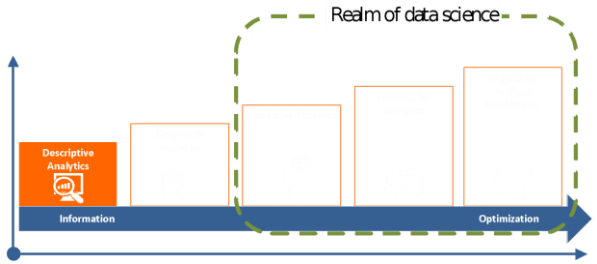
Principles


Programming
Mathematics And Statistics
Machine Learning
Data Science
Software Dev.
Traditional Research
Domain Knowledge

Data analytics maturity model

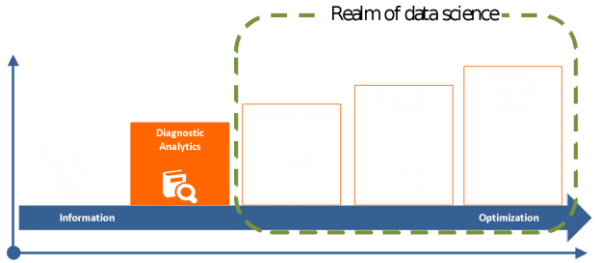


Stage 1: descriptive analytics



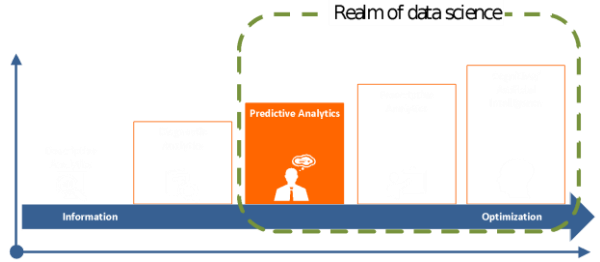
What questions does it answer?	*What has happened in the past?*
How valuable is it?	*Provides some value, but doesn't provide causation or prediction*
How labor intensive is it?	*Easy to deploy provided you have the right data*

Stage 2: diagnostic analytics



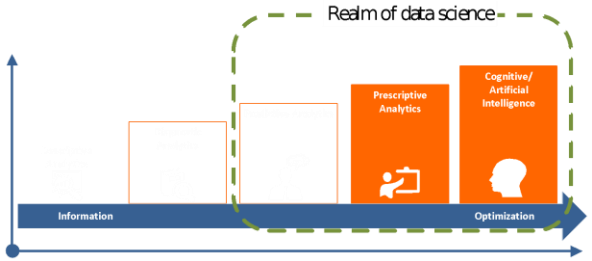
What questions does it answer?	*Why did something happen in the past?*
How valuable is it?	*Provides insights into a particular problem, and can help you identify some root causes for past trends and behaviors*
How labor intensive is it?	*Requires detailed data, but doesn't have to be overly intensive*

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*

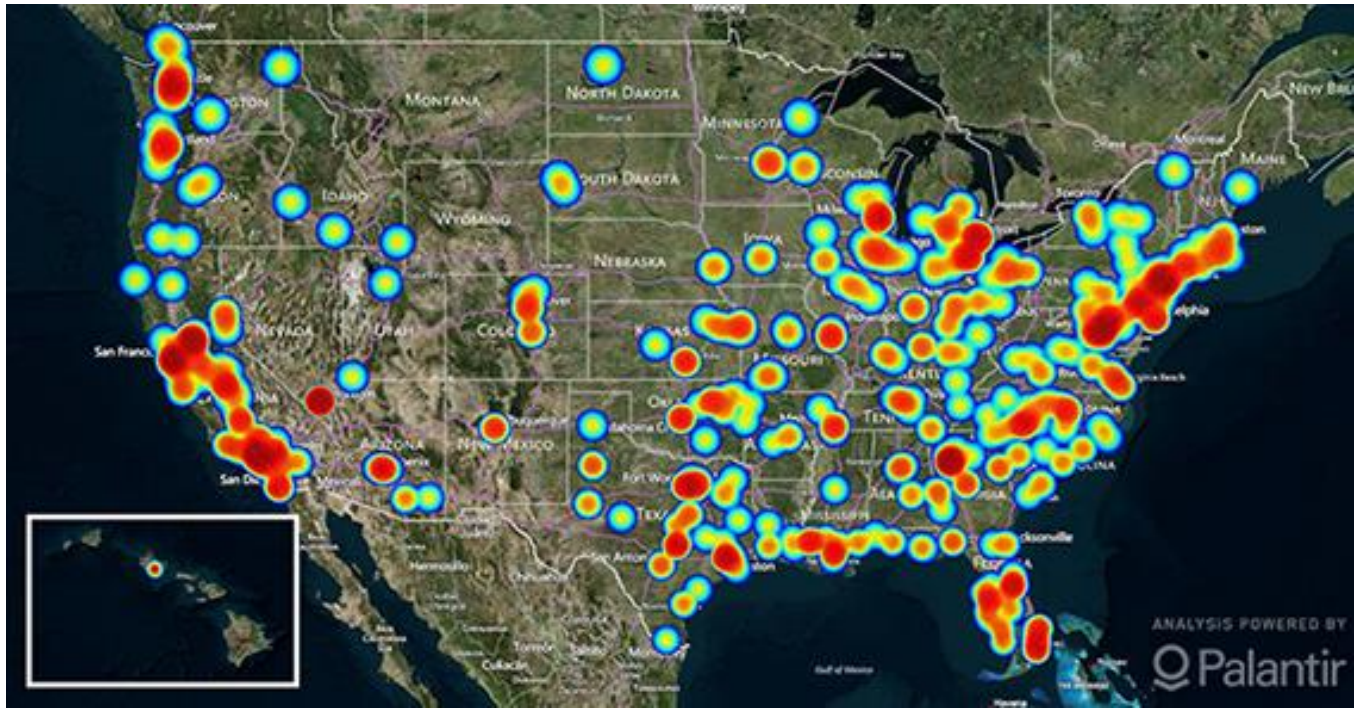
Stages 4,5: prescriptive analytics, AI



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



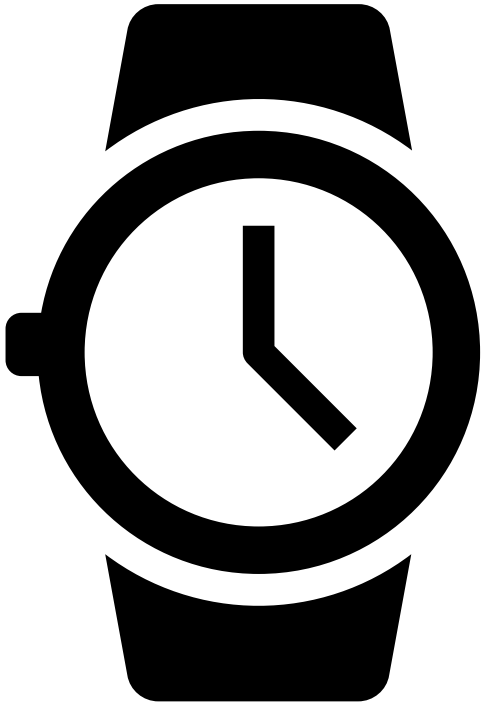
https://www.datanami.com/2016/10/07/data-analytics-fight-human-trafficking/

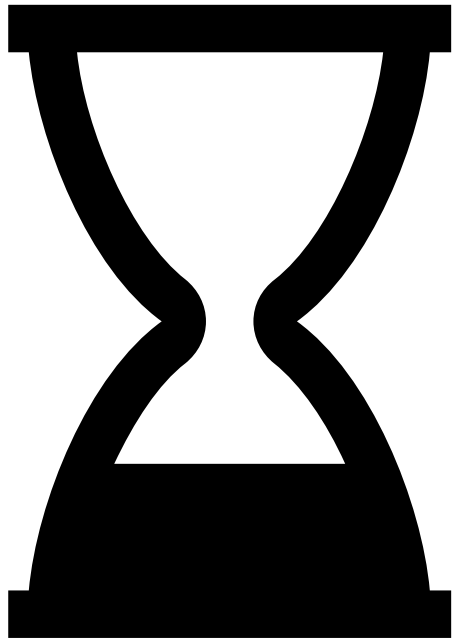
Poll question

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

- Descriptive
- Diagnostic
- Predictive
- Prescriptive

Break





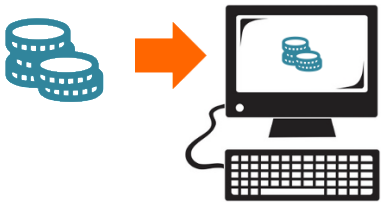
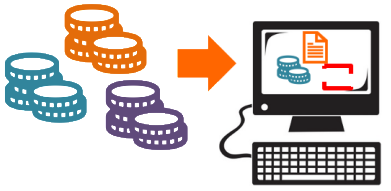
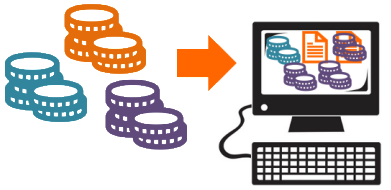
Agenda

Day 1

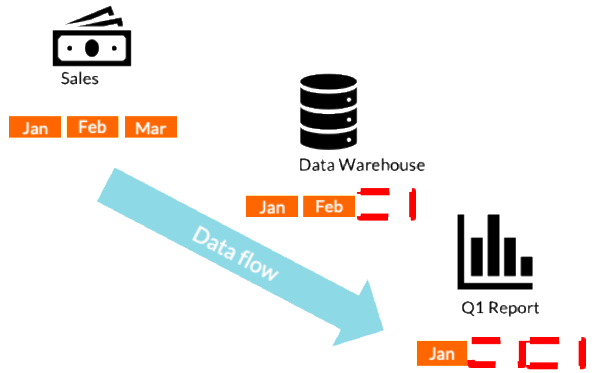
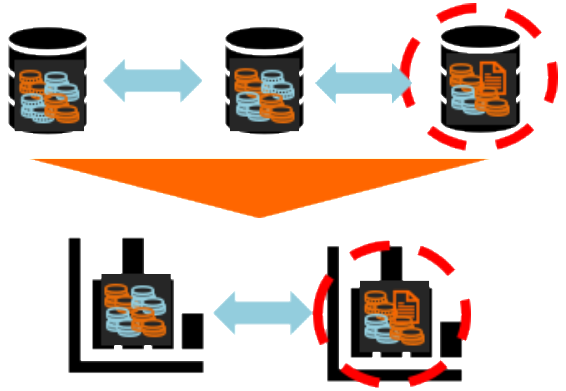
- Data and its uses
- Data analytics overview
- Data governance
- Data tools
- Data teams

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

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	

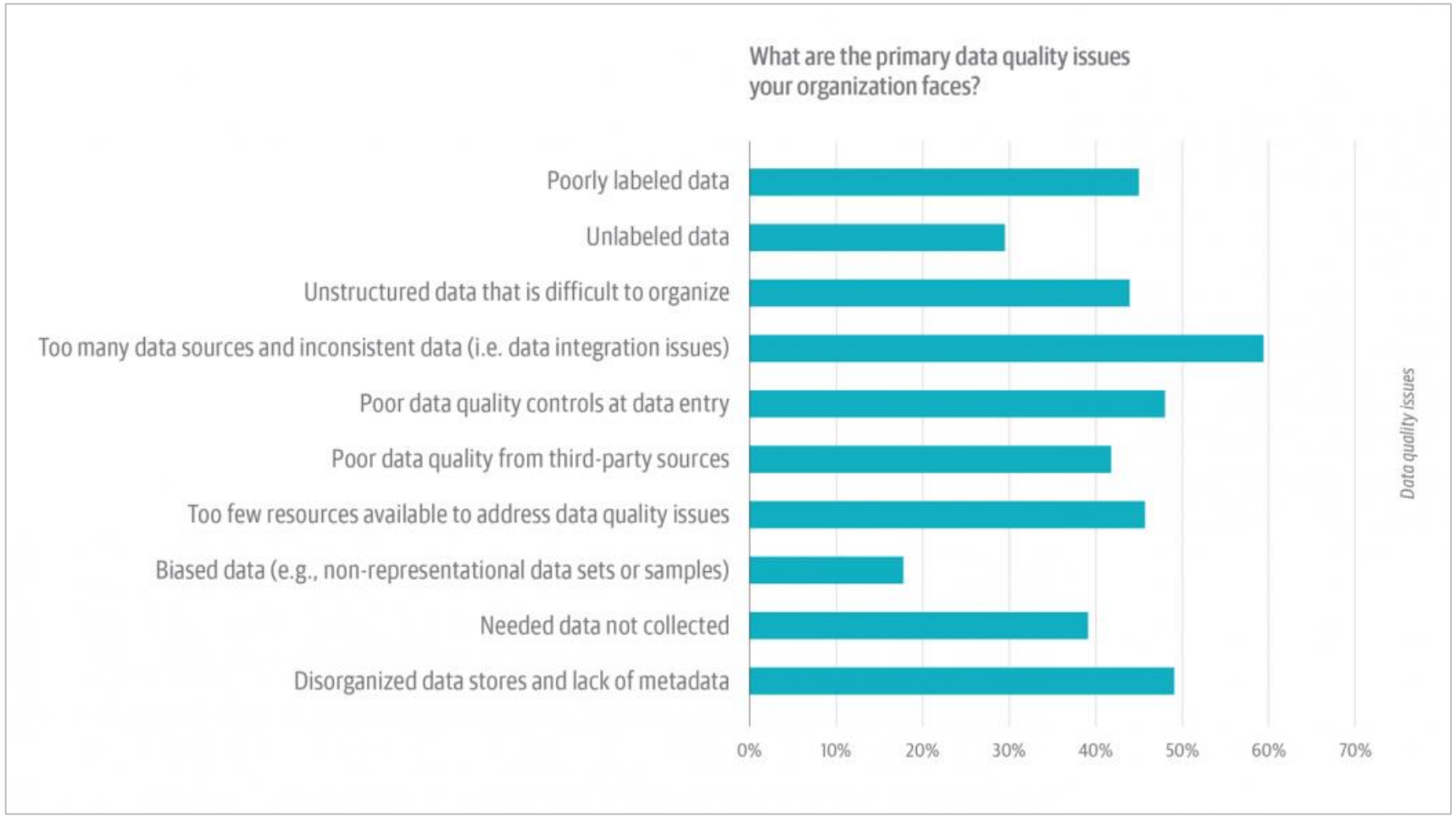
5 components of basic data quality

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	

Acquiring quality data is hard

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

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



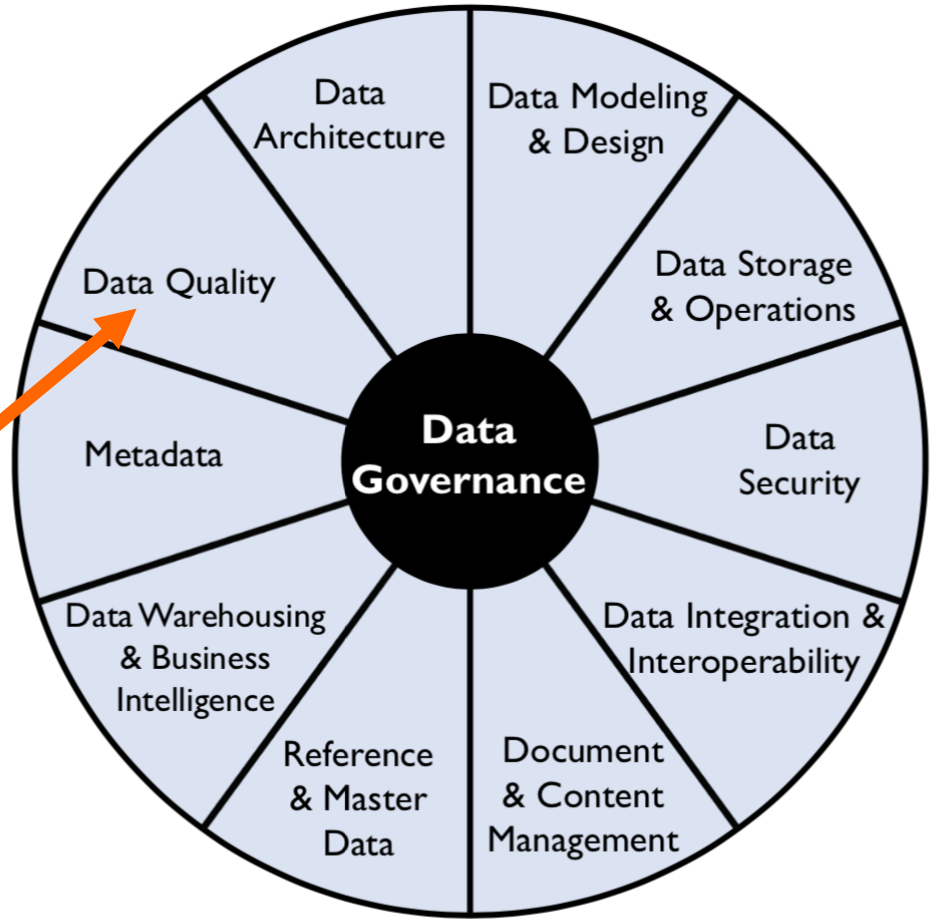
How to prevent data issues?

- 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

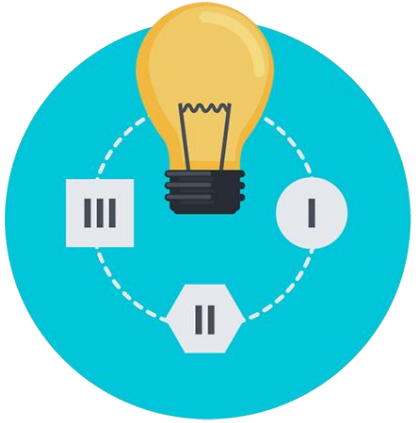
3. Measured – there should be some defined metrics to help demonstrate value to the organization

Data governance strategy

A data governance program might be documented using:

Charters

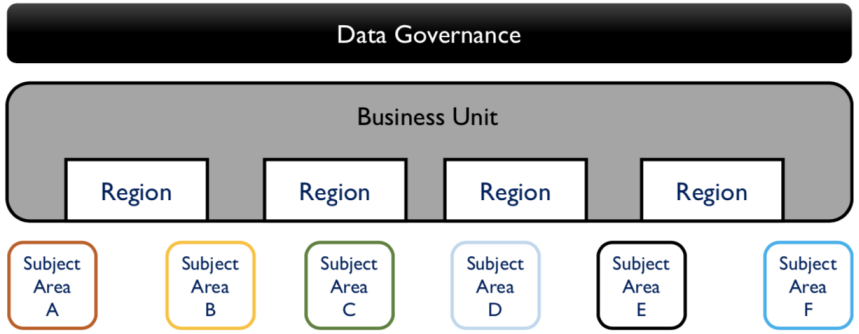
Implementation roadmaps



Operating frameworks / accountabilities

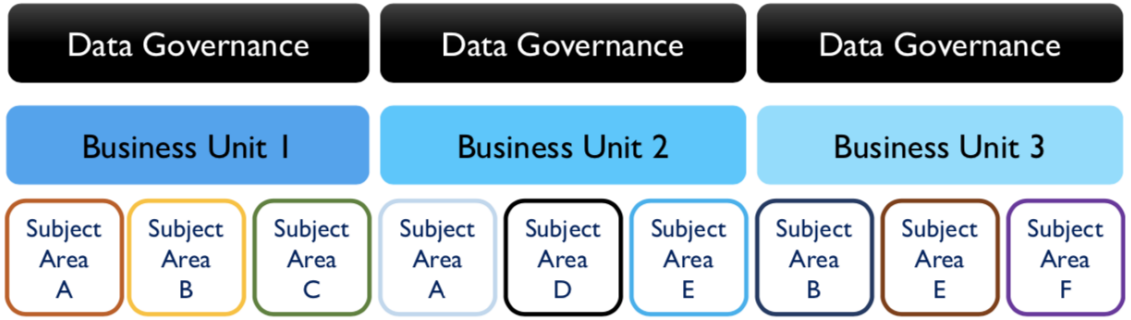
Plans for operational success

Data governance models



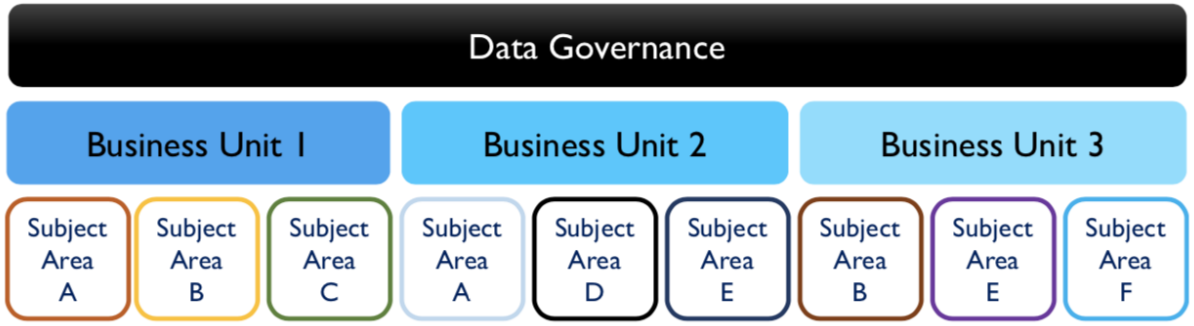
Centralized

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.

Poll question: data governance

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

- Centralized
- Replicated
- Federated
- None of the above

Poll question: data governance

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

- Centralized
- Replicated
- Federated
- None of the above

Activity: evaluate yourself!

- Turn to your participant guide to the **Data governance assessment**, which begins on page 3, to see how far along you and your team are in the data governance cycle.

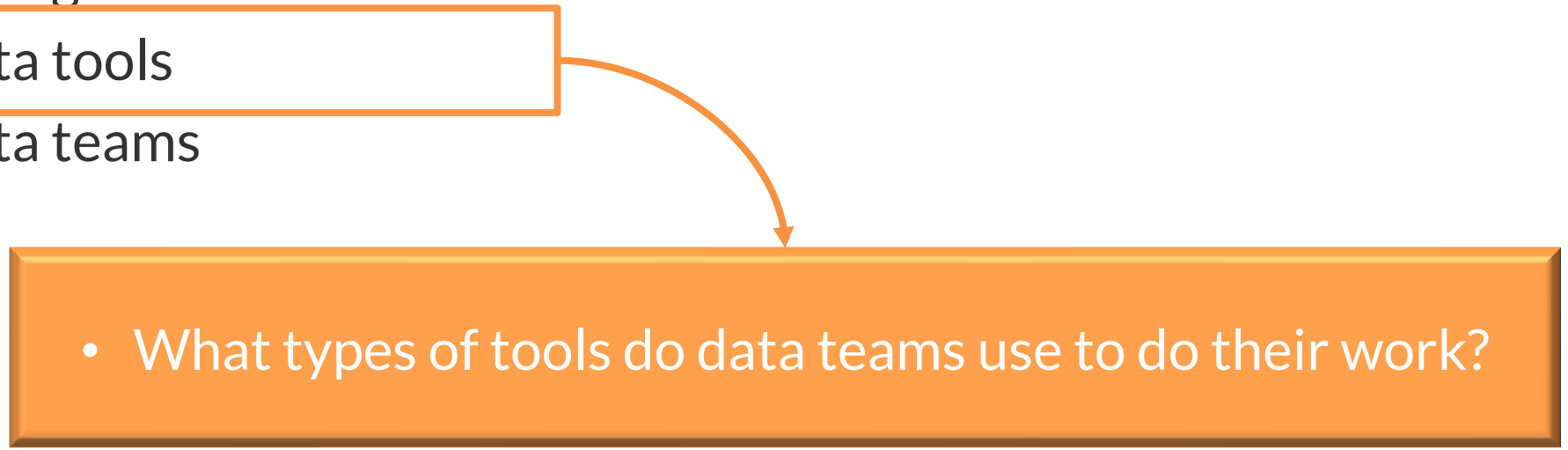
- You'll measure the foundational components, such as **awareness**, **formalization**, and **metadata**, as well as the project components of **stewardship**, **data quality**, and **master data policies.**

- Then, assess your progress and set goals for where you want your team.

Agenda

Day 1

- Data and its uses
- Data analytics overview
- Data governance
- Data tools
- Data teams



Tools


python


git


Spark


GitHub


tableau


Microsoft
Azure


THE
C
PROGRAMMING
LANGUAGE


iCloud


Google Drive


MySQL


RStudio


AWS GovCloud (US)


amazon
web services


MATLAB


sas


box


R


Microsoft
SQL Server


hadoop


mongoDB


JavaScript


Scala


X


Microsoft
Power BI

Poll question

What data tools does your organization use?

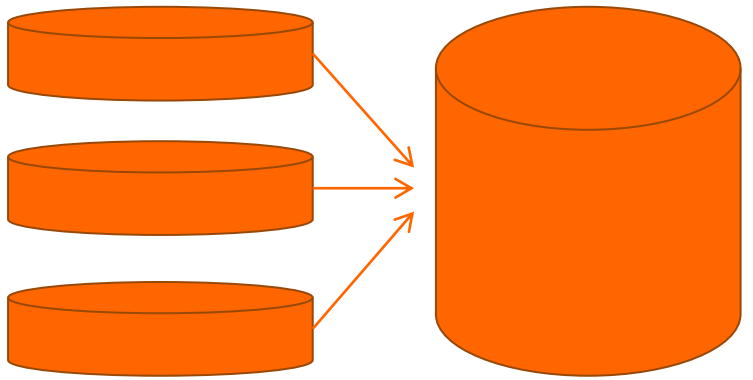
Spark, Python, R, Scala, Excel, PowerBI, Tableau, Hadoop, mongoDB

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



Extract, Load, Transform (ETL)

- ETL is the process of migrating and transforming data from one system into another often for downstream analysis
- 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, Informatica, Java, Spark, ApacheAirFlow, etc.

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.

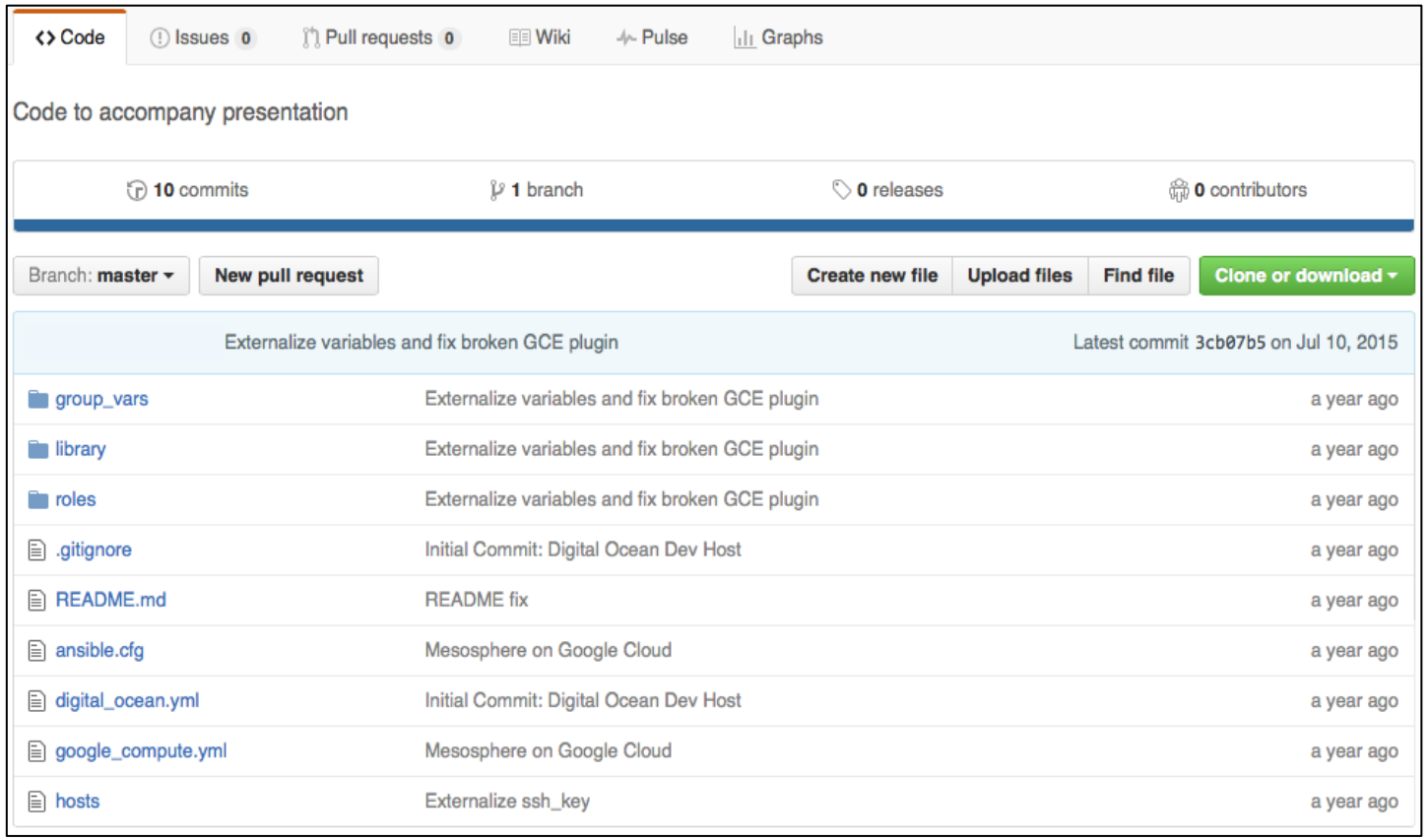
- Example tools: Excel, R, Python

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?

DATA SOCIETY®

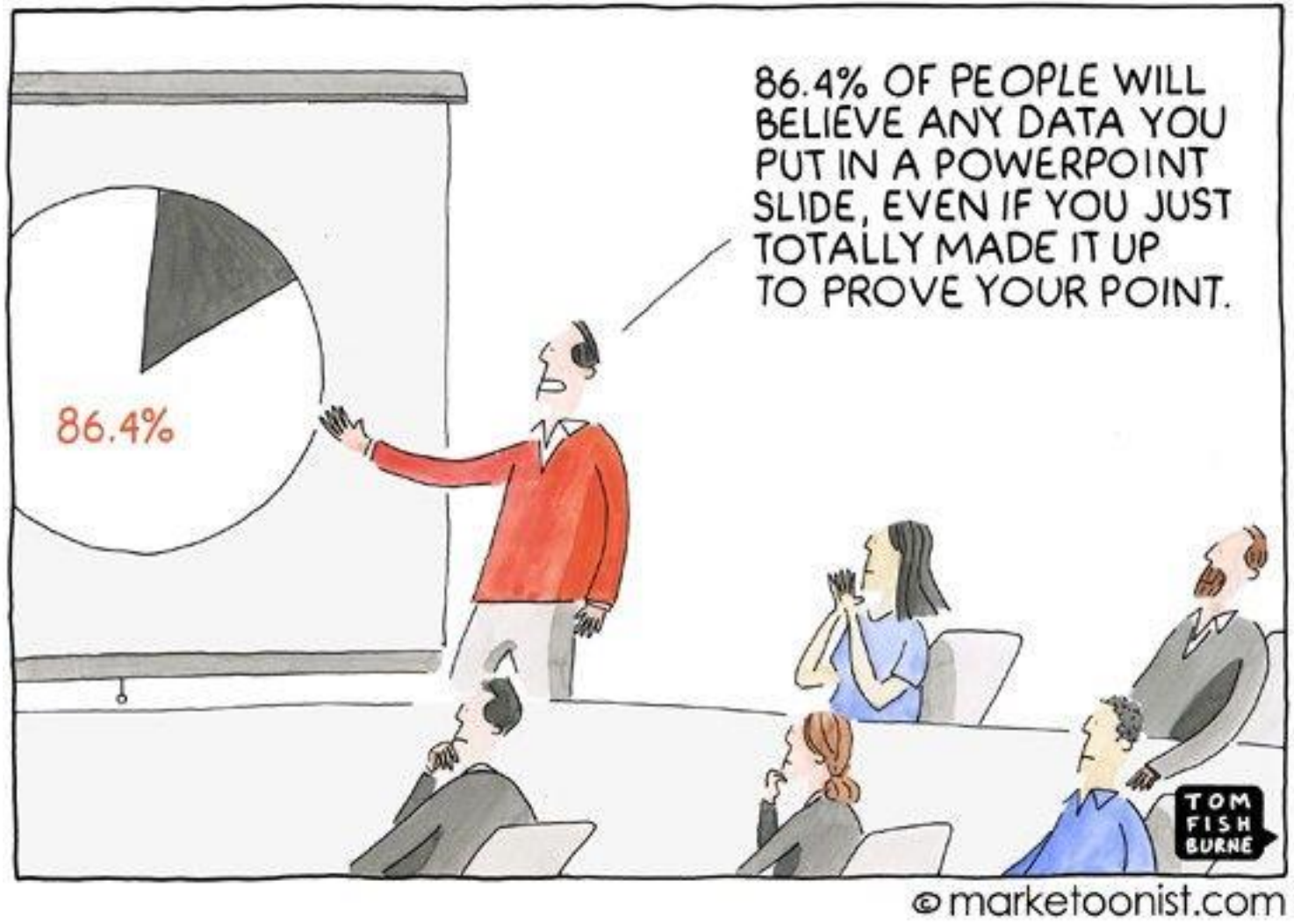
Questions?

End of Day 1

DATA SOCIETY®

Day 2

Welcome back!



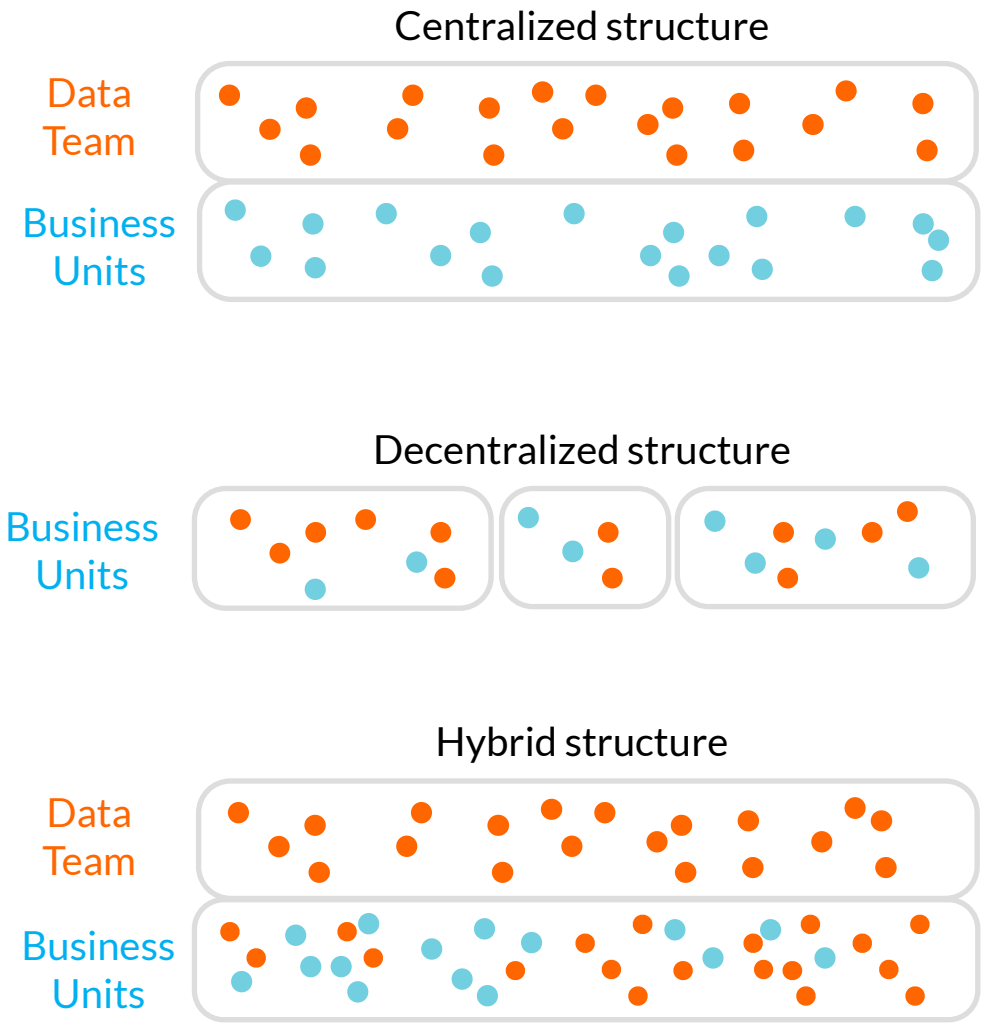
Agenda

Day 2

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

- How are data teams structured?

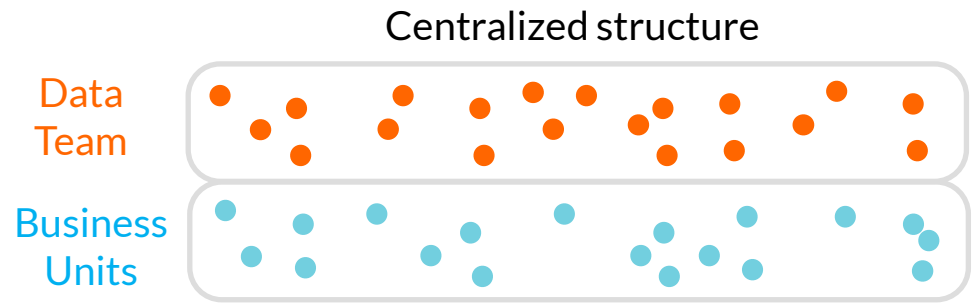
Team structures

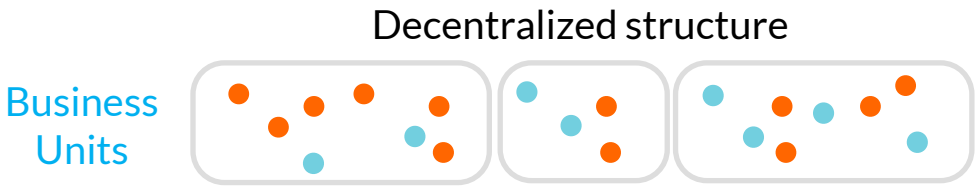


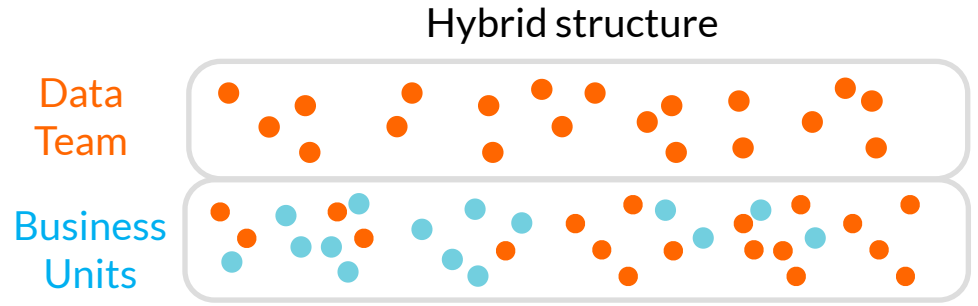
Poll question

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

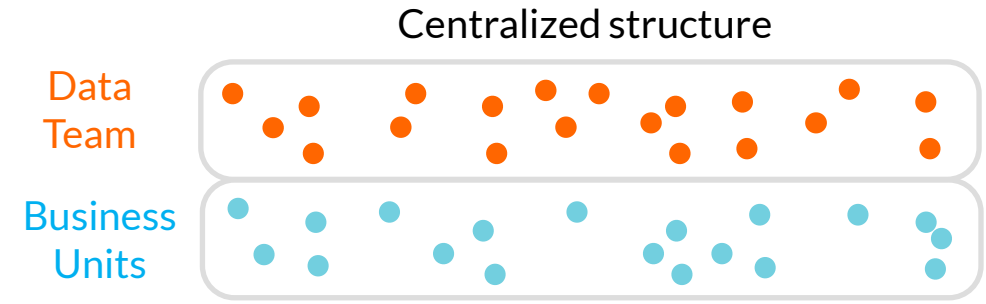
- Centralized
- Decentralized
- Hybrid



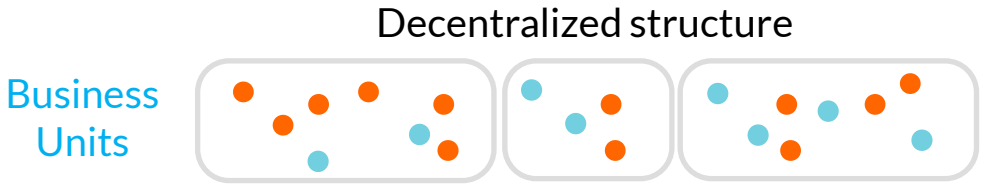




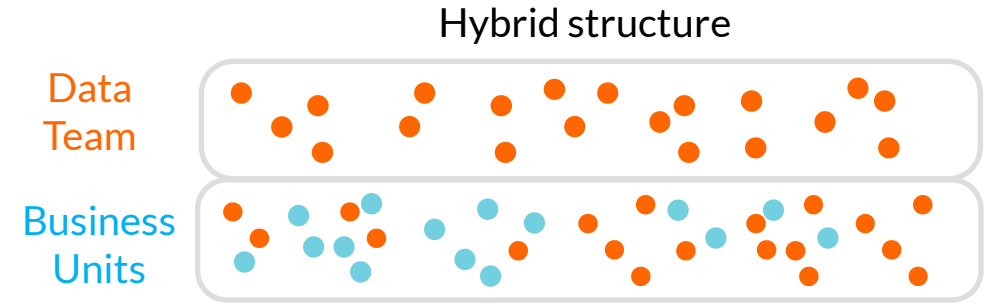
Pros and cons



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



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



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

Another option… contracting

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

Poll question

- What would be the best option for your organization?
  - Contracting a team
  - Hiring a team

- What are the key factor(s) in making that decision?
  - Recruitment/ training time
  - Cost
  - Internal know how
  - Flexibility
  - Not depending on outside forces

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

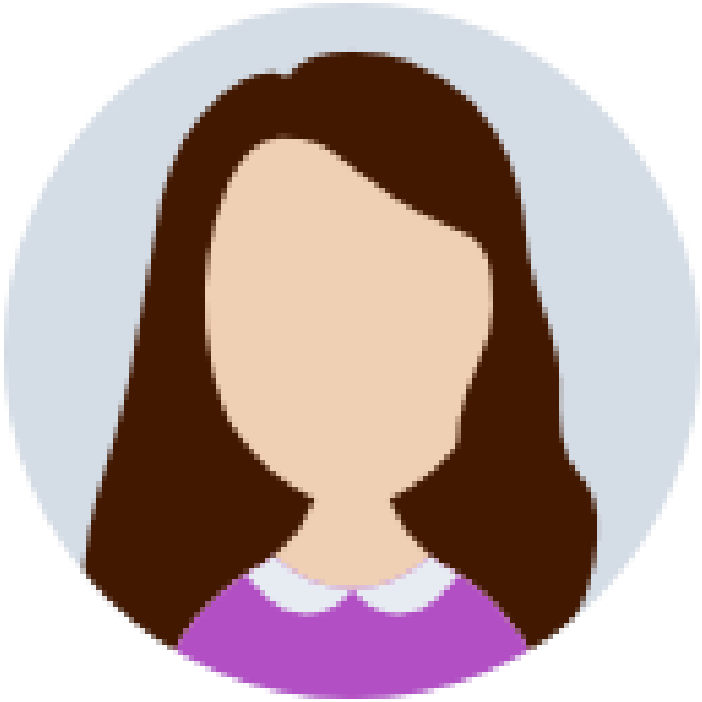
- Core skills to include: SQL, R / Python, Tableau / Power BI

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

ML Engineer

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

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

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

Data science manager

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

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

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

Agenda

Day 2

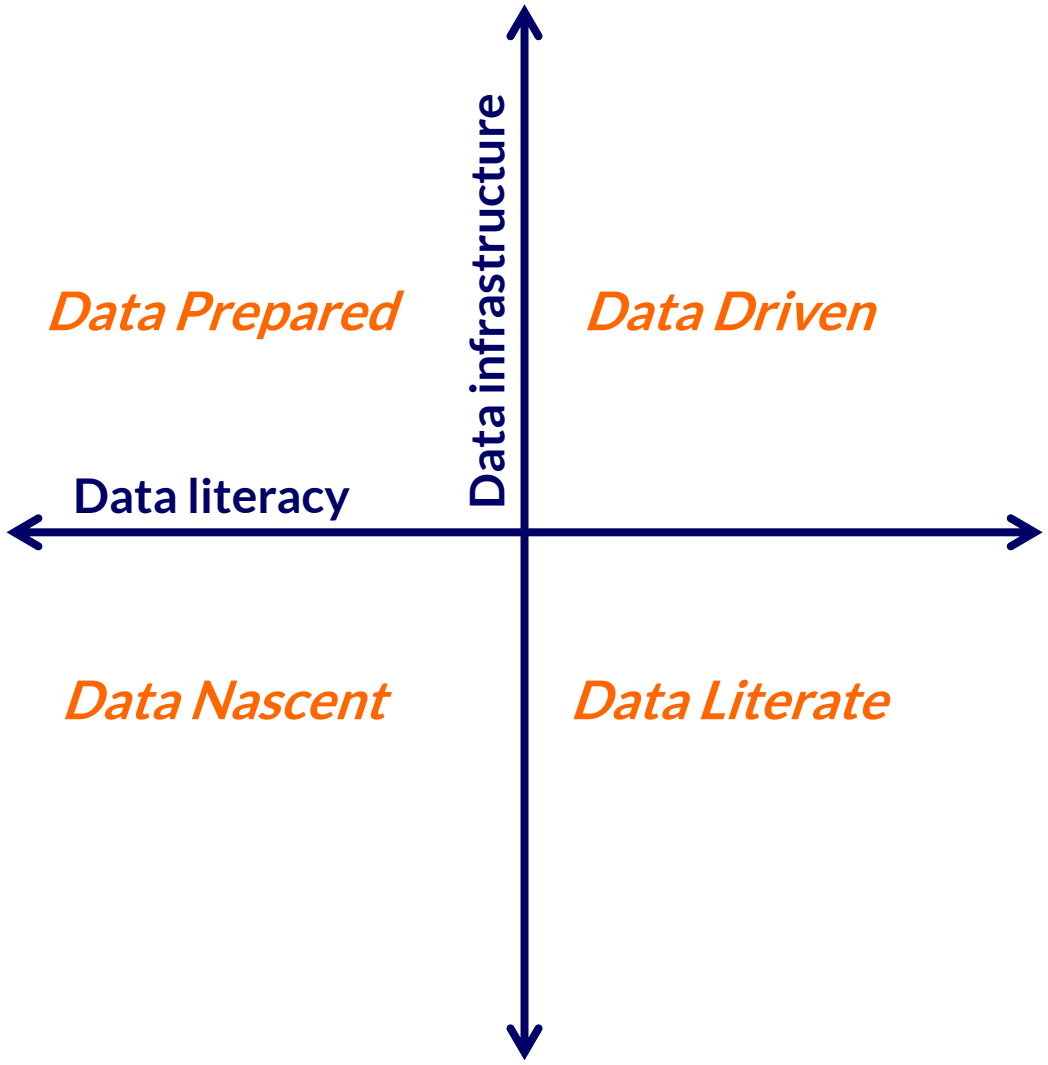
- Building a data-driven culture
- Data ethics
- The data science process
- Putting together a project

- What is a data-driven culture and why is it important?
- How can I build awareness about the importance and uses of data?
- How can I make data-driven decision-making routine?

What is a data-driven culture?

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

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



Data infrastructure

- Components of data infrastructure include:

DATA ACCESS

Can staff access data easily and in a timely manner?

DATA STORAGE

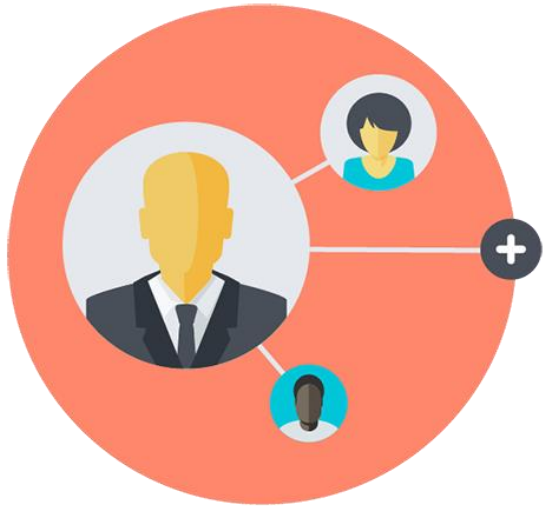
Is the data stored securely with a backup?

DATA COLLECTION

Is data collected in a timely and clean way?

Data literacy

- Components of data literacy include:



DATA LEADERSHIP

Do executives champion data usage?

DATA GOVERNANCE

Are staff aware of data standards and practices?

DATA KNOWLEDGE

Does staff understand how to ask questions of data?

Why is it important to be data driven?

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

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

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

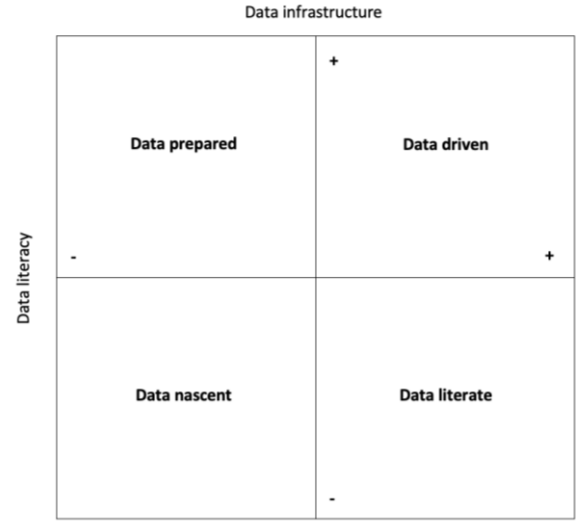
A study from the MIT Center for Digital Business found that organizations driven most by data-based decision making had 4% higher productivity rates and 6% higher profits.

Activity: Are you data driven?

- Turn to pages 6-8 of your participant guide to the **Data-driven culture assessment** to evaluate your team.

Component: Data infrastructure		
		Points
Data access	1. I can easily access the data I need without asking others for help. a) Not at all b) Only for some colleagues c) Only for some teams d) Organization-wide	
	2. I can easily access the data I need in a timely manner. a) Not at all b) Only for some data c) Only for data in my team / related teams d) Organization-wide	
Data collection	3. Data is automatically collected and stored on a continuous basis. a) Not at all b) Only at someone's request c) Regularly, a few times a year d) There is continuous data collection	
	4. The data we have is accurate and good quality (few missing entries, few duplicates, accurate measurements). a) Not at all b) Only for some data c) Only for data in my team / related teams d) Organization-wide	
Data Storage	5. Our data is stored securely either internally or offsite. a) Not at all b) Only for some data c) Only for data in my team / related teams d) Organization-wide	
	Total	

Component: Data literacy		
		Points
Data literacy	1. My company routinely offers data trainings and other educational opportunities. a) Not at all b) Occasionally c) Regularly, a few times a year d) There are continuous learning opportunities	
	2. Most of my colleagues understand the importance of data. a) Not at all b) Only for some colleagues c) Only for some teams d) Organization-wide	
Data governance	3. Our organization has a set of data standards that reviews how data should be collected, stored, and analyzed. a) Not at all b) Only for some colleagues c) Only for some teams d) Company-wide	
Data leadership	4. My organization emphasizes the importance of using data to track initiatives. a) No one b) A few people across the company c) Some teams across the company d) Organization-wide	
	5. I am expected to present data metrics when I explain conclusions and decisions. a) Not at all b) Only for some colleagues c) Only for some teams d) Organization-wide	
	Total	



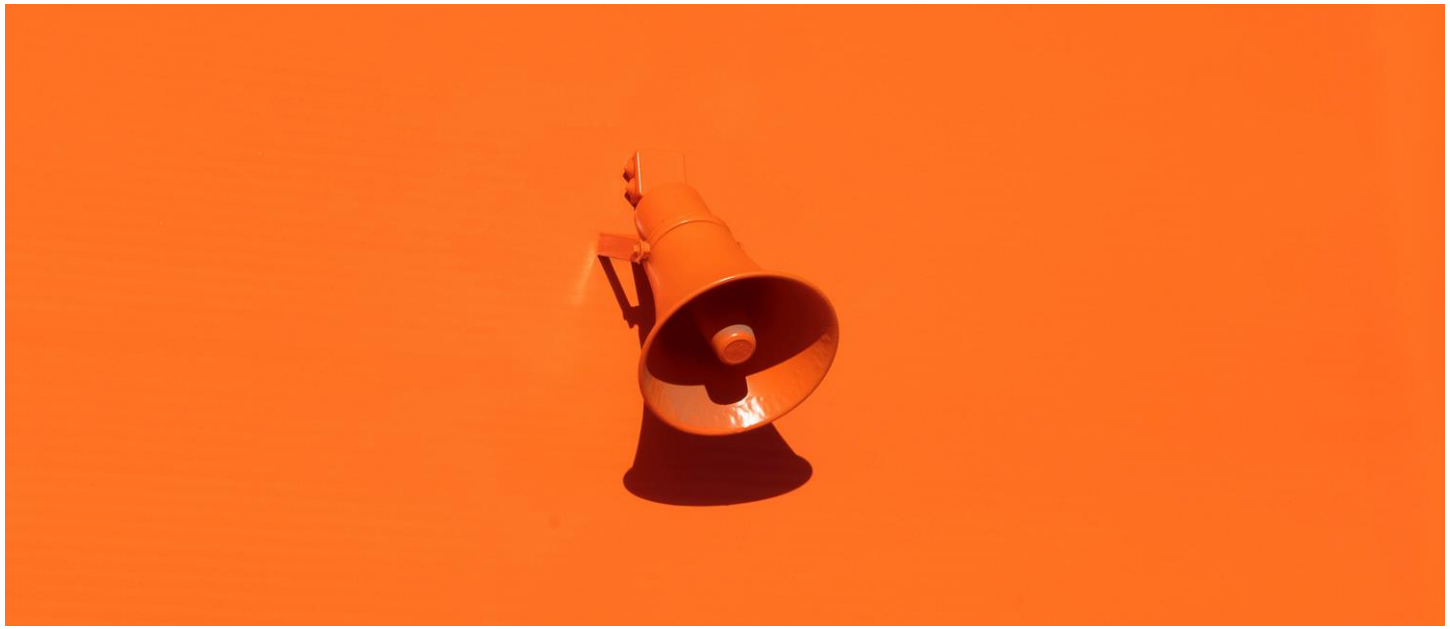
Building awareness

- Step 1: find a champion (or be the champion!)

Building awareness

- Step 1: find a champion (or be the champion!)

- Step 2: identify a successful analytics project / team, and highlight their success through a newsletter, event, or lunch and learn



Building awareness

- Step 1: find a champion (or be the champion!)
- Step 2: identify a successful analytics project / team, and highlight their success through a newsletter, event, or lunch and learn
- Step 3: once there is more interest, offer additional data trainings (like this one!) to develop a common data vocabulary and empower staff

Building awareness

- Step 1: find a champion (or be the champion!)
- Step 2: identify a successful analytics project / team, and highlight their success through a newsletter, event, or lunch and learn
- Step 3: once there is more interest, offer additional data trainings (like this one!) to develop a common data vocabulary and empower staff
- Step 4: build upon the community of practice that will develop from the trainings to bring wider awareness and more buy-in from executives and managers across the organization

Make it routine

- Lead by example!
- How can you adjust your practices now to reflect a data-driven mindset?

1. Ask for the metrics / analysis summary behind conclusions and reports.
2. Demonstrate data-driven decision-making during meetings.
3. Highlight data-driven team members or successful analyses.

What else can you do?

1. Bring in external / internal experts for "lunch and learns."

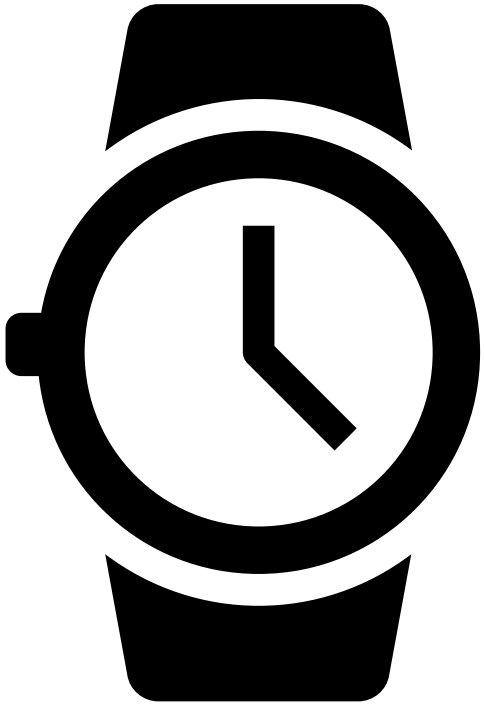
2. Attend and send team members / colleagues to data conferences.

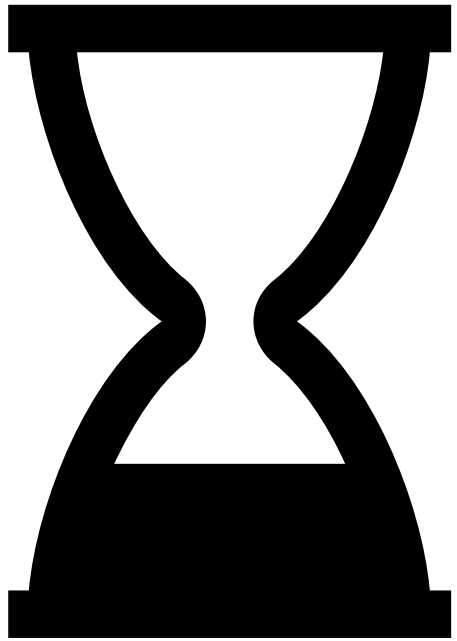
3. Plan an event, such as a data competition.

Remember!

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

Break





Agenda

Day 2

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

- What is data ethics? What role does it play in good data governance?

What is data ethics?

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

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

Source: University of Oxford

Why data ethics?

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

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

Source: University of Oxford

FDS: Ethical Governance

Uphold Ethics	• Monitor and assess the implications of federal data practices for the public. Design checks and balances to protect and serve the public good.
Exercise Responsibility	• Practice effective data stewardship and governance. Employ sound data security practices, protect individual privacy, maintain promised confidentiality, and ensure appropriate access and use.
Promote Transparency	• Articulate the purposes and uses of federal data to engender public trust. Comprehensively document processes and products to inform data providers and users.

FDS: Ethics Framework

- By December 2020, GSA will develop a data ethics framework to help agencies systematically identify and assess the potential benefits and risks associated with the data they acquire, manage, and use.

- Initial work has already begun with a review of data ethics frameworks developed by other countries, organizations, and advocacy groups to identify common elements and themes.

Federal Data Strategy

Leveraging Data as a Strategic Asset

Existing frameworks

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

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

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

Data Society guidelines

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

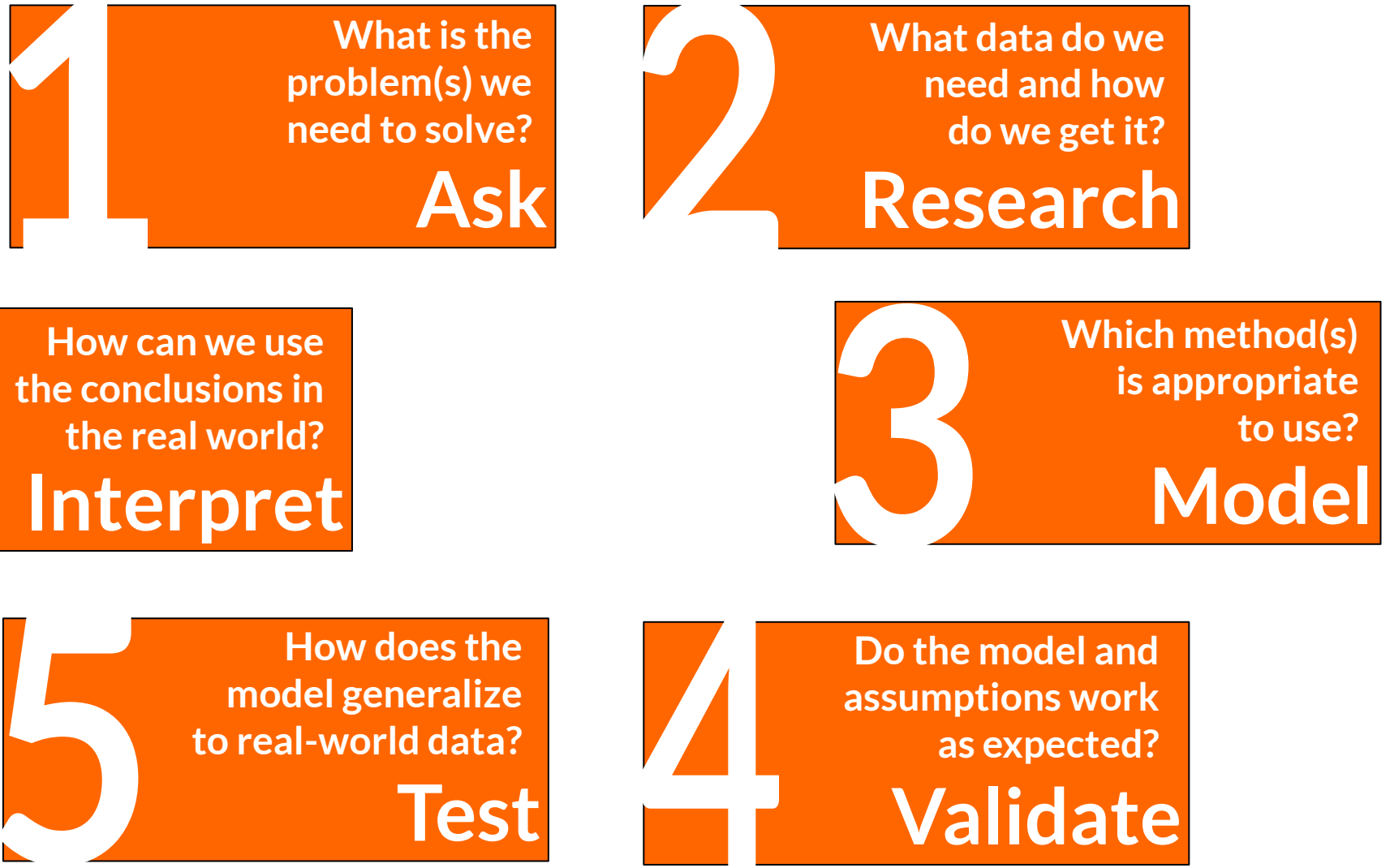
Agenda

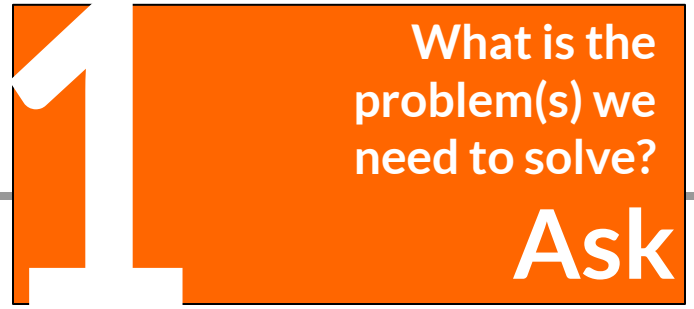
Day 2

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

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

Typical data science process





- The question should be specific, measurable, and objective. This is where 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

- Be specific about what type of data you need in order to get a relevant answer.
- *Is it already collected, or do you need time to get it? What format is it in?*

Examples

I'm sure we have the data somewhere.	→	We'll use the datasets from the policy report that can be found in X repository.
I'm sure the data is good enough as is.	→	Where can I read about how the data was collected and how the metrics are defined?

3 Which method(s) is appropriate to use?

Model

4 Do the model and assumptions work as expected?

Validate

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

Test

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

6

How can we use the conclusions in the real world?

Interpret

- Look at what the results are telling you—not what you were expecting the results to be.
- Make recommendations based on data and domain knowledge.
- *Who are the stakeholders? How should I visualize the results? What story can I tell with this data?*

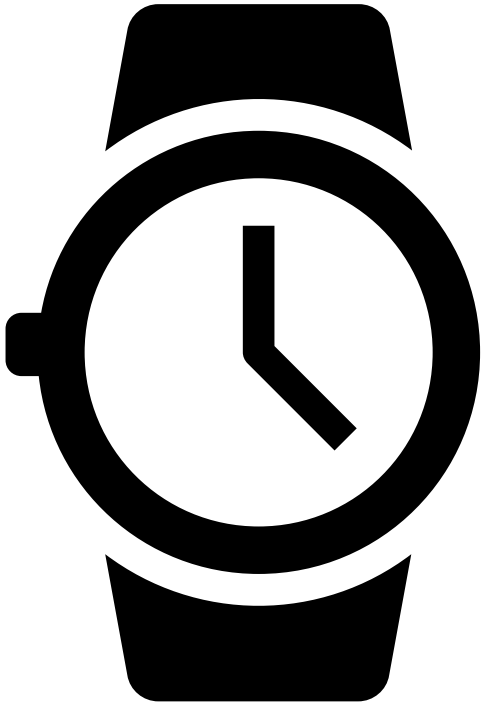
Example

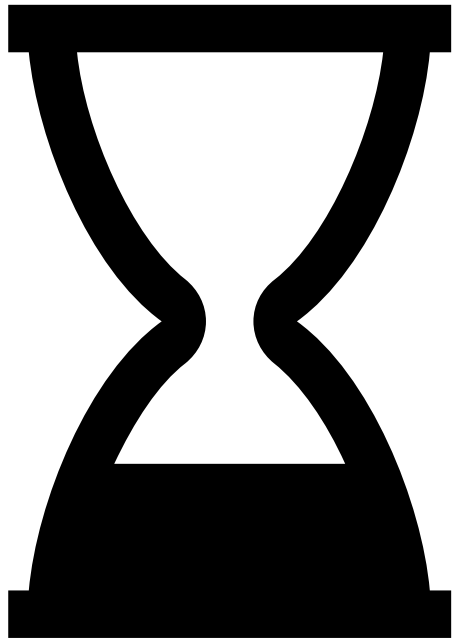
I'll put the results in the same format as I usually do. → How can I best convey the results that matter most to my end users?

Poll question

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

Break





Agenda

Day 2

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

- How do I identify feasible and impactful data projects?

Activity: brainstorm ideas

- Turn to pages 9-11 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.

DATA SOCIETY®

Questions?

End of Day 2

DATA SOCIETY®

Day 3

World's Smartest Home

Agenda

Day 3

- Foundational data science methods
- Advanced data science methods

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

Why learn about these methods?

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

2. To direct data science projects and make recommendations

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

What's an algorithm?

What is machine learning?

- 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 are categorized as either supervised or unsupervised

Supervised learning

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

Unsupervised learning

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

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

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

Poll question

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

Do you think this statement describes supervised machine learning?

Poll question

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

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

Before we go further…

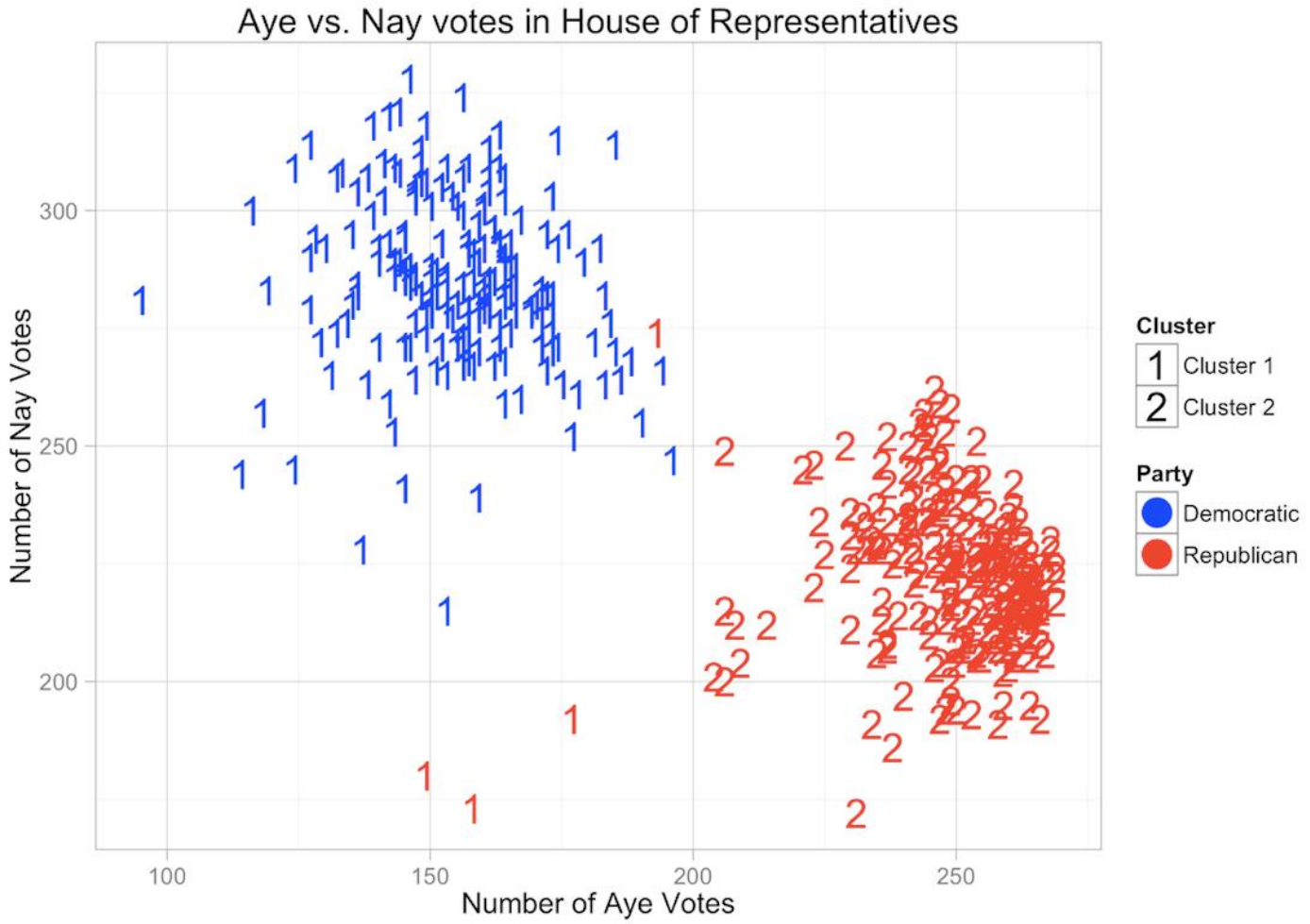
- Remember that most data science projects combine a few methods to extract the full picture.

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

Clustering

Clustering

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



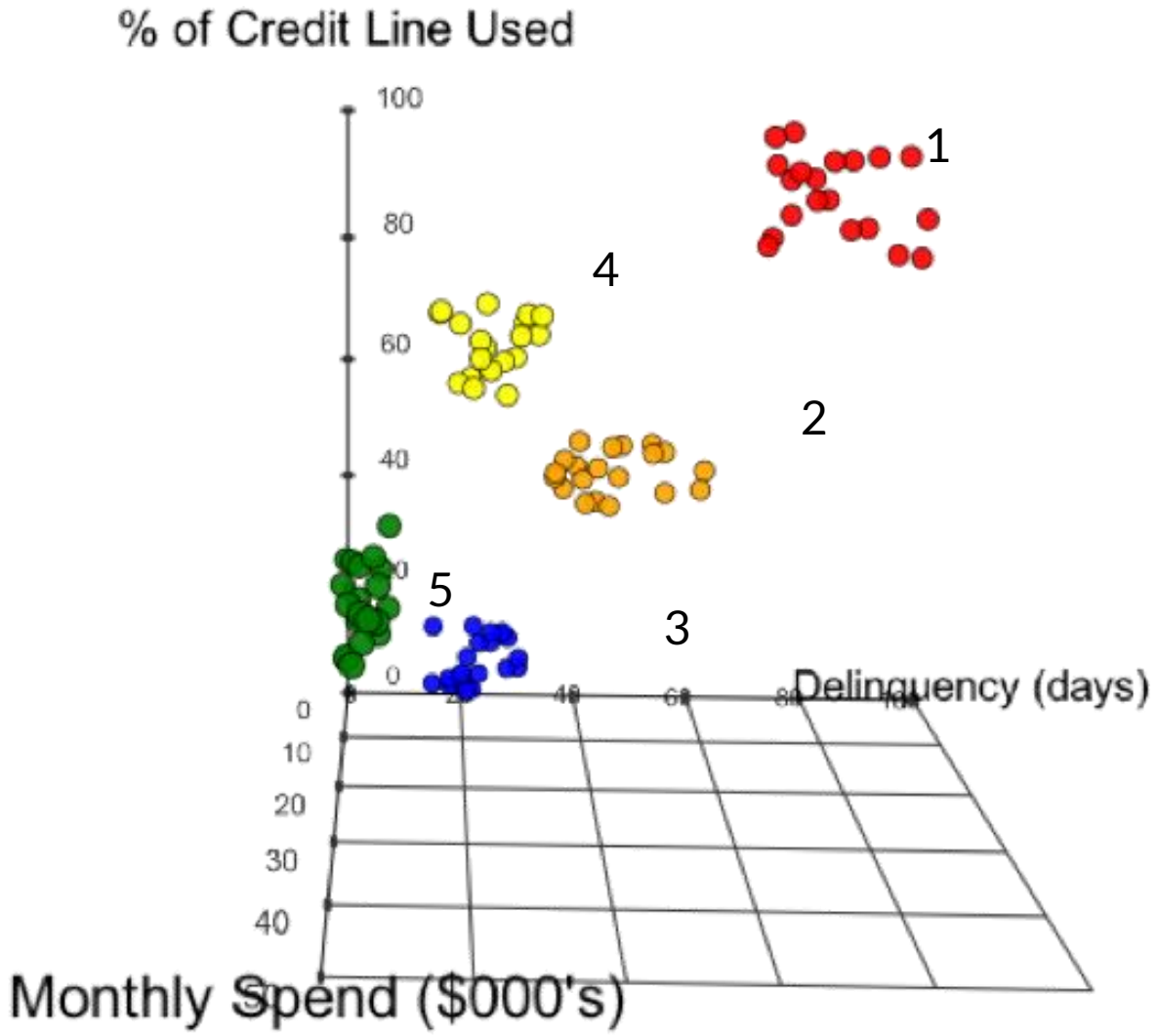
How can you use clustering?

- Clustering answers the questions:
  1. To whom or what is this person/object similar?
  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!
  - Based on the results of clustering analysis, we could model the average effectiveness of marketing campaign by demographics.
  - Clustering can identify unintuitive groups or pairings.

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.

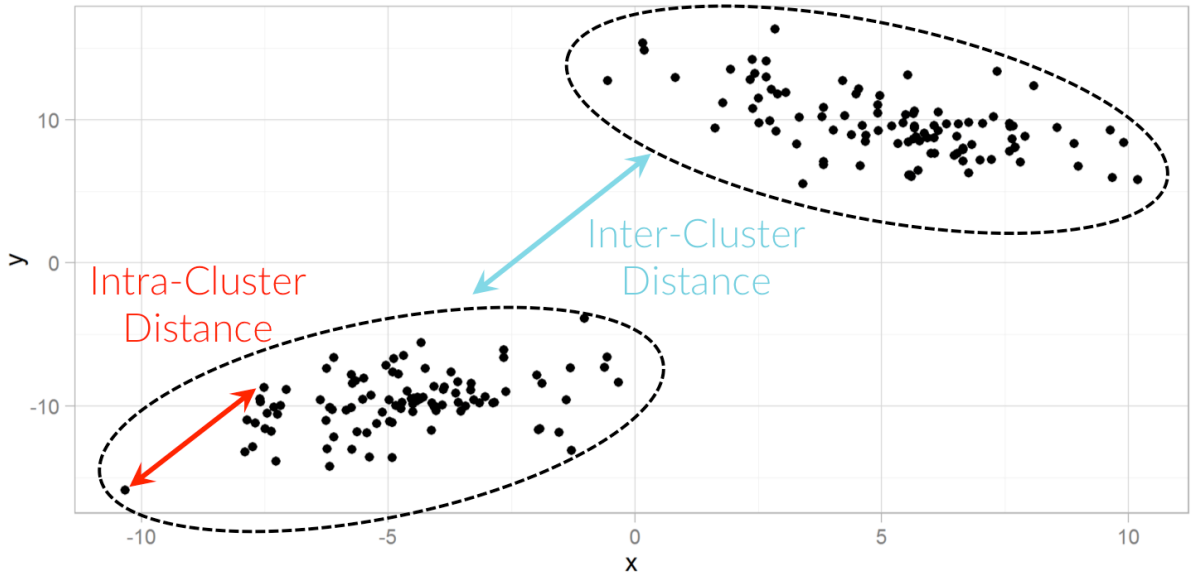


Evaluating the accuracy of the model

- 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

Variation explained by clusters

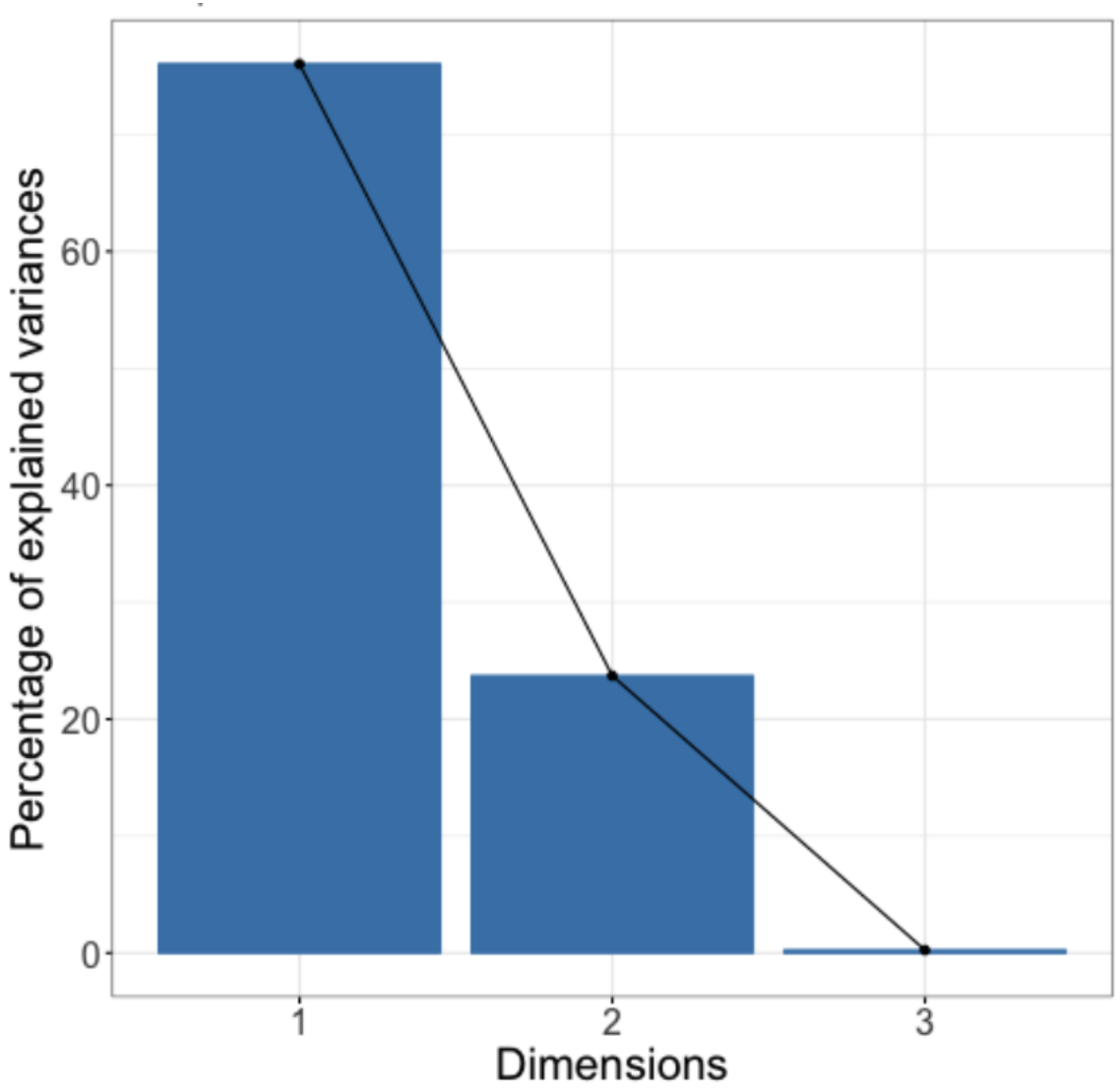
$$\frac{\text{inter-cluster variance}}{\text{total variance}}$$



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?

Implications of clustering

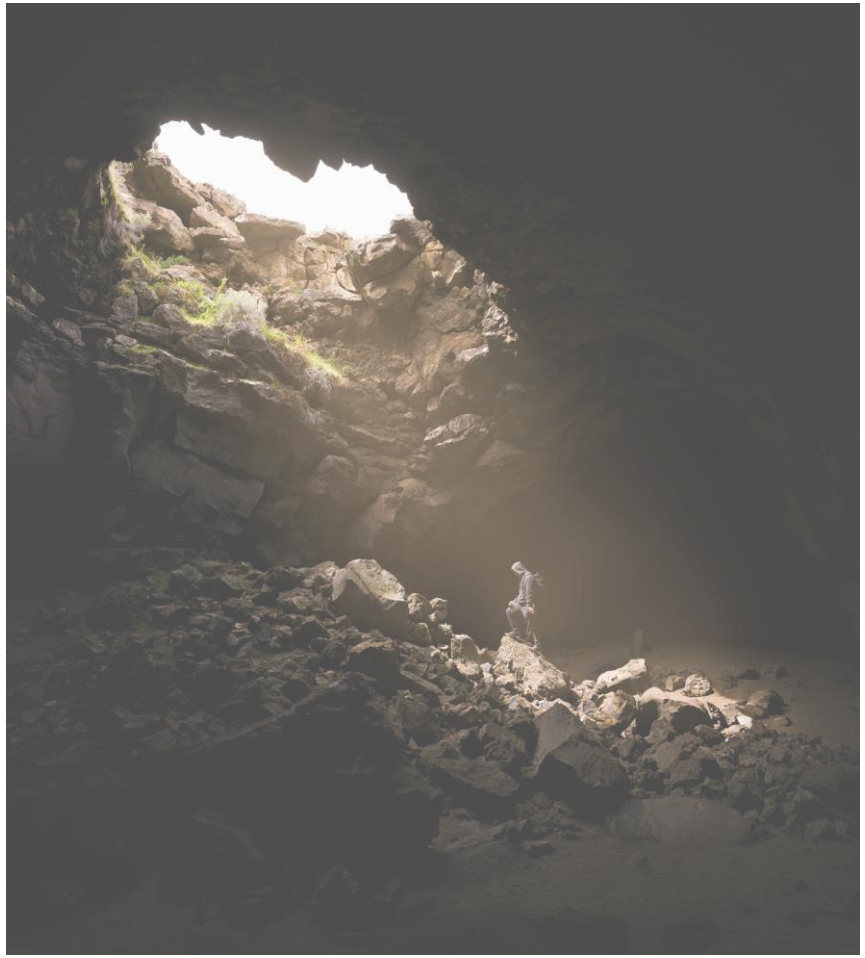
- When a problem is not well defined in advance, domain knowledge is key to extracting relevant conclusions.
- Clustering only highlights patterns.
- Other methods need to be used to confirm your hypotheses.

Common pitfalls with clustering

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

- Different data types need to be formatted correctly (i.e., mixing categorical data with numerical data may not be the best way to find similar points).

- Make sure you use the right clustering model for the data!



Recap: when should you use clustering?

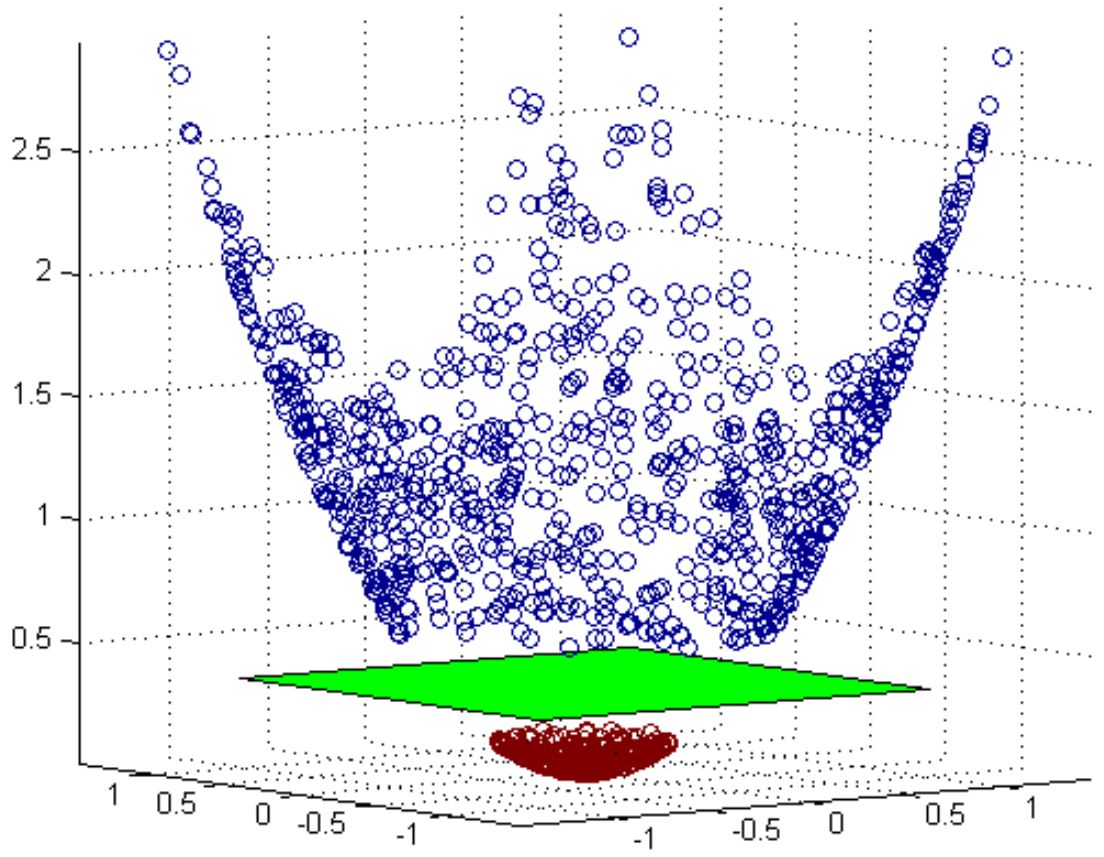
- Use clustering when:

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

Classification

Classification

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



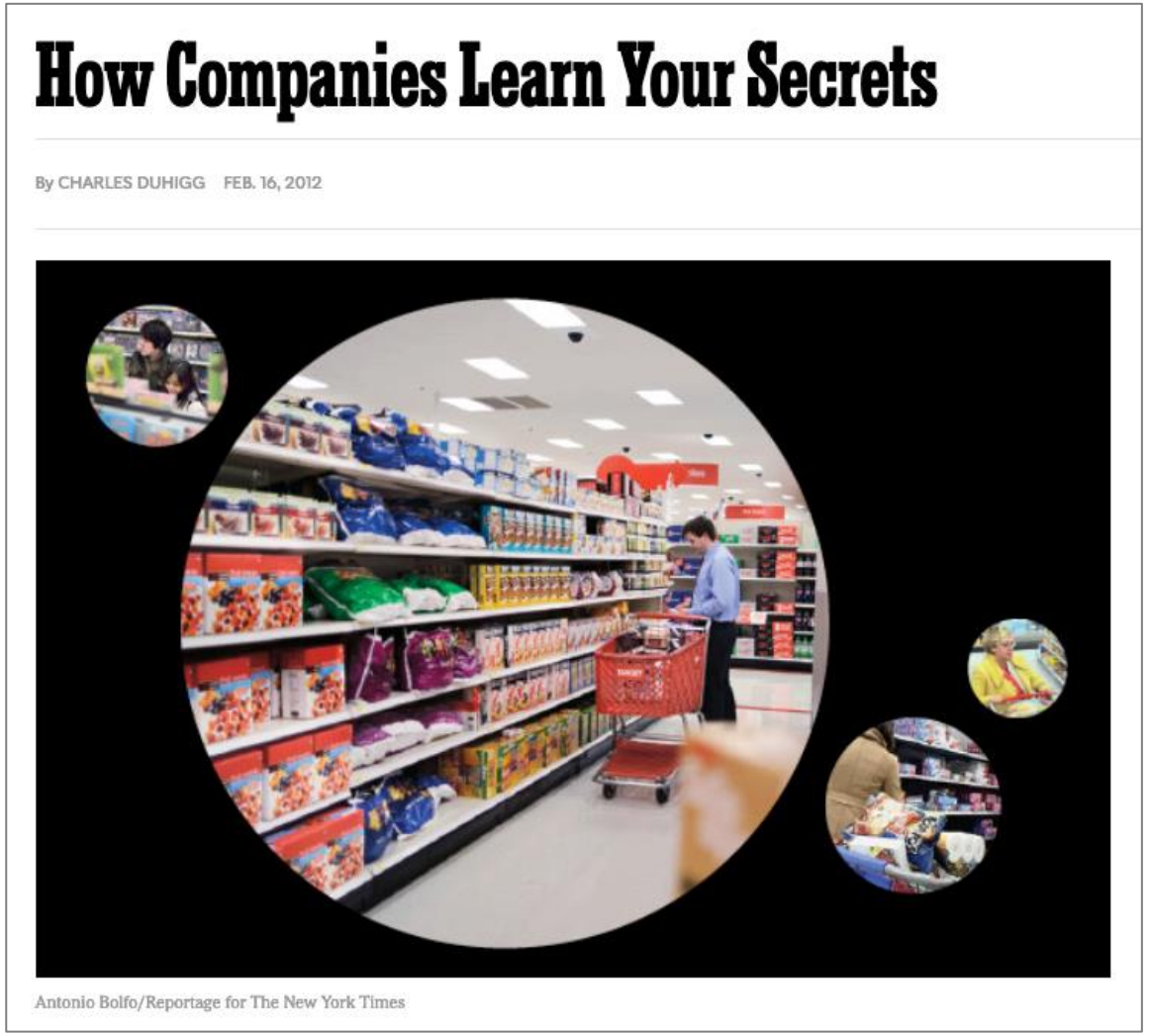
How can you use classification?

Classification answers the questions:

1. Which is the probability of an object / person being in a particular group?
2. What category is this person , object, or data in?

Example: predicting pregnancy

- In 2002, Target implemented data analytics to analyze buying patterns in customers.
- New parents often get bombarded with advertising offers, so Target wanted a way to anticipate who is expecting in order to get ahead of the competition.
- They were able to predict pregnancy of their customers based upon their purchases and sent out targeted coupons.



How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

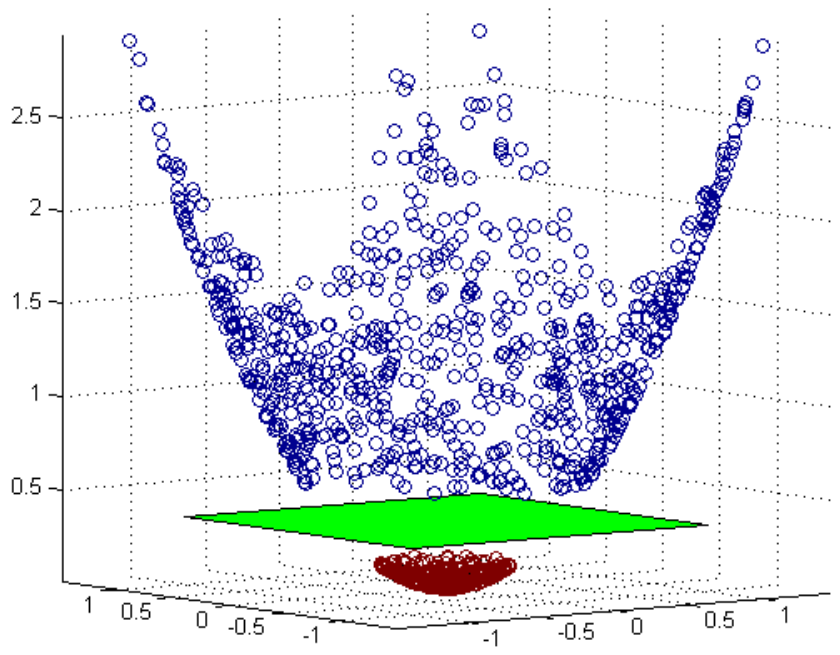
Antonio Bolfo/Reportage for The New York Times

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

Common classifiers

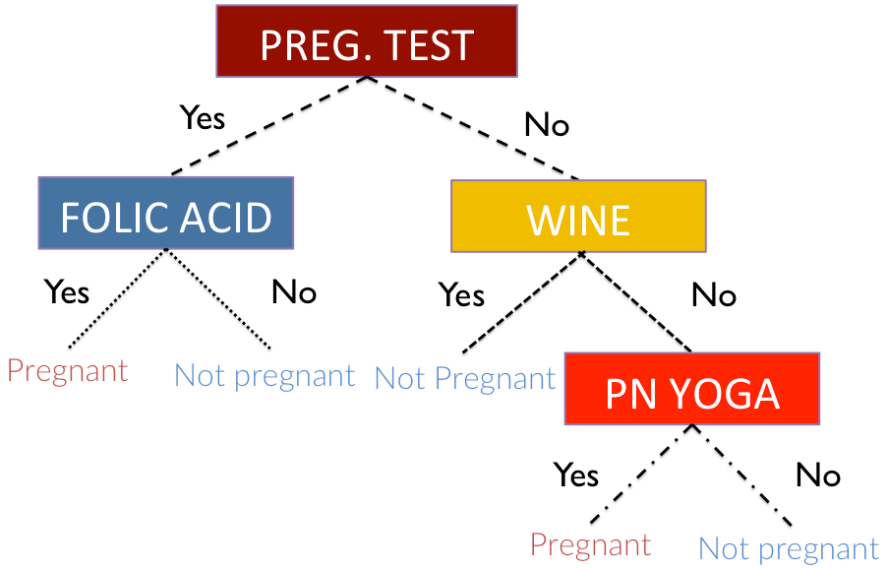
- Logistic regression – determines the probability of a data point to be part of a certain class or not

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

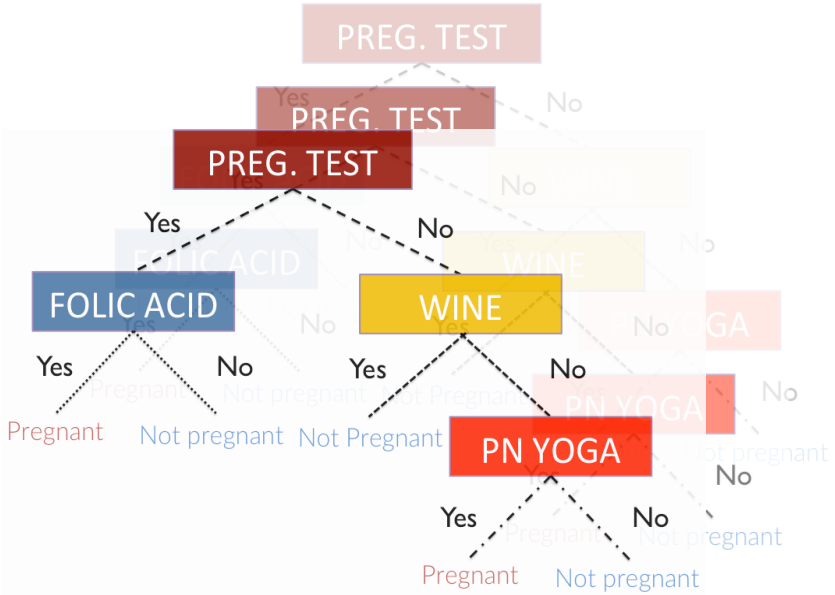


Common classifiers

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



- Random Forests – ensemble learning method that constructs multiple decision trees to receive a more accurate prediction



Evaluating accuracy of a model

- In order to determine the accuracy of the model, you need to split your data into a training set and a test set.

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

Confusion matrix

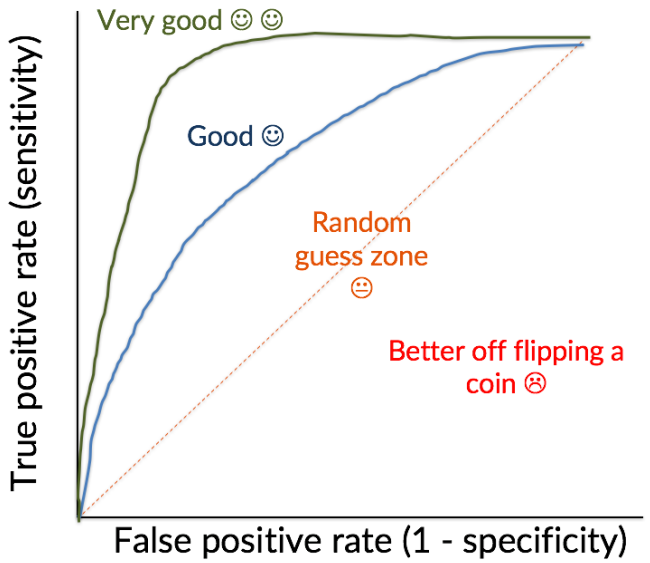
- Confusion Matrix describes how well a classification model performs on a set of test data

- Confusion Matrix is used to measure error

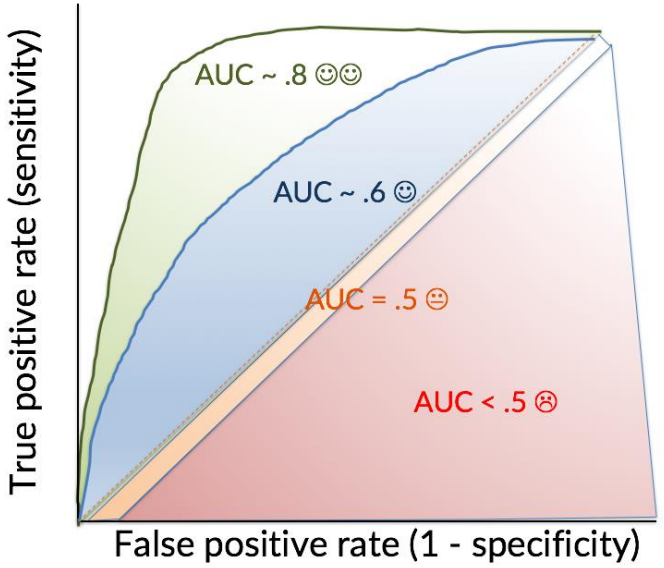
	Y1	Y2	Predicted totals
Predicted Y1	True positive (TP)	False positive (FP)	Total predicted positive
Predicted Y2	False negative (FN)	True negative (TN)	Total predicted negative
Actual totals	Total positives	Total negatives	Total

Accuracy, cont'd.

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



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



Questions managers should ask

1. How many classification techniques did you attempt?
2. Why did you select a specific algorithm?
3. Did you scale the data appropriately?
4. How did you split the test and training data?
5. What thresholds did you use for AUC and ROC?

Recap: when should you use classification?

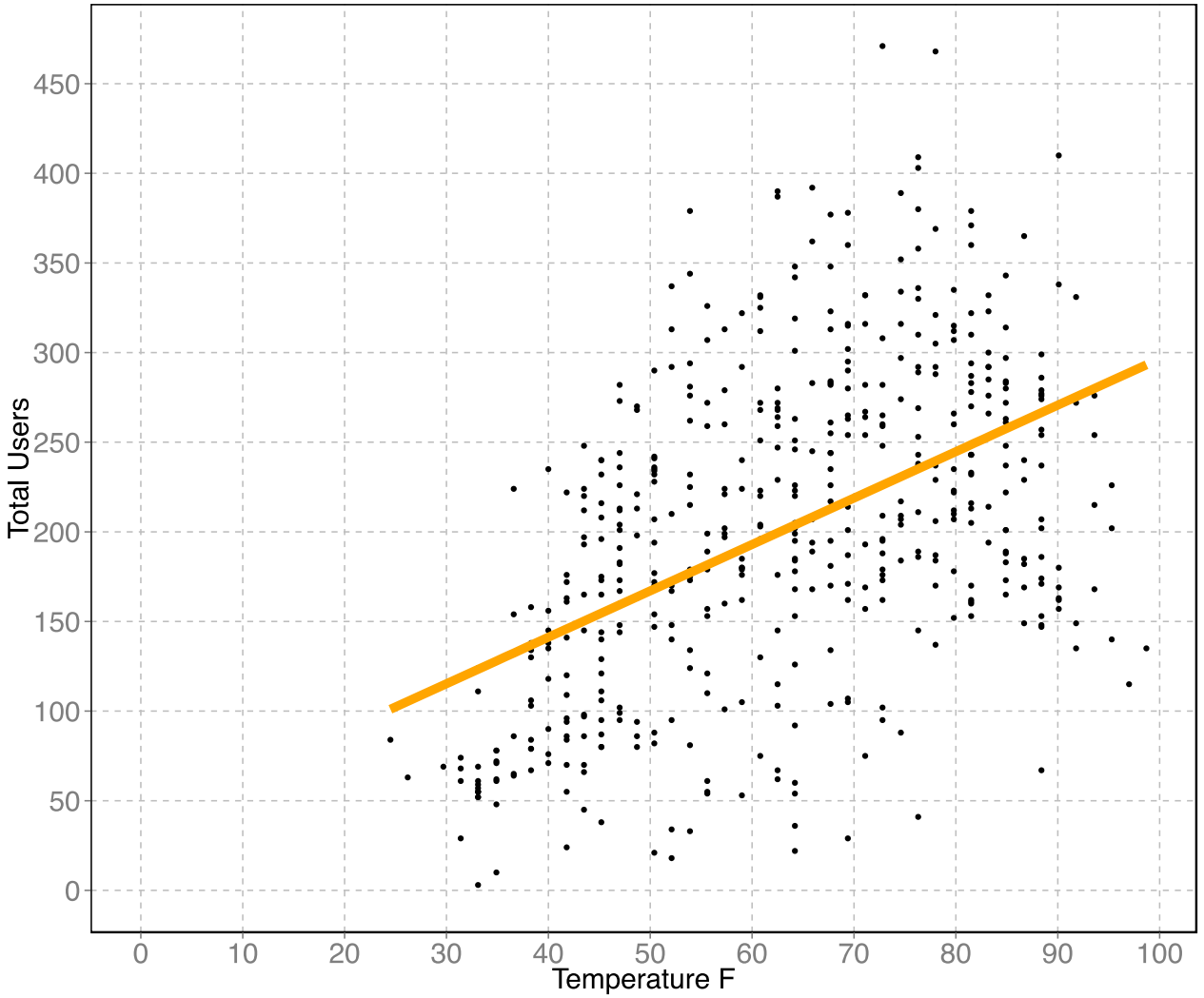
Use classification when:

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

Regression

Regression

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



How can you use regression?

- Regression answers the questions:
  1. Which factors matter most?
  2. Which can we ignore?
  3. How do those factors interact with each other?
  4. How certain are we about these factors?

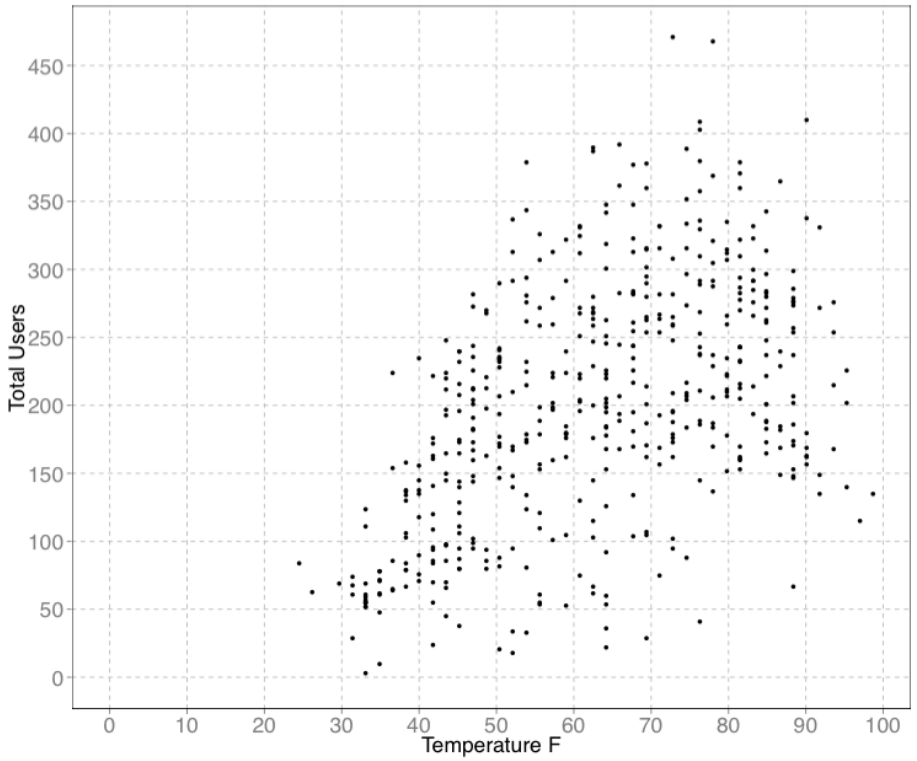
- Domain knowledge is key!
  - We can predict political instability in countries
  - We can predict how tourism season affects a country's economy

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.

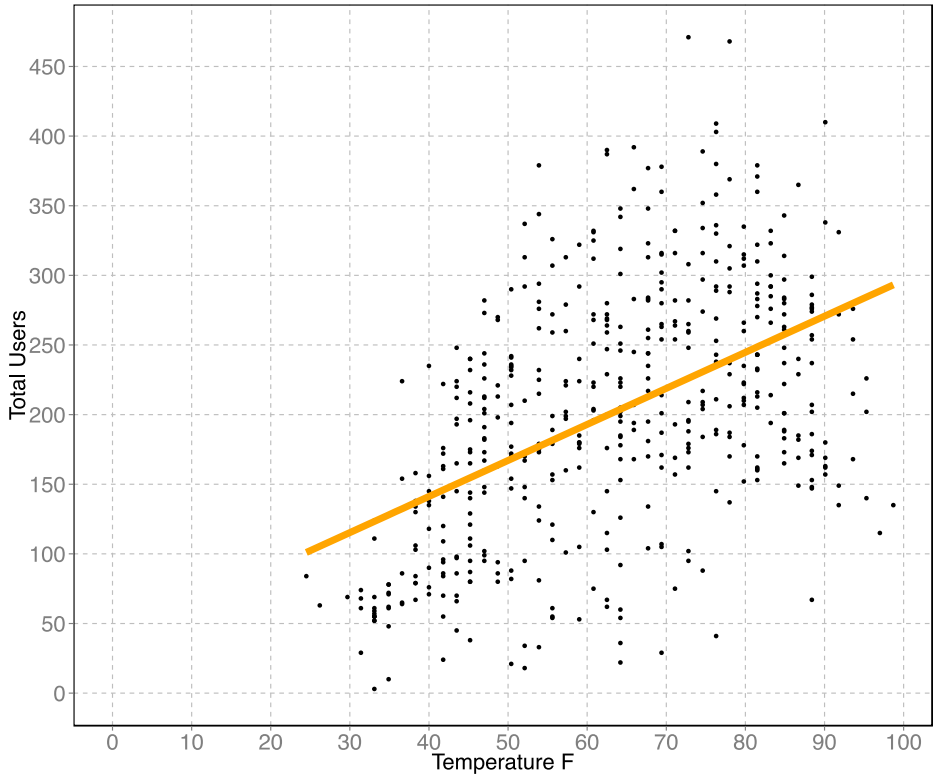
Simple linear regression

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

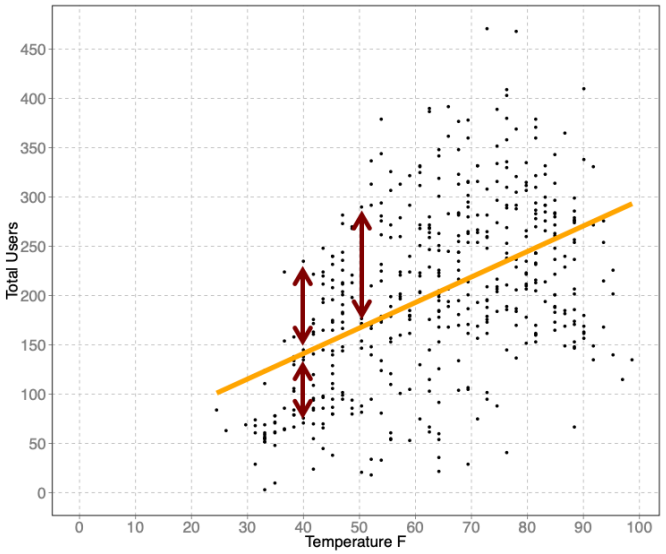


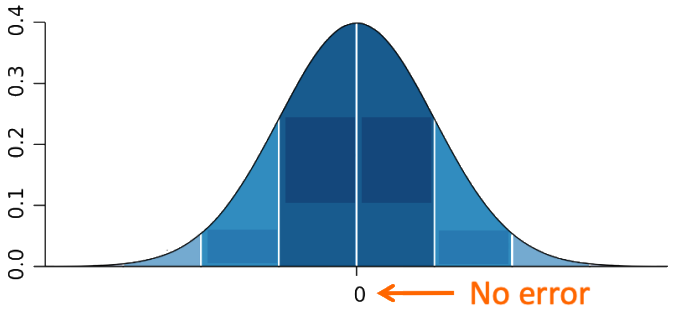
$y = mx + b$

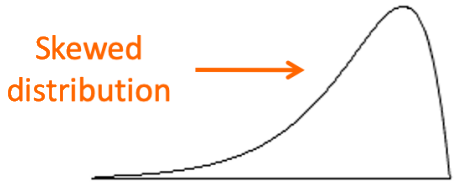
Number of bike users
=
2.6 * (Temperature)
+ 37.6

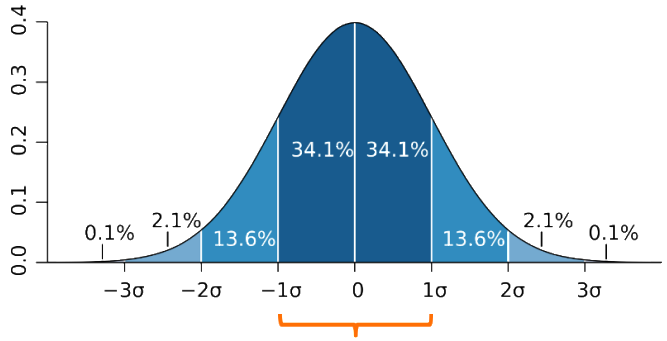


Measure error









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

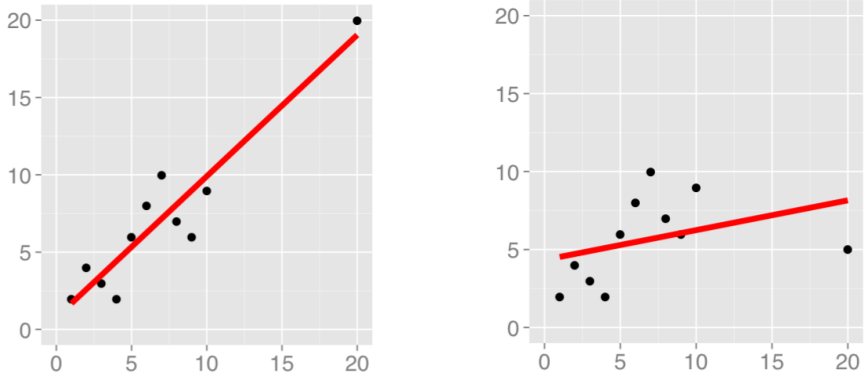
Variance. How widely dispersed is actual data from the expected data?

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

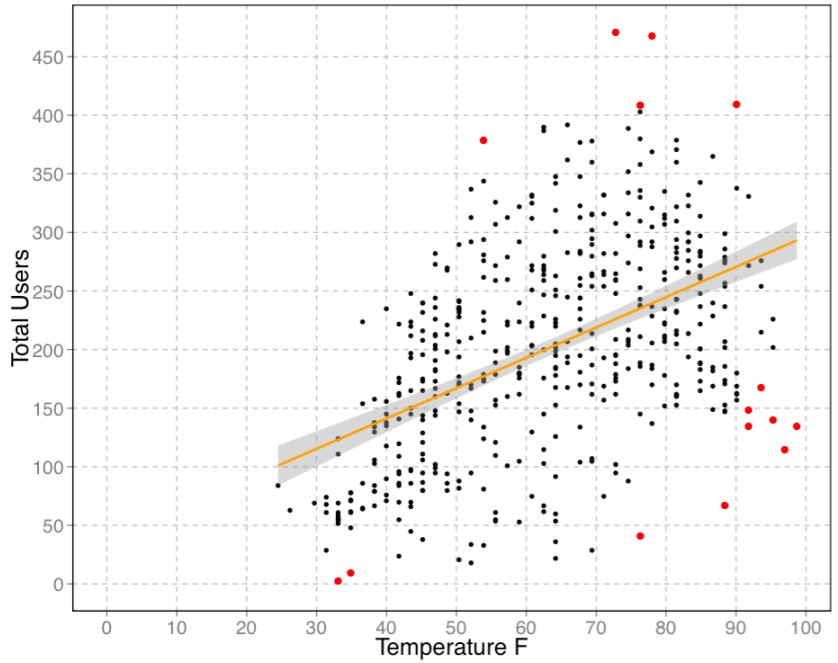
Standard deviation/Certainty. What proportion of data points fall within a given range? How likely is a value to be in that range?

Deal with outliers

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



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



Evaluate model

- 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 tell us what percent of the target's variation is explained by our model.
  - e.g., "about 40% of the variance in the number of bike users is explained by the temperature"
- Mean Absolute Error (MAE) tells us the average difference between the predicted value and the actual values

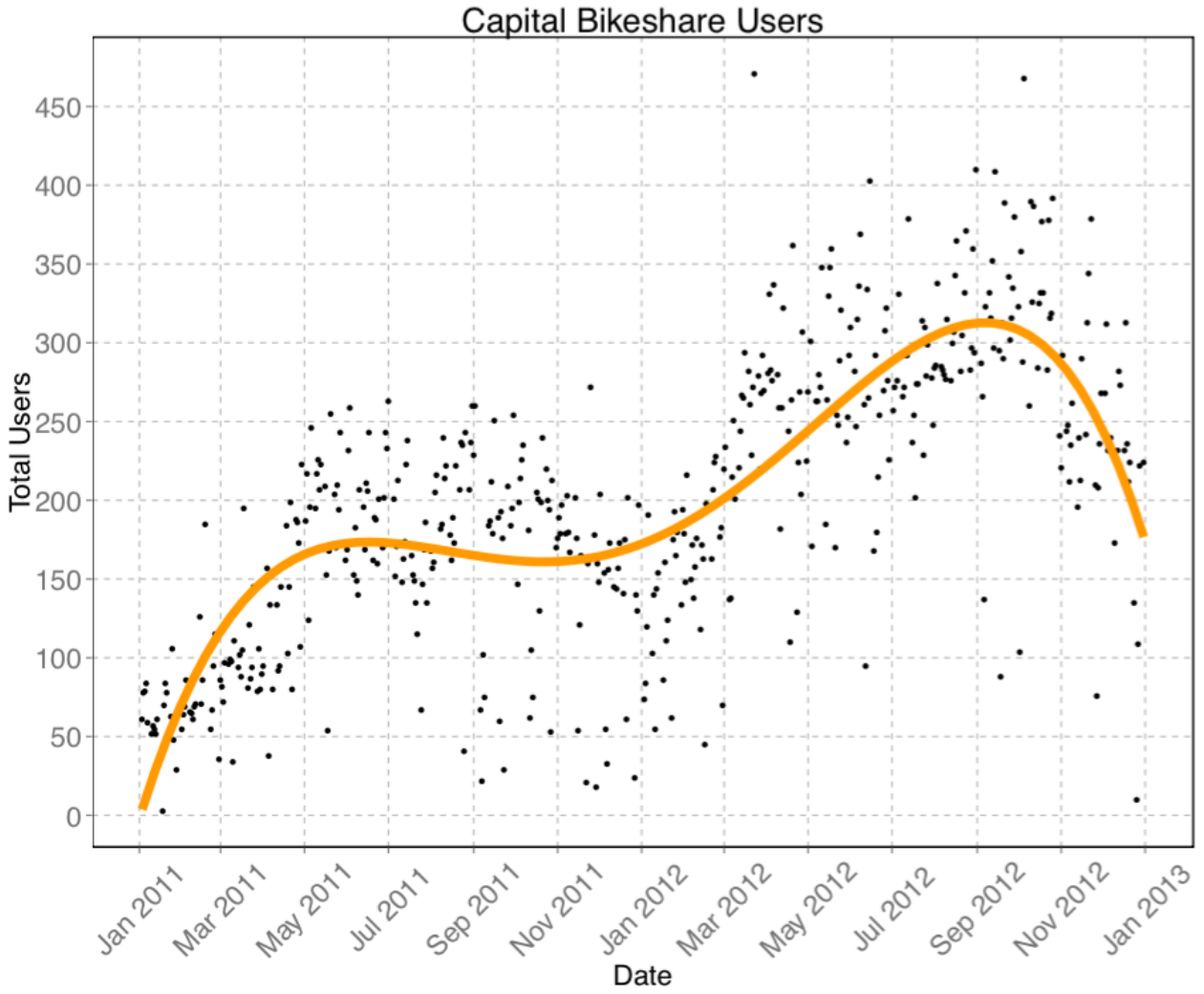
Multiple linear regression

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

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

Other types of regression

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



Questions managers should ask

1. How well do we understand the underlying data distribution?

2. Did you identify any outliers? Were they significant? Did you remove them?

3. Did you test the variables for multicollinearity so as not to double-count their effects?

4. What were the R^2 and MAE metrics?

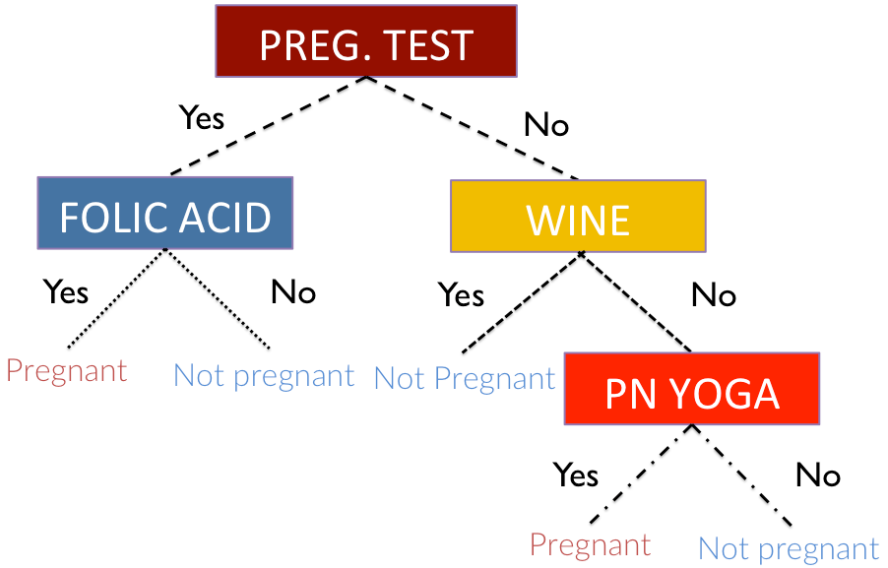
Recap: when should you use regression?

- Use regression when:

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

Poll question

Do you think decision trees are a type of classification?

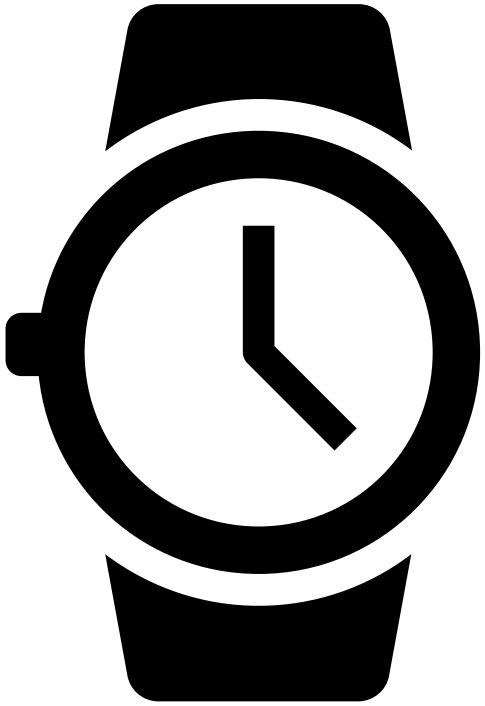


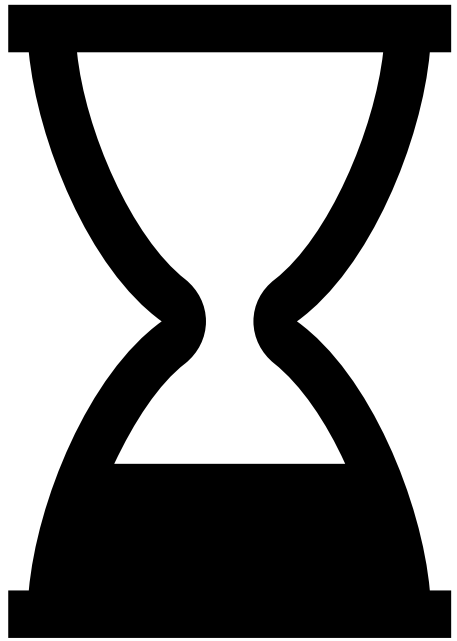
Poll question

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

How Machines Learn

Break





Agenda

Day 3

- Foundational data science methods
- Advanced data science methods

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

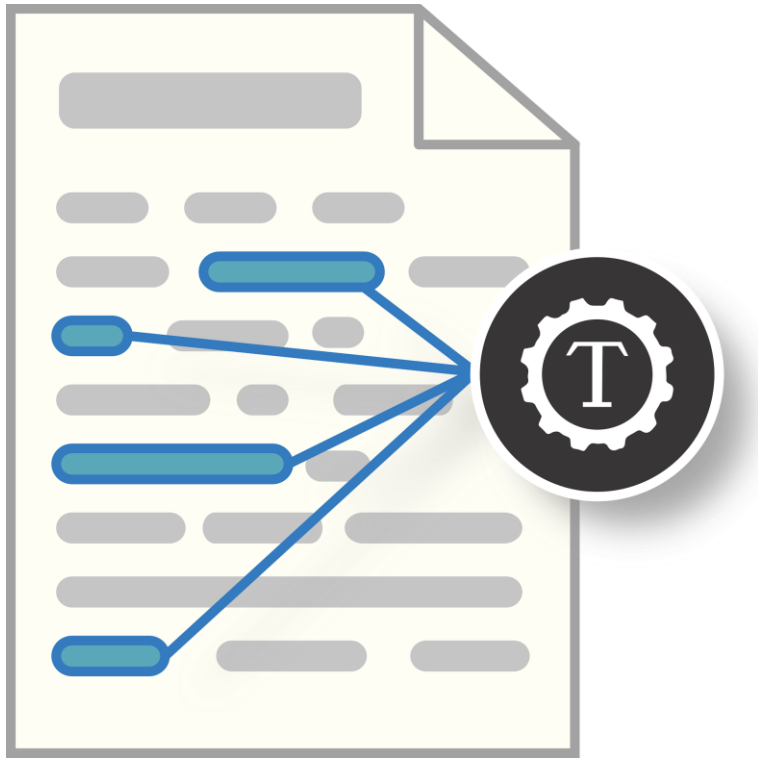
Text mining

What is text mining?

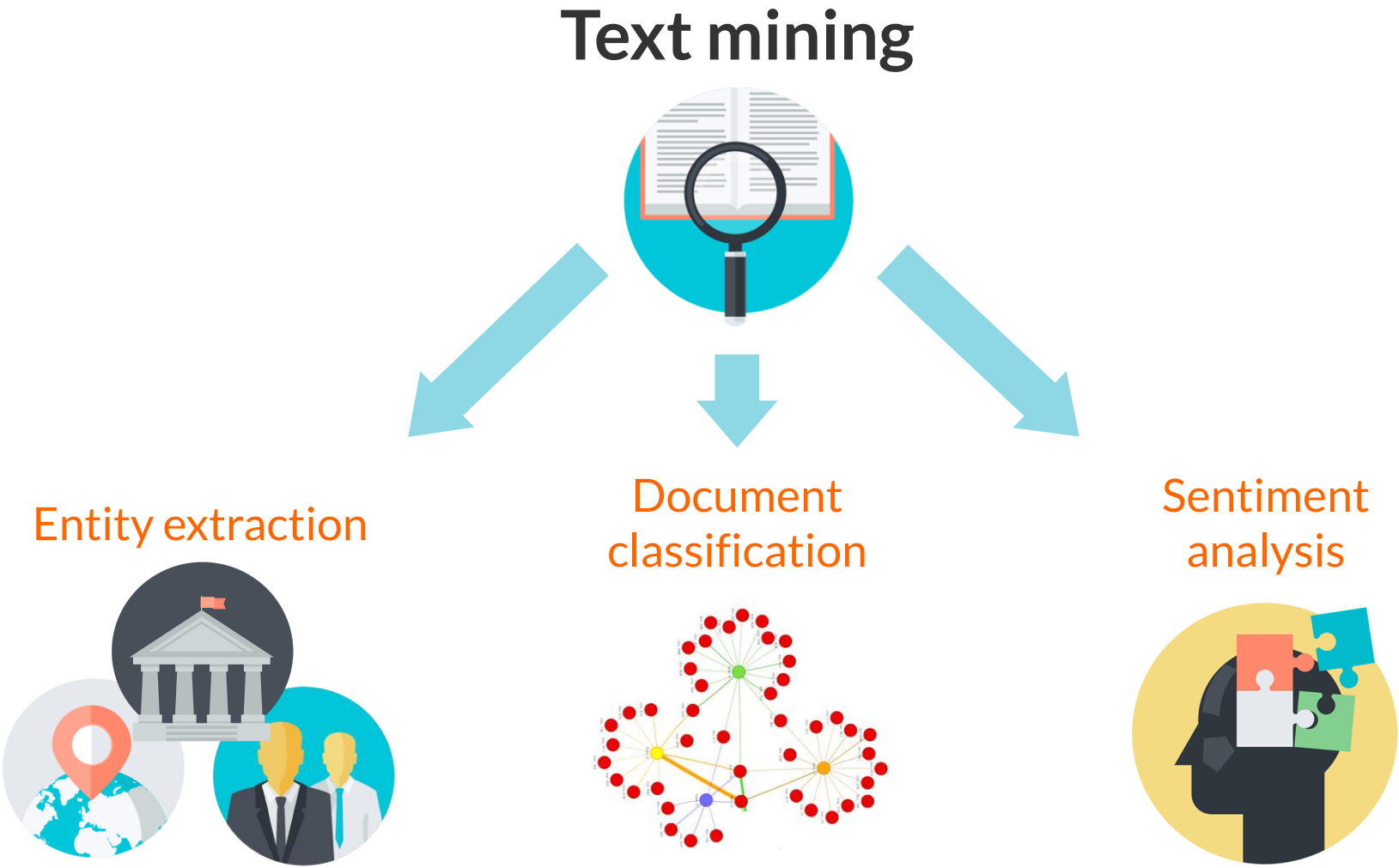
- Text mining is the process of getting insightful and valuable information out of text data.

- 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.

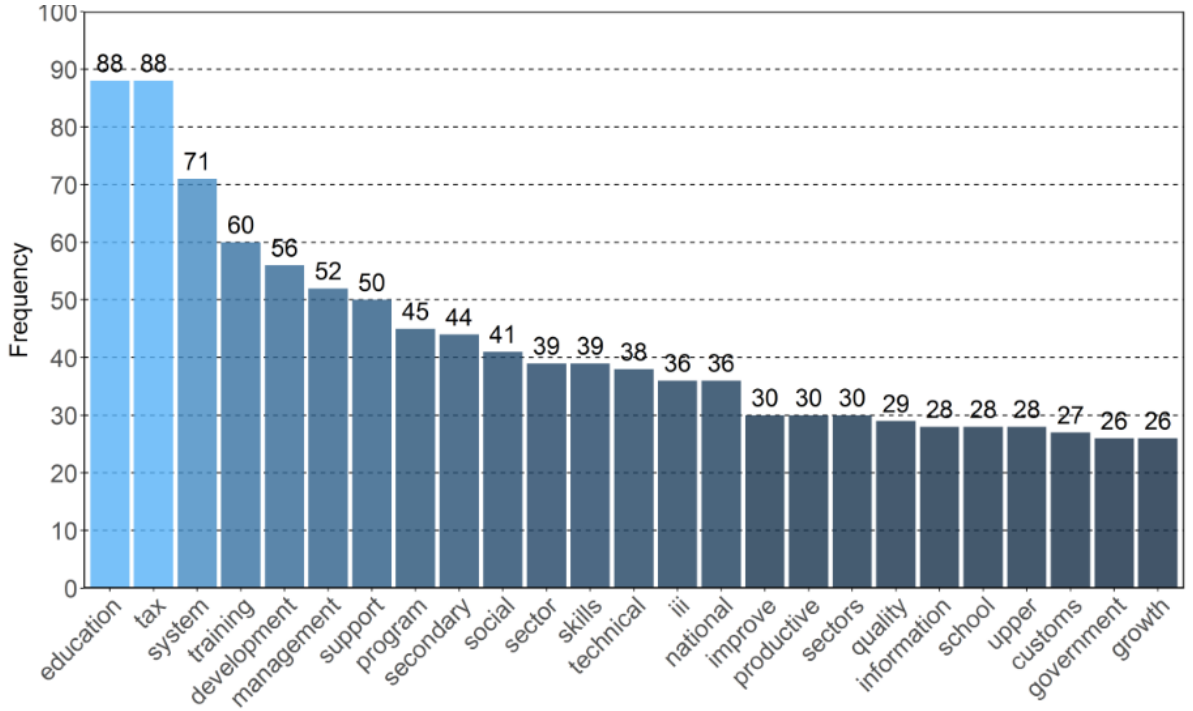


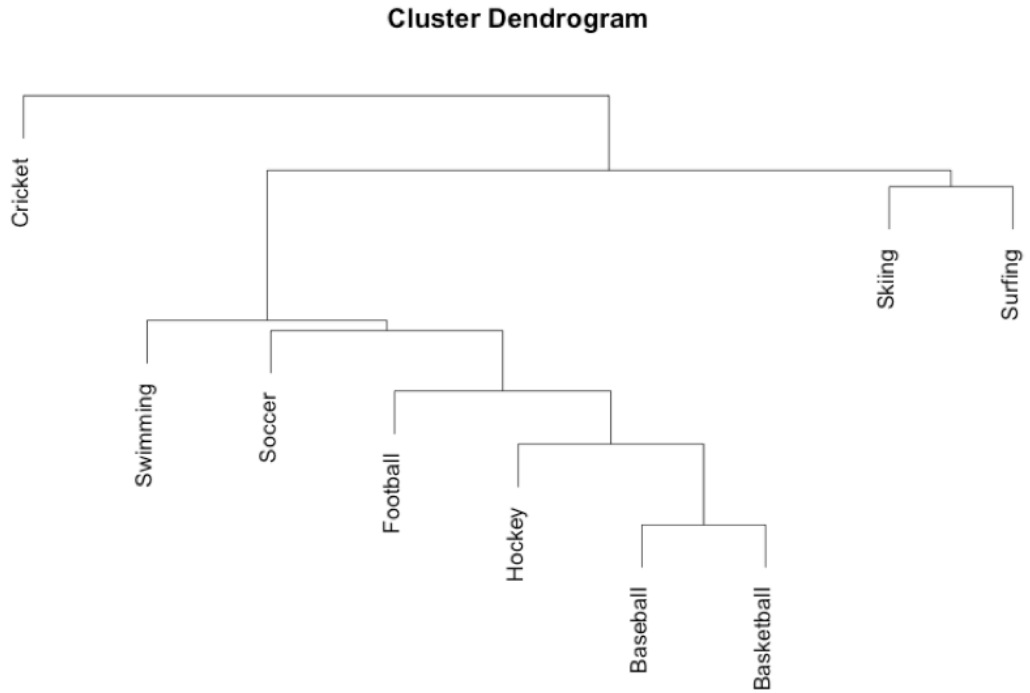
Text mining branches



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.

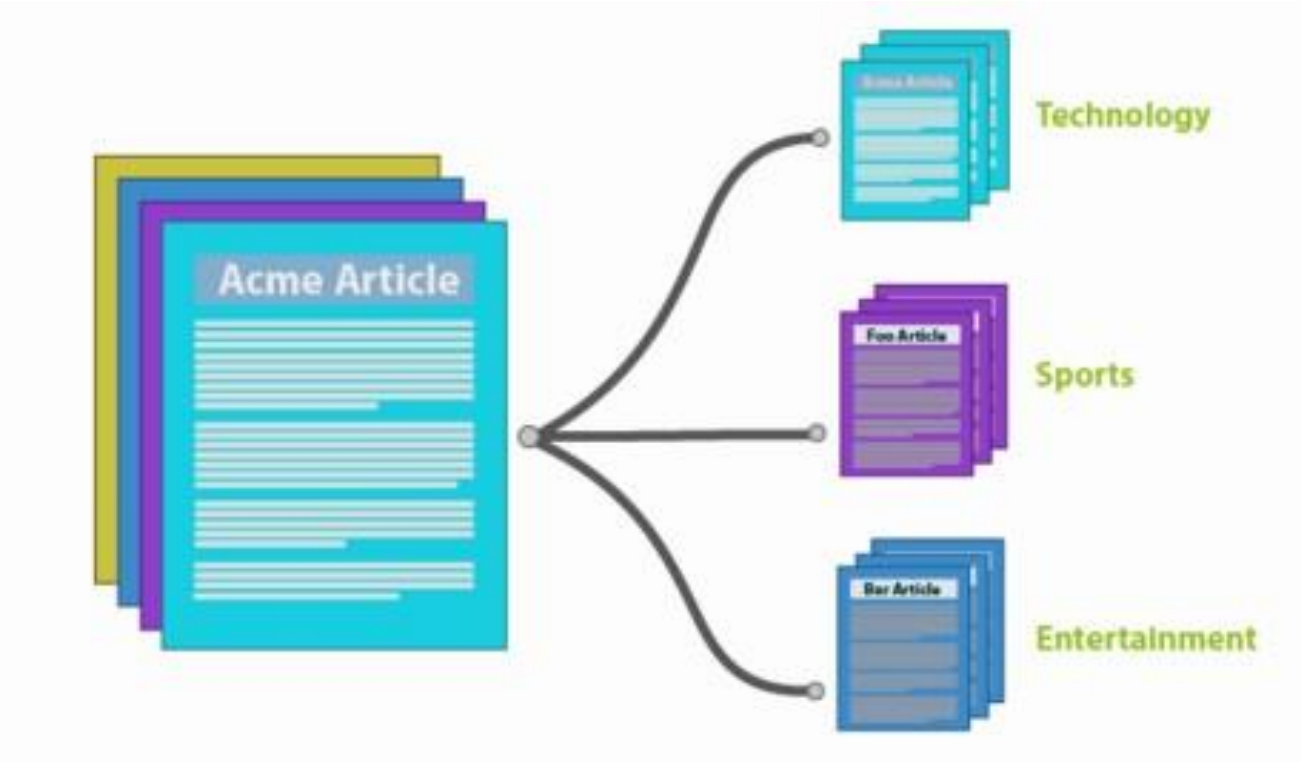




Document classification

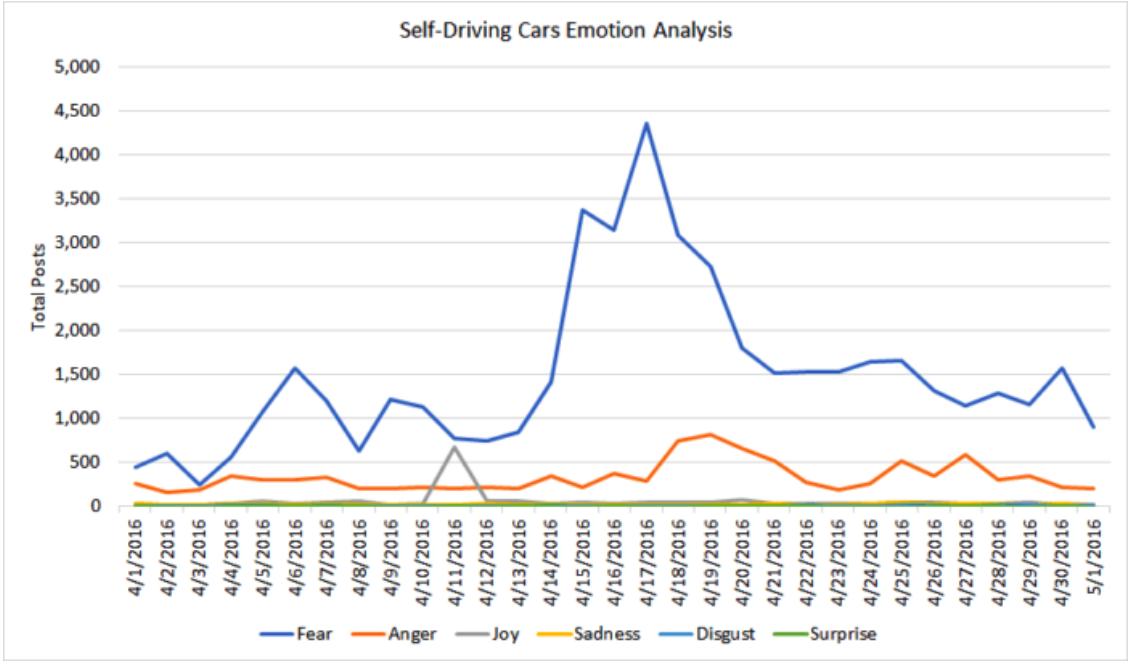
- Use document classification when you want to sort through documents and identify groups of similar articles.

- Based on similarity of topics / other metrics



Sentiment analysis

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



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

Text mining process

Scrape / collect

Clean & organize

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

Visualize

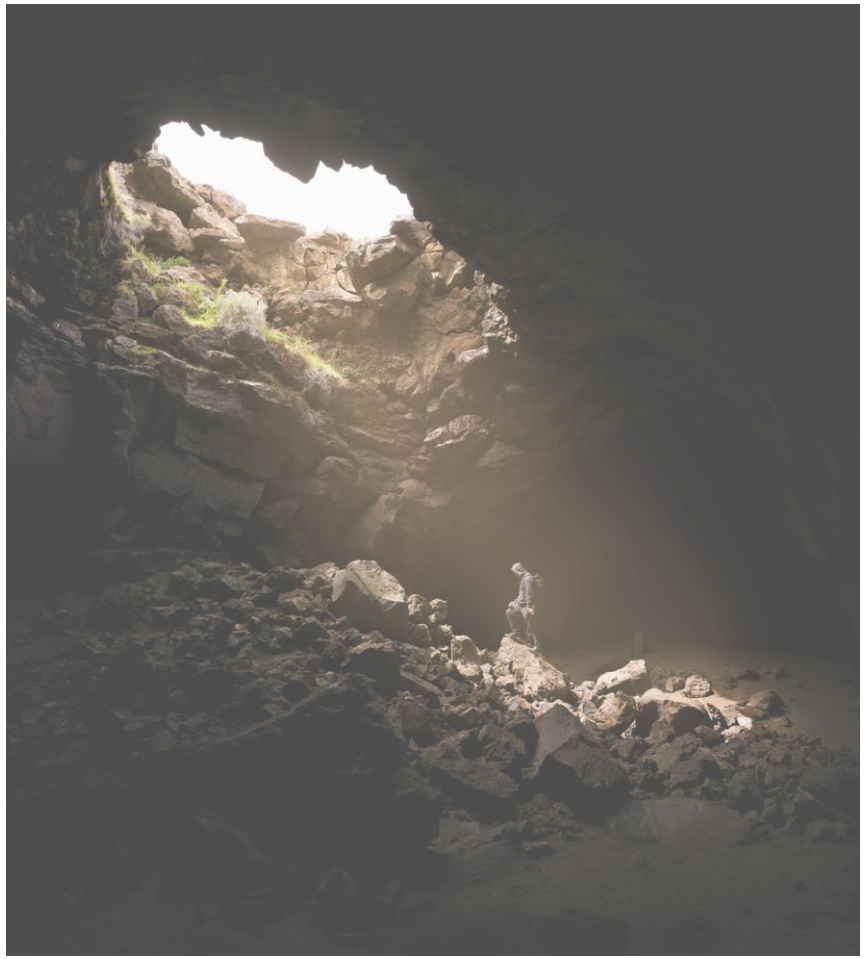
Analyze

Evaluating accuracy of our model

- This is a tricky subject!
- Text analysis and text mining rely on other methods that we've introduced in this class, such as clustering and classification. You'll need to use the evaluation methods for those particular models.
- In terms of sanity-checking the text mining process, look for unhelpful stop words (frequent words that don't provide additional information) and see if the topics generally make sense.

Common pitfalls with text mining

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



Questions managers should ask

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

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

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

4. How do the topics change over time?

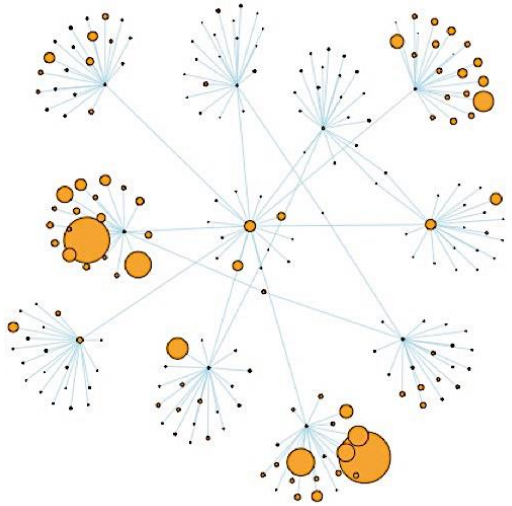
Graph analysis

Graph analysis

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

- 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?



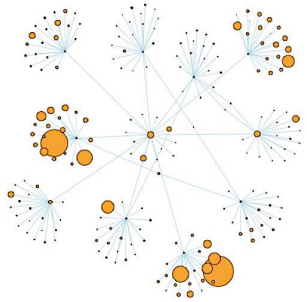
Example: IBM & a volcano

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

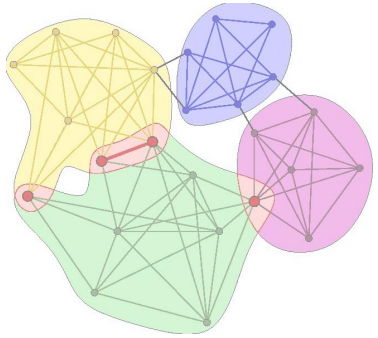
Source: Big Data Driven Supply Chain Management by Nada R. Sanders

Types of graph analysis

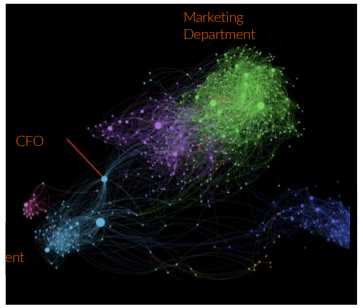
Graph analysis



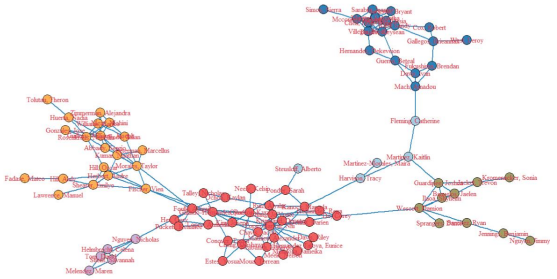
Community detection



Centrality metrics



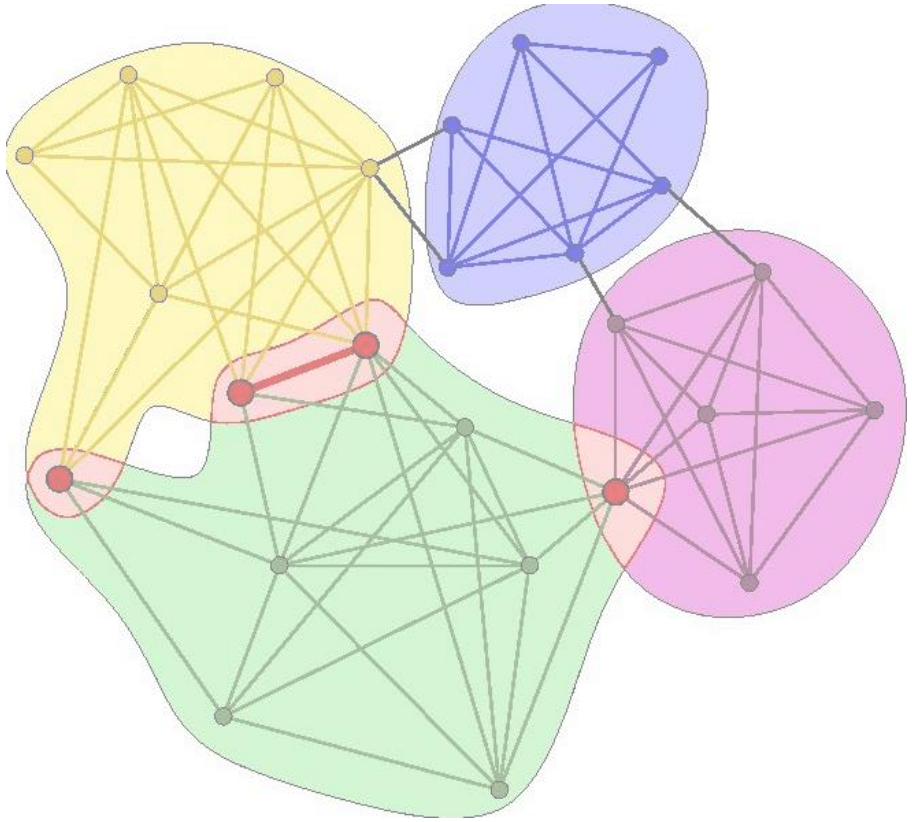
Social Media



Community detection

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

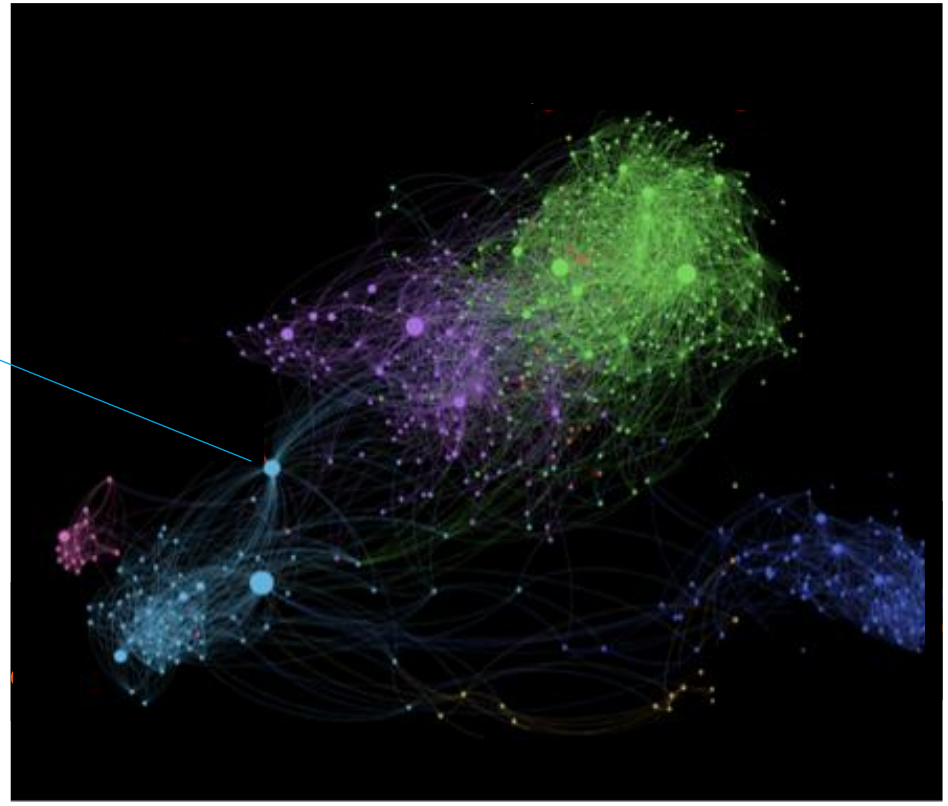
- Identifies groups of individuals / nodes that belong together; can detect latent connections and communities.



Centrality metrics

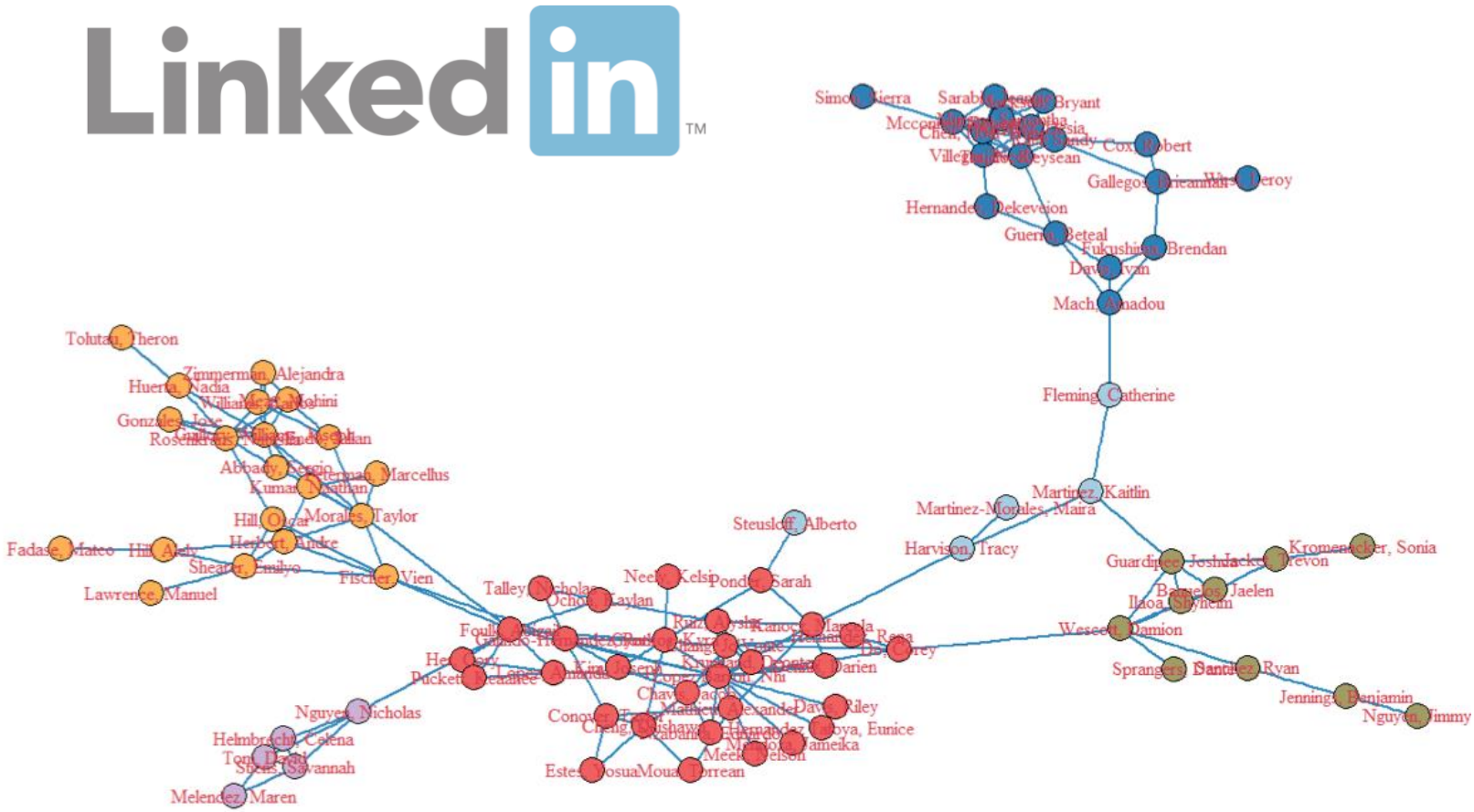
- Use centrality metrics when you want to look at an overview of a network and identify key nodes.
- Identifies the most important nodes, most central nodes, shortest paths, etc.

This email network shows how a company communicates.



Social media

- Use social media when you are using data from social media platforms.
- Identifies how an idea travels across social media platforms and how individuals are connected.



Ways to measure networks

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

Evaluating a Graph Analysis

- This is a tricky subject!
- Graph analysis relies on other methods that we've introduced in this class, such as clustering and classification. You'll need to use the evaluation methods for those particular models.
- In terms of sanity-checking the process, look at how the nodes are accounted for in each community and determine what threshold makes the most sense for your analysis.

Questions managers should ask

- What aspect of the relationship are you most interested in (i.e., who is the most connected, who has the strongest connections, who is most important)?
- Does the data you're using account for a large amount of a relationship? How much is in the numbers versus not collected?
- What metrics did you use to evaluate the proximity between nodes / communities?

Neural networks

Activity: field trip

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

Note: A clickable link is available on page 12 of the participant guide.

What are neural networks?

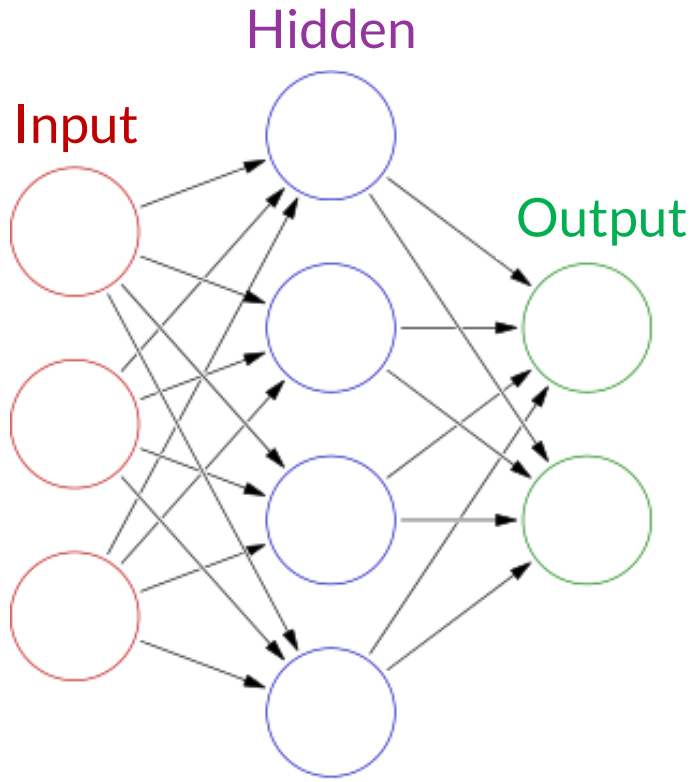
- A neural network is born ignorant and builds on itself to get smarter and smarter.
- It starts out with a guess, and then tries to make better guesses as it learns from its mistakes.
- Neural networks cover the same topics that we've reviewed previously. In theory, you can apply them to almost any method!
- You should typically only use neural networks when you have large volumes of text, image, or audio data.

Intuition: neural networks

- Artificial neural network is a graph of neurons similar to the human brain.

- A simple neural network 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.



What data do you need?

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

2. **Properly classified**: Labeled data are need for a deep-learning solution. If a labeled dataset is not available, someone needs to label the raw data.

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

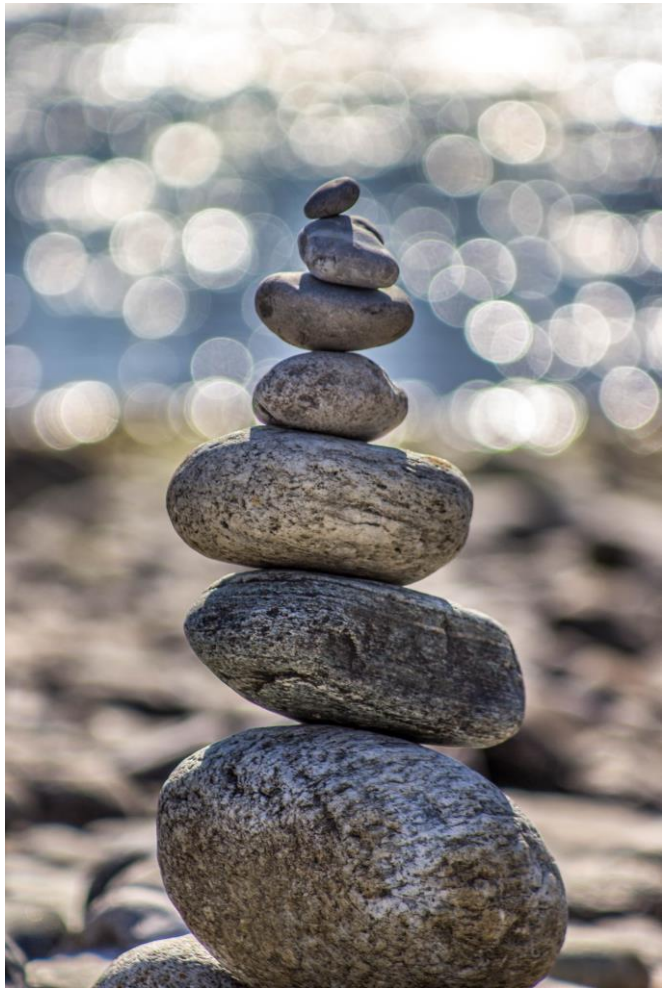
4. **Minimum data requirement**: this may vary with the complexity of the problem, but 100,000 instances in total across all categories is a good place to start

5. **Pre-trained models**: *these models were trained on large datasets that reflect similar questions that we are interested in solving.*

Neural networks: pros and cons

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

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



Poll questions

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

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

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

Key points

- Don’t accept an analysis at face value – you need to ask the right questions!
- Most data analyses incorporate multiple methods in order to determine which one is the most accurate.
- Remember! The two big components that drive the decision for which method to use are: the question you’re asking, and the data you have.

DATA SOCIETY®

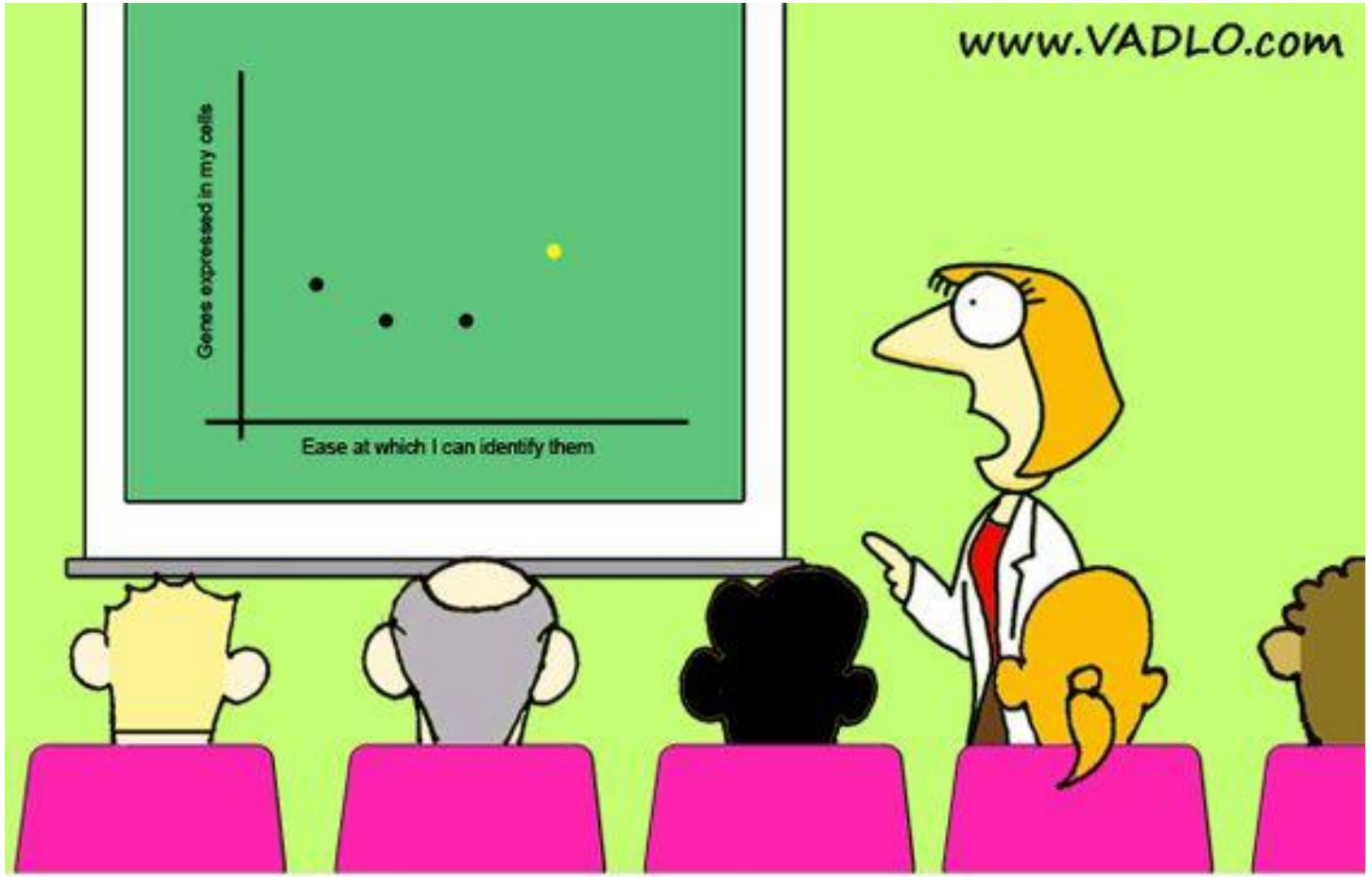
Questions?

End of Day 3

DATA SOCIETY®

Day 4

Welcome back!



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

Agenda

Day 4

- Data visualization
- Data storytelling

- Why visualize?
- What tools can be used for visualization?
- How do I select which graphs and charts to use to present results?
- What design principles apply to data visualization?

Which is more effective?

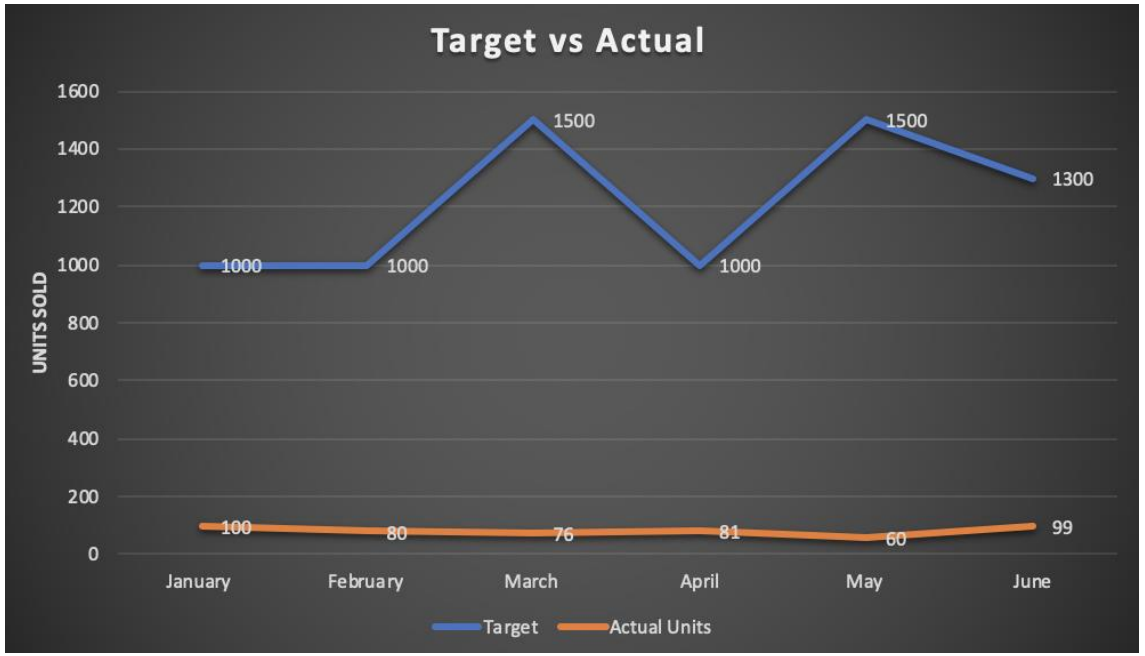
This is a picture of a puppy. This puppy is a golden retriever. The golden retriever puppy is sitting on the green grass next to a green ball. The ball is a lighter hue of green than the grass. Laying next to the green ball are two yellow tubes. The puppy, the ball, the grass, and the tubes are in front of a metal fence.

https://www.tapclicks.com/resources/blog/blog-what-is-data-visualization/

Which is more effective?

- Imagine that you are the CEO of a cabinet company that installs beautiful cabinets and doors in people's houses
- 5 months ago you launched a new product, customized “Rosemary-Gold” cabinets, which had 20% higher production costs than the other cabinets
- A newly joined analyst comes to you in a meeting and says, “Sir, we need to disable the production of Rosemary-Gold as I feel its not profitable.”

Do you agree with the analyst? What if he presents this instead?

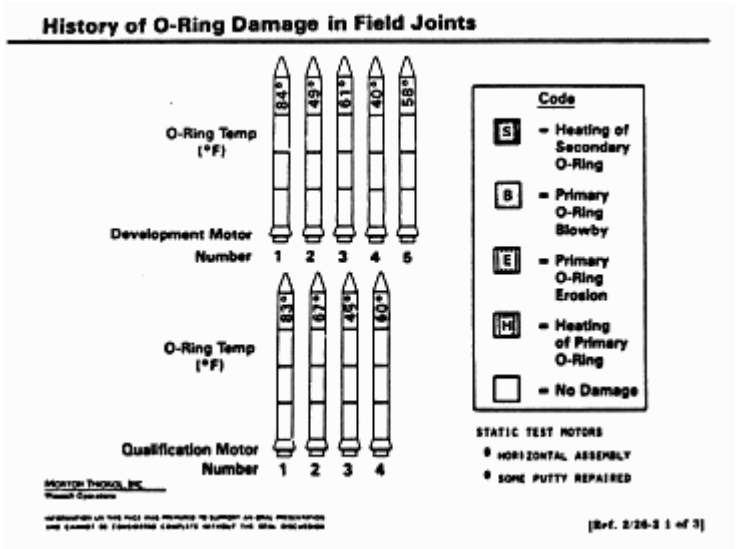


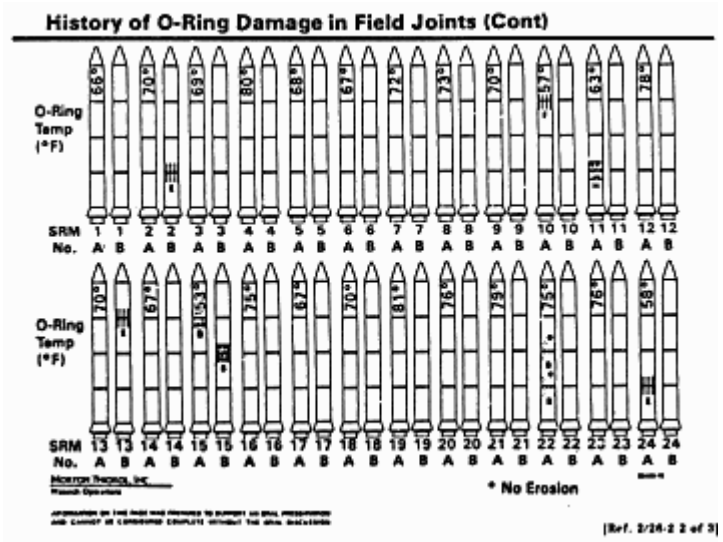
What is data visualization?

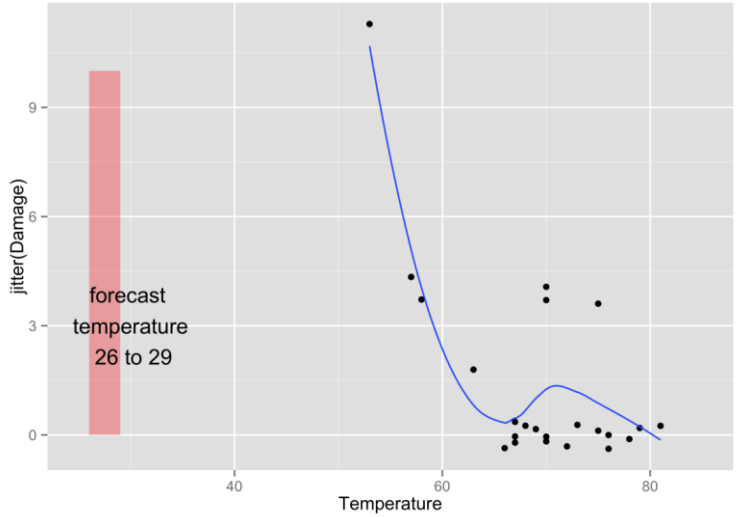
- Data visualization is any attempt to make data more easily digestible by rendering it in a visual context (e.g., charting, graphing, etc.).
- We use data visualization to transform raw data into something compelling.
- Using visualizations incorrectly can cause you to lose your audience, lose the value in your data, and ultimately lead to poor decision making.

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







Poll: data viz importance

Is data visualization an important part of your job?

What types of data visualization does your organization produce?

- Histogram, Box Plot, Pie Chart, Line Chart, Heat Map, Scatterplot

How to visualize data?

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. Use a visual that conveys the information in the best and simplest form for your audience. (How)

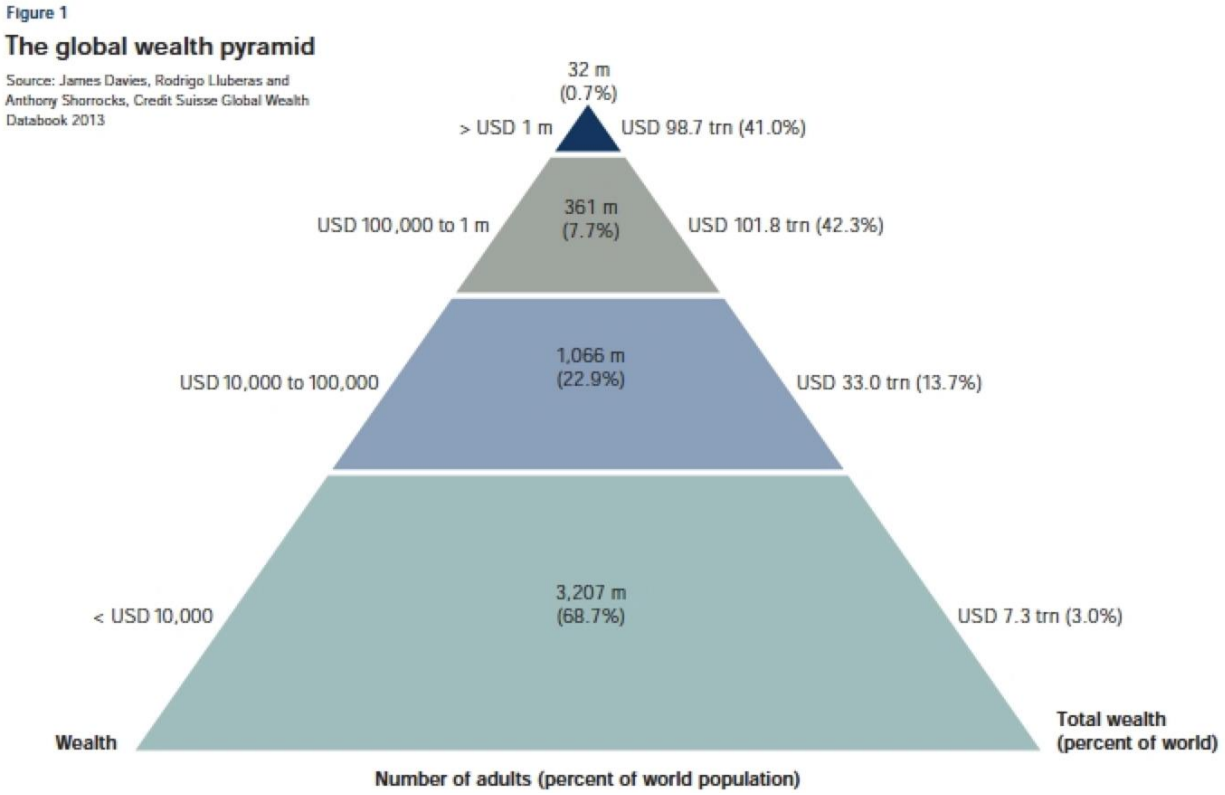
Who

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

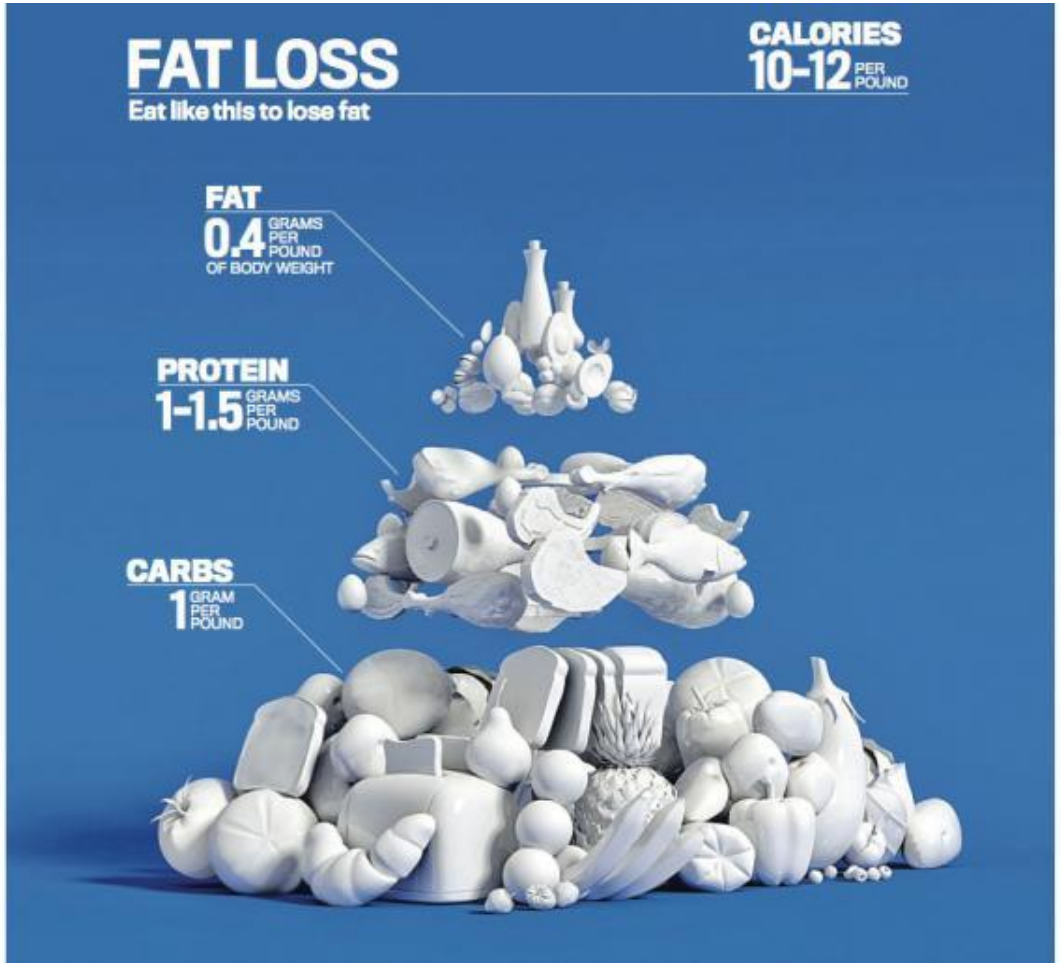
- Consider how the visualization will be used:
  - *Is it for executives to use to make decisions?*
  - *Is it to inform the public?*

Which do you like better?

Example 1



Example 2

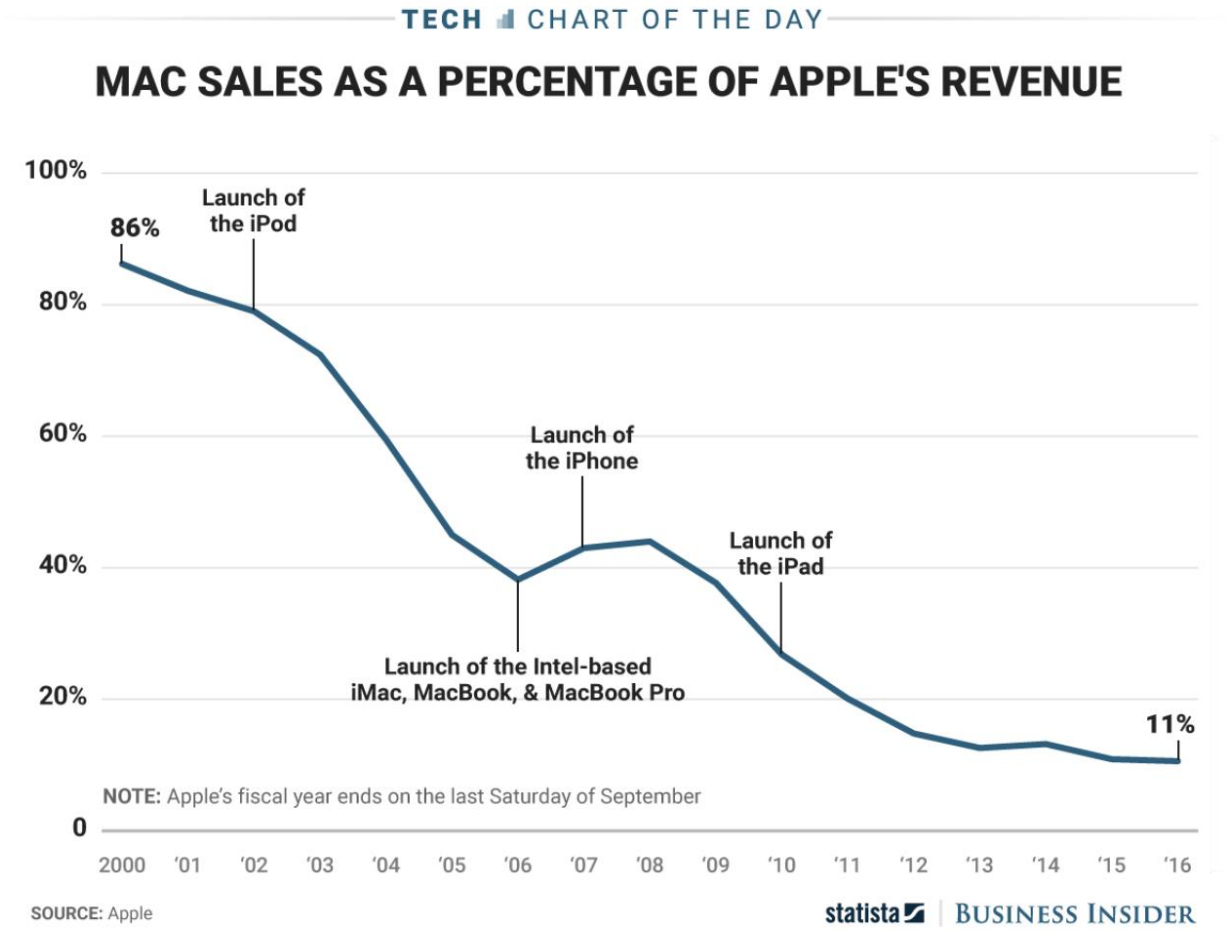


https://www.oreilly.com/library/view/learning-highcharts-4/9781783287451/ch08s04.html

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

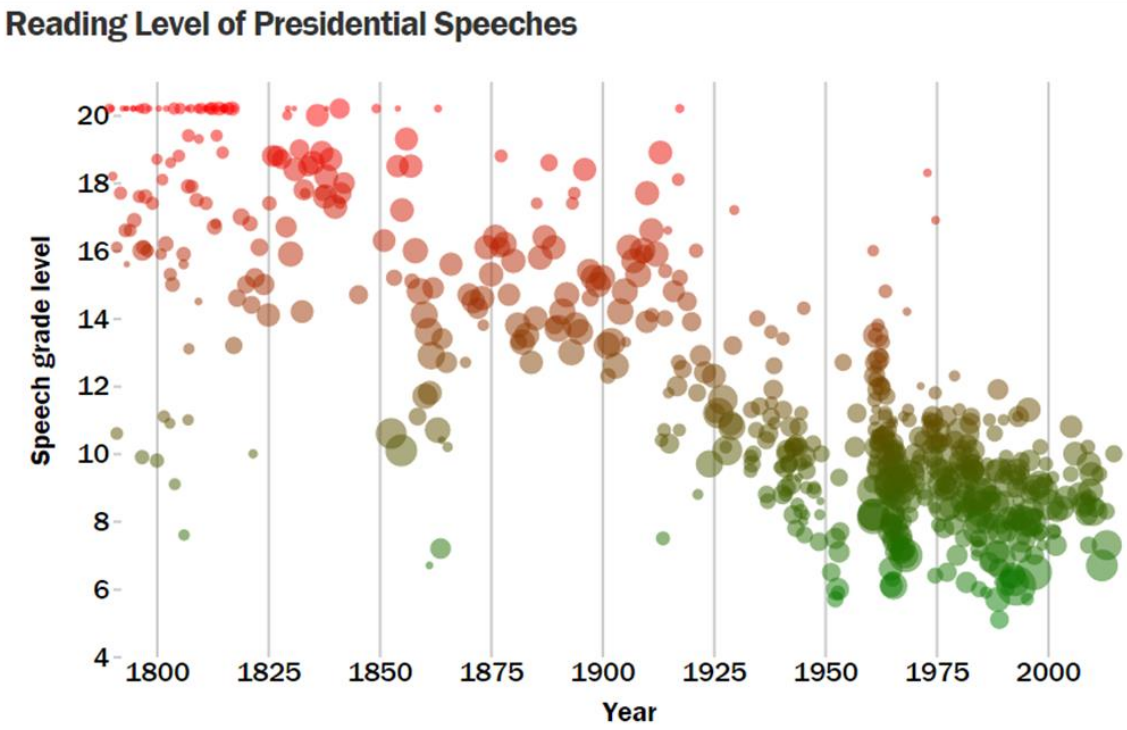
Which do you like better?

Example 1



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

Example 2



http://dadaviz.com/i/1276

What

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

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

How

- Once we know who the target audience is and what we want to communicate to them, we need to determine how we will communicate it.

- If the message is for:
  - **exploratory analysis**, you might provide a complete dataset with interactive elements to help the audience find interesting stories
  - **explanatory analysis**, you might use standard charts and graphs to accompany text or graphs with interactive layers that let users focus on their own areas of interest

Types of visuals

Choosing the type of visual

- Defining the who and what is often easier than defining the how.

- *What is the right visual for my situation?*
  - It is always the same answer: whatever will be easiest for your audience to read.

- But, the type of visual you use depends primarily on two things:
  - the data you want to communicate
  - what you want to convey about that data

Text and tables

- Don’t overcomplicate!
- Simple text works well when there is just a number or two to share.
- Tables are great when communicating to a mixed audience who will look to a particular row of interest, or to show different units of measure.

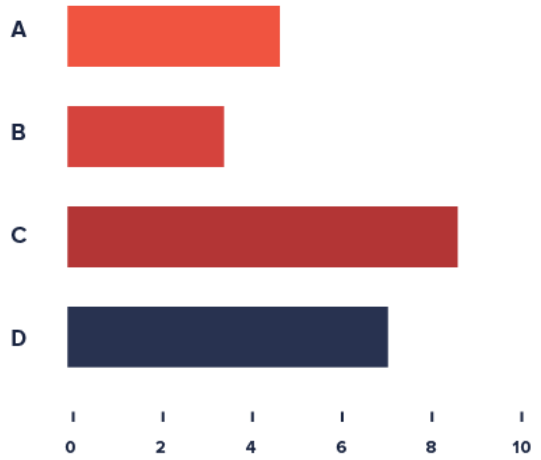
Common messages

- Often, we aim to show the following with our visualizations:
  - Comparisons - evaluate and compare values between two or more data points
  - Relationships - shows the relationship, correlation, or connection of two or more variables and their properties
  - Composition - how individual parts make up the whole
  - Distributions - combines the comparison and composition

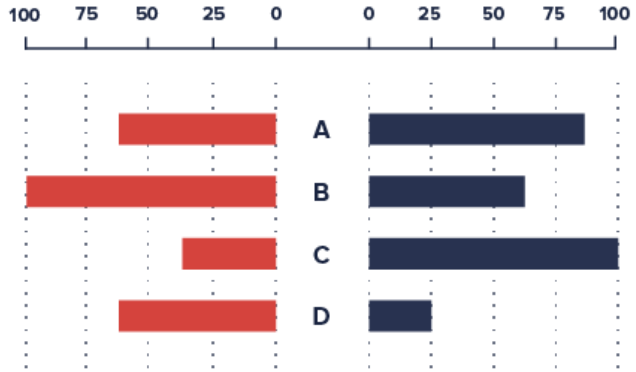
Comparisons

- Can detail over time ranging from many periods to a few periods
- Can group into one or two variables per item and go further by providing a range of categories
- Consider bar charts, pie charts, heat maps, treemaps, & more

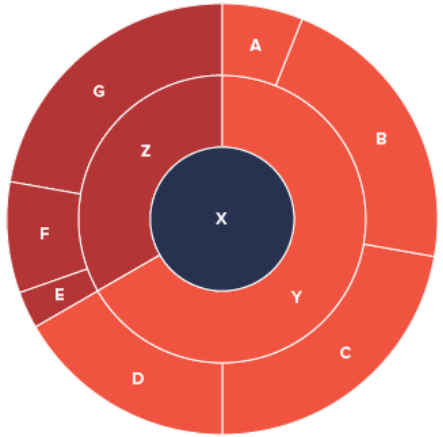
https://datavizproject.com



Bar Chart



Butterfly Chart

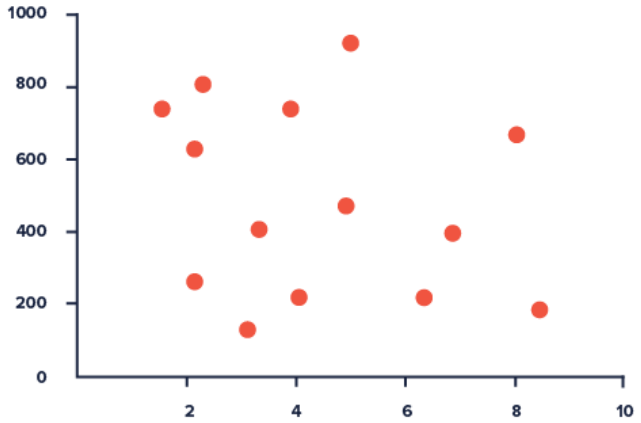


Multi-level Pie Chart

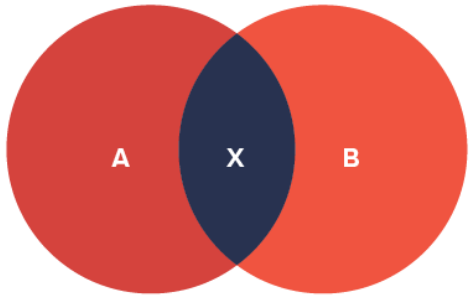
Relationships

- Can choose number of variables
- Consider scatter plots, Venn diagrams, bubble charts, & more

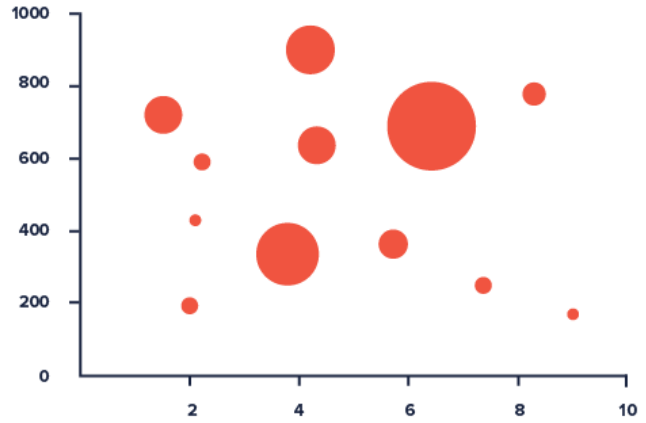
https://datavizproject.com



Scatter Plot



Venn Diagram

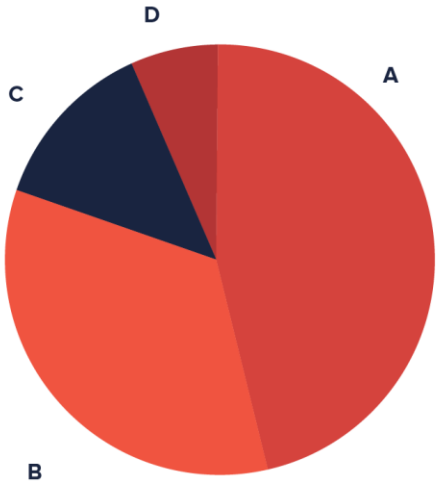


Bubble Chart

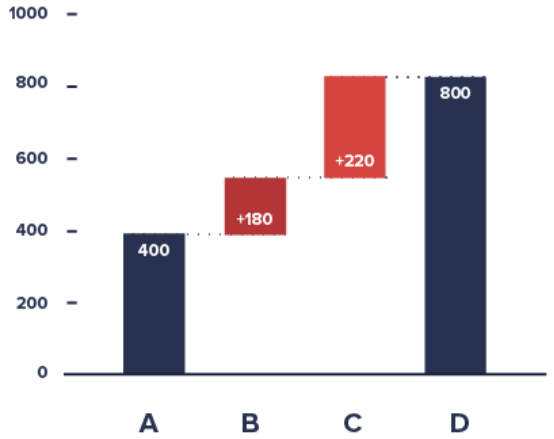
Composition

- Can be grouped by change over time or by static groupings
- Can visualize over different periods with relative and absolute differences
- Consider pie charts, waterfall charts, stacked column charts, & more

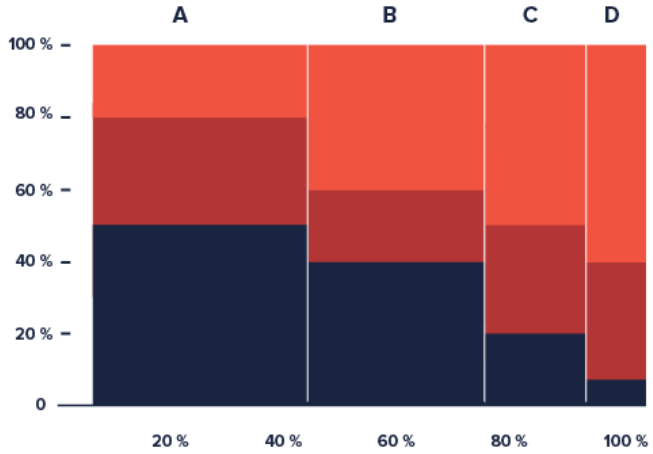
https://datavizproject.com



Pie Chart



Waterfall Chart



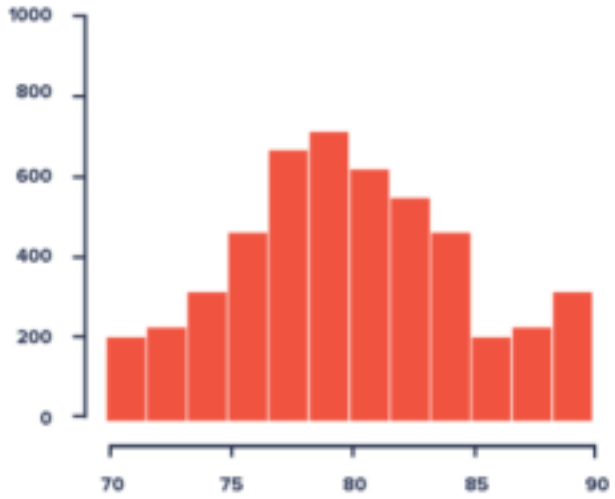
Stacked Column Chart

Distributions

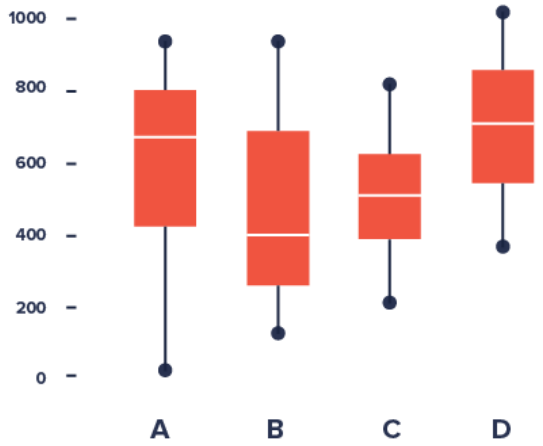
- Can choose one to three variables
- Can choose a few data points or many
- Consider word clouds, histograms, box plots, & more

https://datavizproject.com

Word Cloud



Histogram



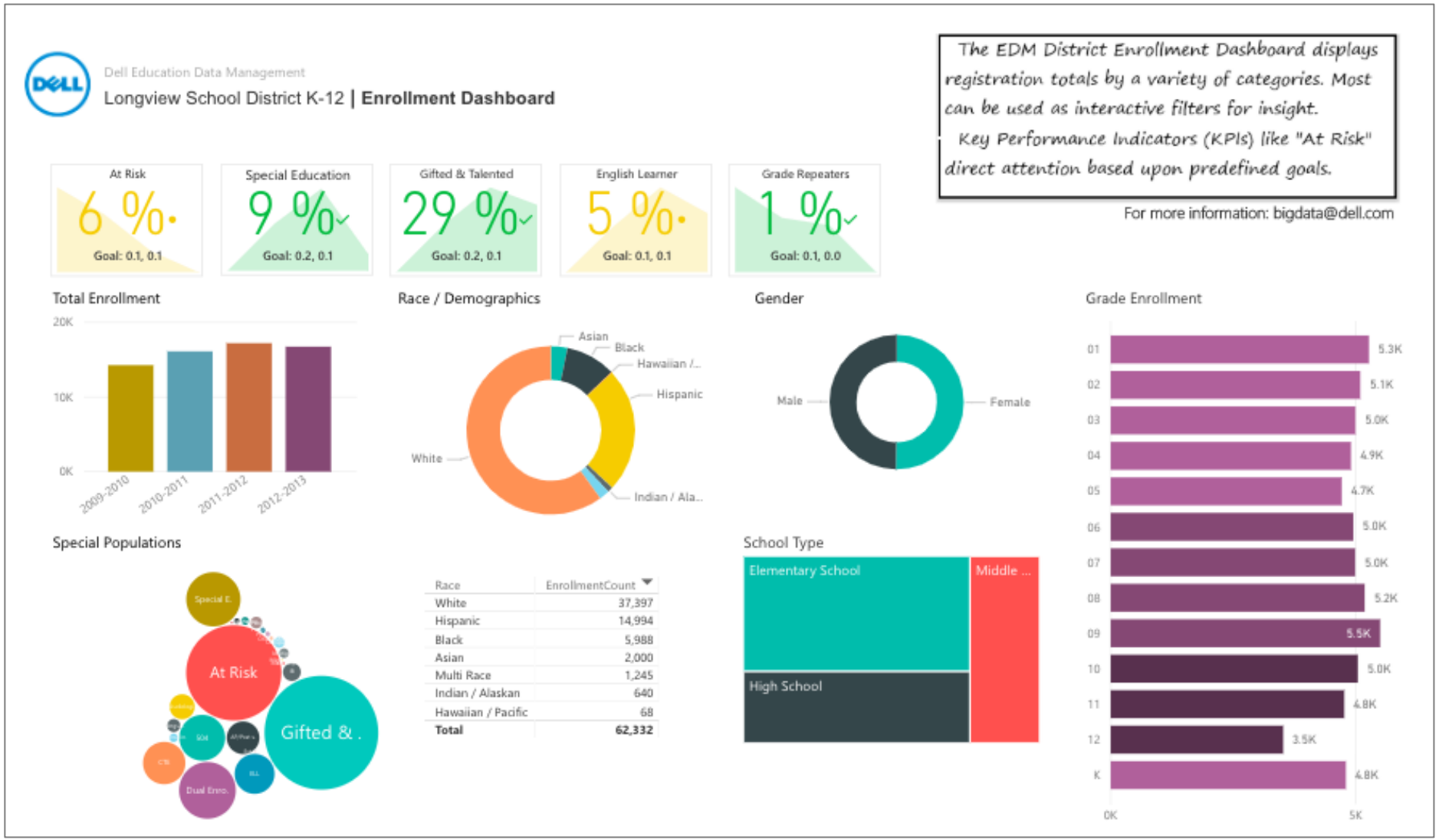
Box plot

What if it's more complicated?

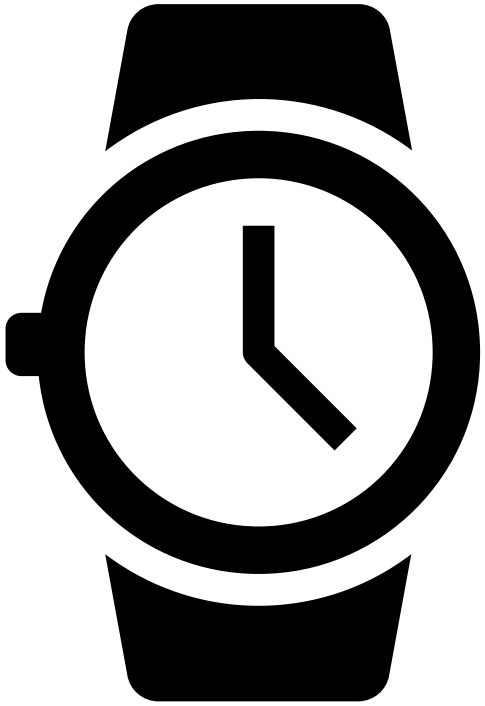
- A **dashboard** is a collection of visual reports that display important metrics and KPIs, usually in real-time.

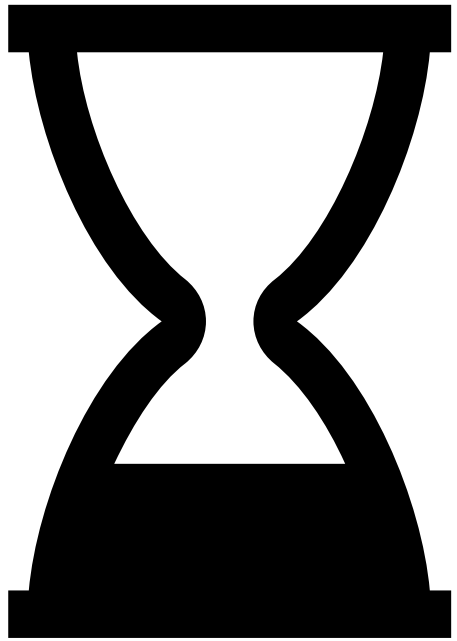
Data visualization	Data dashboard
• a visual representation of your data, such as a chart • can be static or dynamic • typically shows data for a single metric, such as electricity usage	• a collection of data visualizations assembled into a single, unified view • might display data visualizations for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

Sample dashboard



Break





Design

Designing compelling visuals

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

Visual design theory

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

Example: buying oranges

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

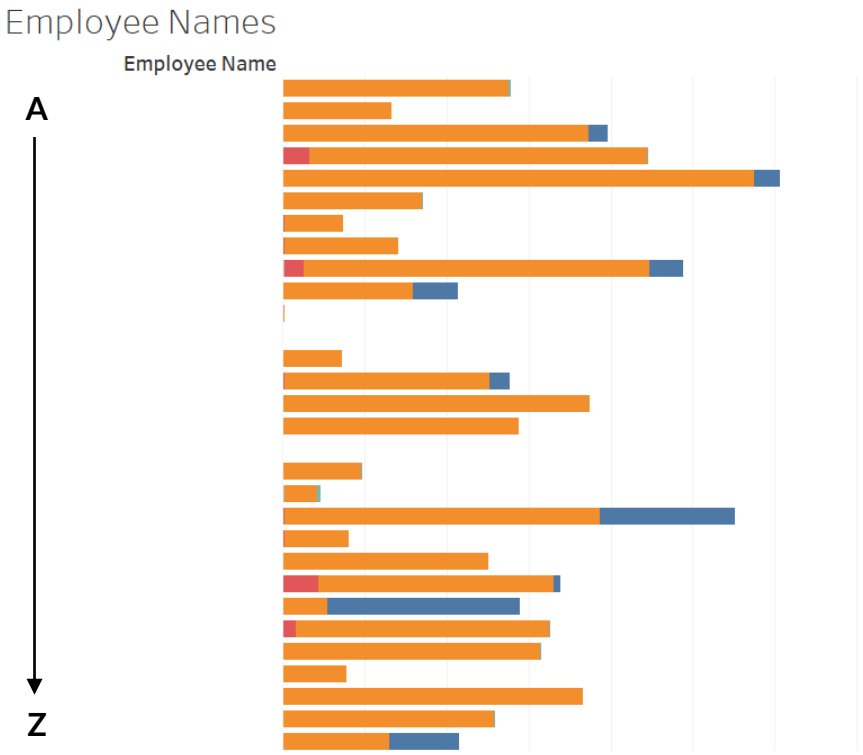
Theory

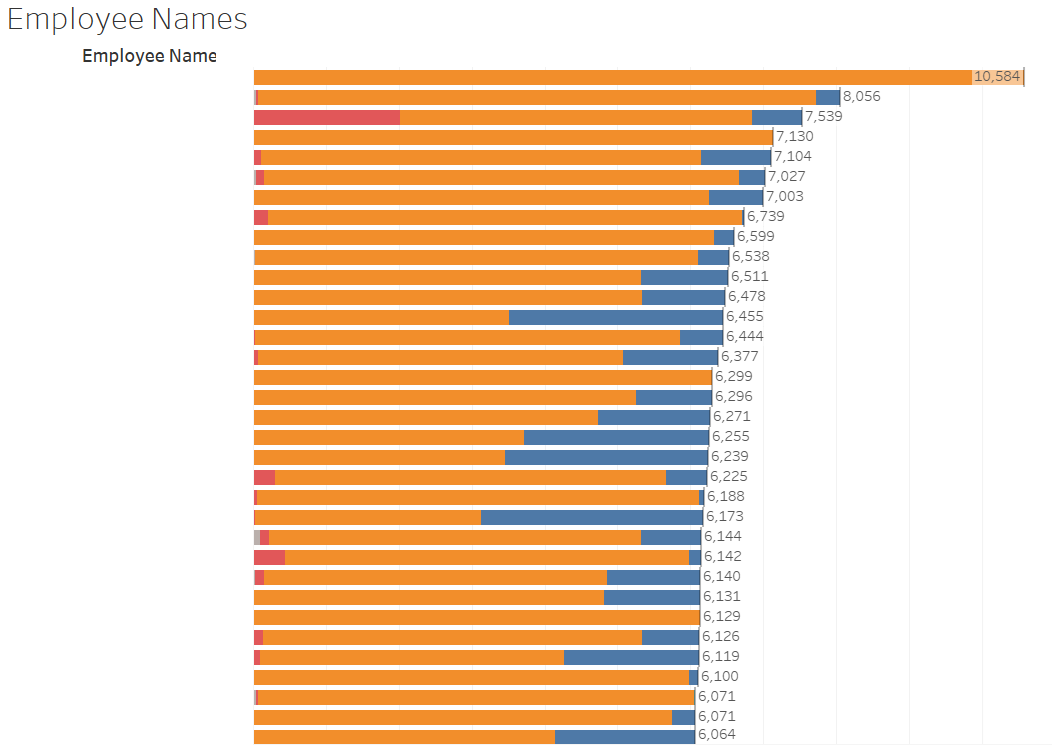
- The visual design tips we'll review today draw on theory such as:
  - the building blocks of visual design described by the Interaction Design Foundation
  - the four categories of preattentive visual attributes described in Colin Ware's book, *Information Visualization: Perception for Design*
  -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

Tip 1

Make position meaningful

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





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

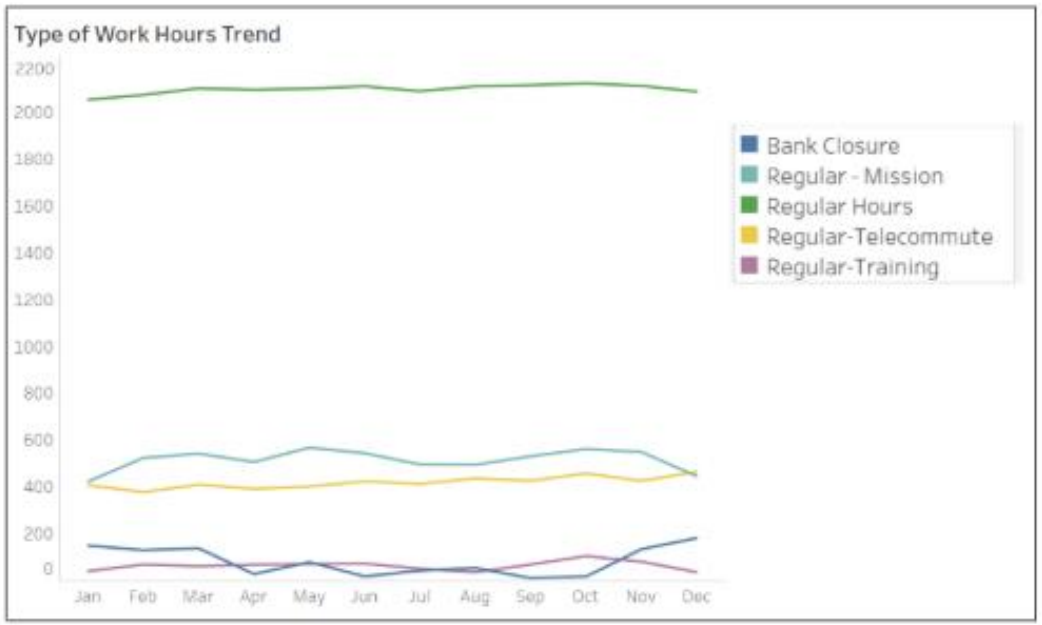
When would you use one over the other?

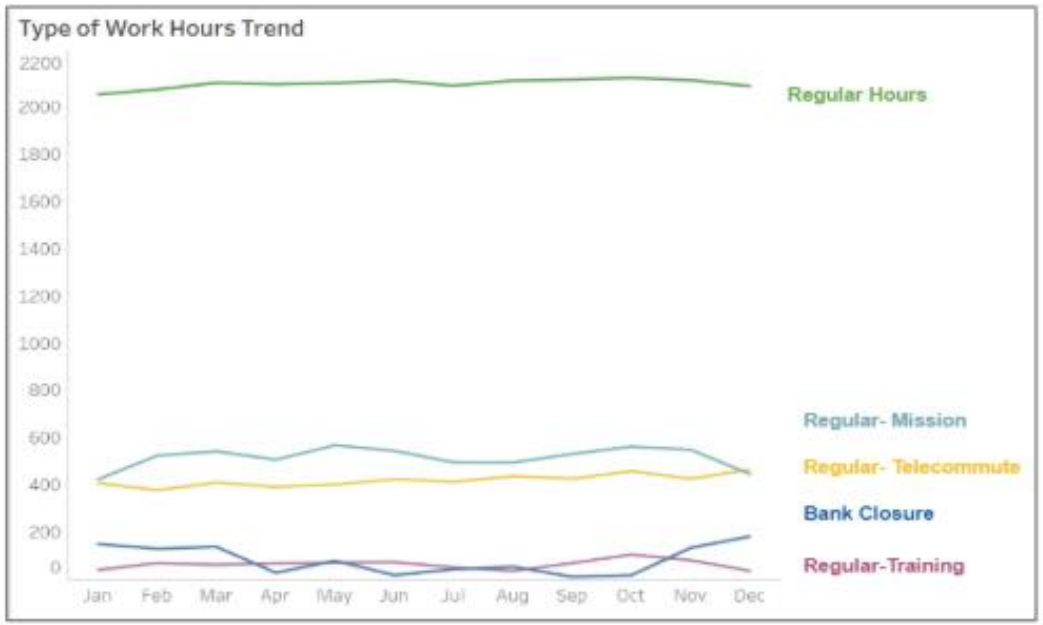
Tip 2

Group related items

- Things that are closer appear to be more related than those that are spaced farther apart.

- In fact, proximity overrules the similarity of other factors (e.g., shape, color).



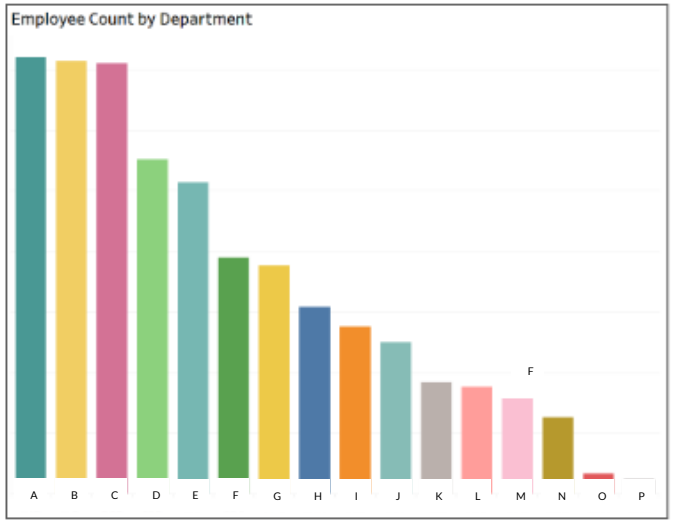


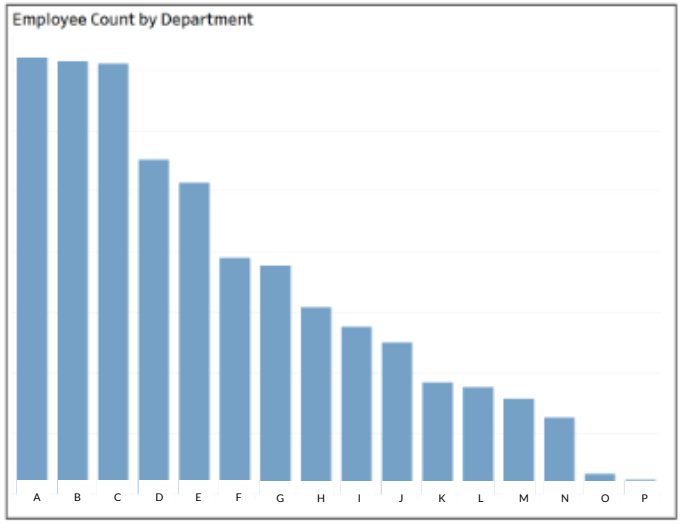
Tip 3

Distinguish different items

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

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

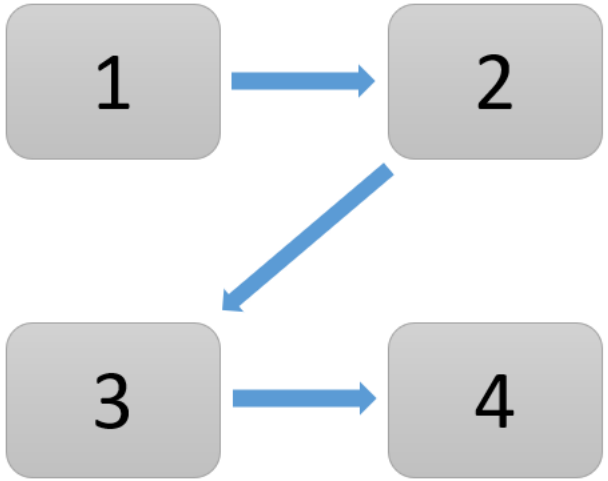




Tip 4

Use natural positioning

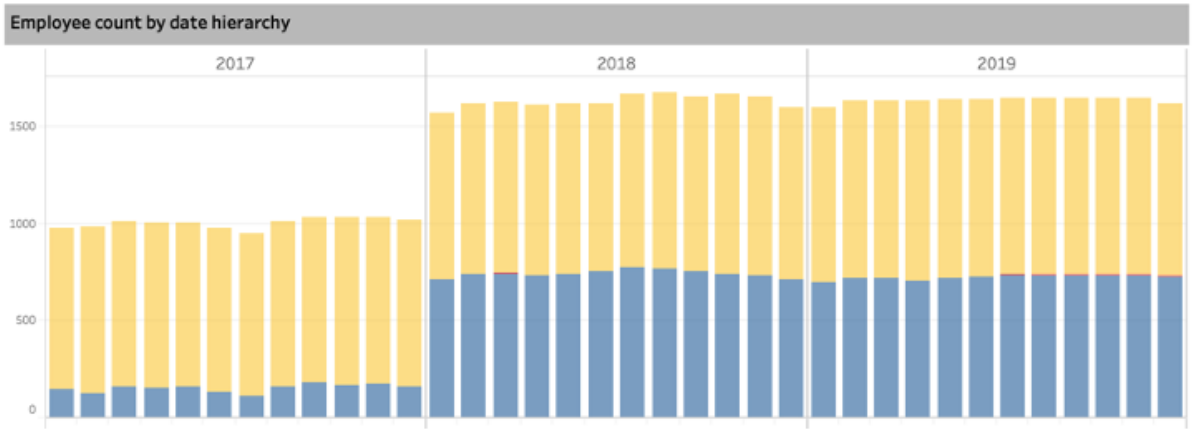
- People usually tend to start at the top left of the visual and scan in **zig-zag** motions across the page forming a **Z-pattern**.
- Aim to position elements in a way that will feel natural for users to consume.
- Also, remember that the top of the page is the most precious.

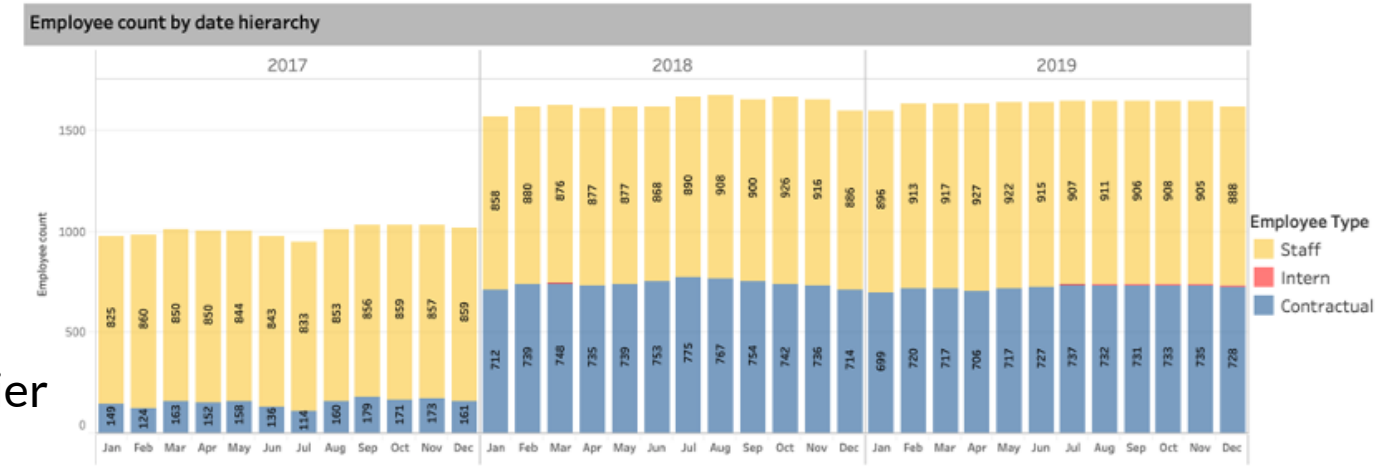


Tip 5

Use labels and legends

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



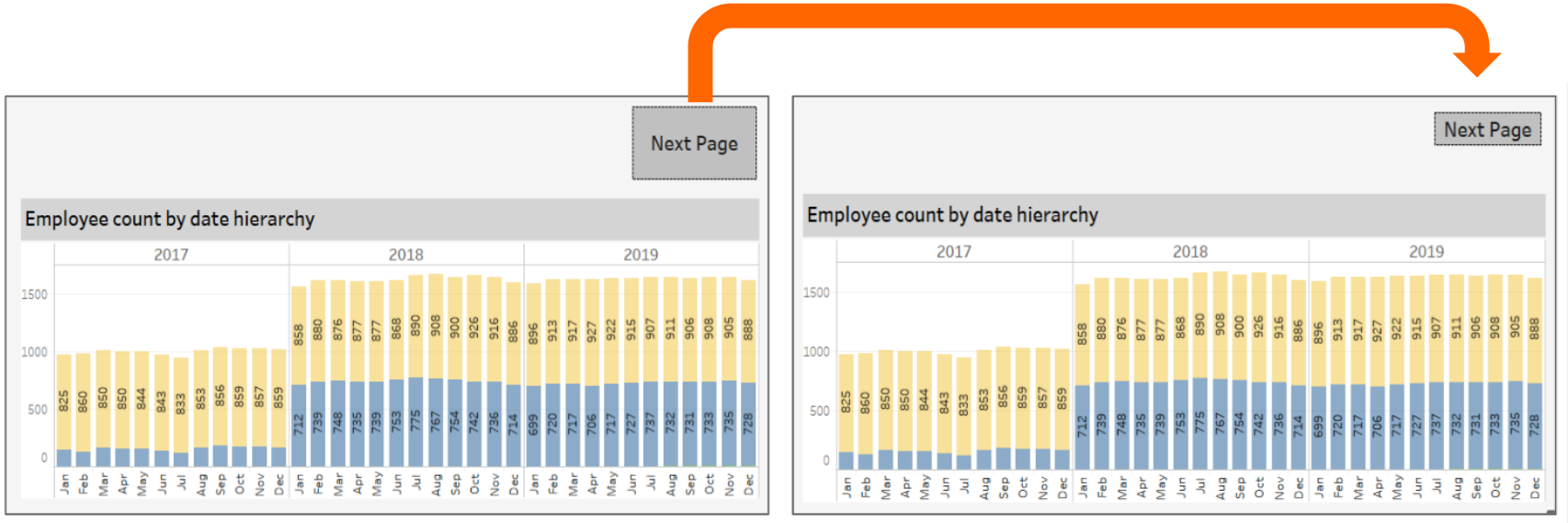


The labels and legends used in the bottom chart makes it easier to understand.

Tip 6

Use size to show importance

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



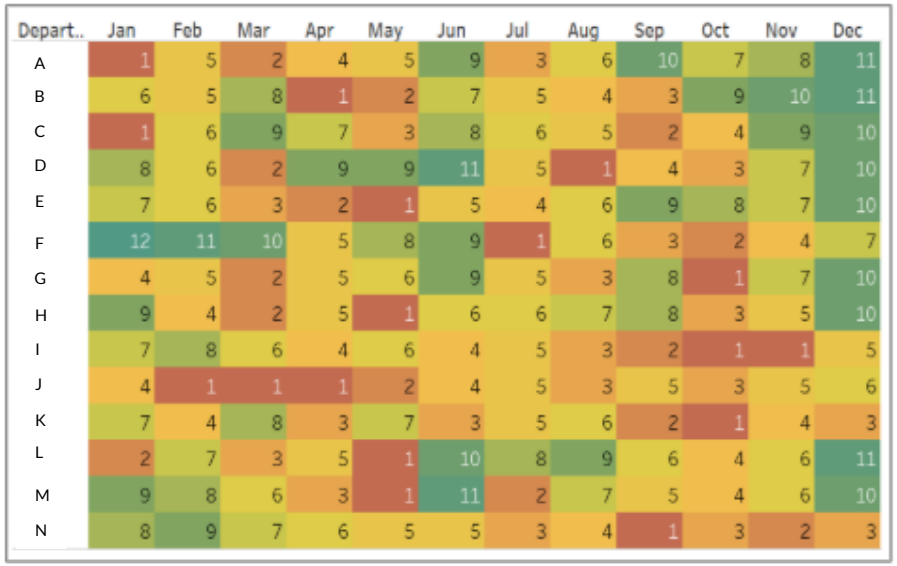
Resizing the "Next Page" button deemphasizes its importance.

Tip 7

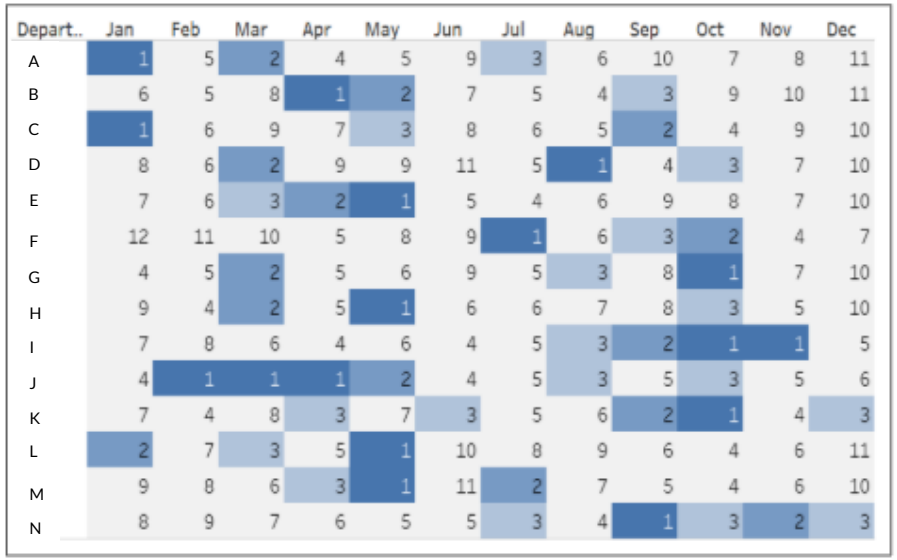
Use color to grab attention

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

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



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



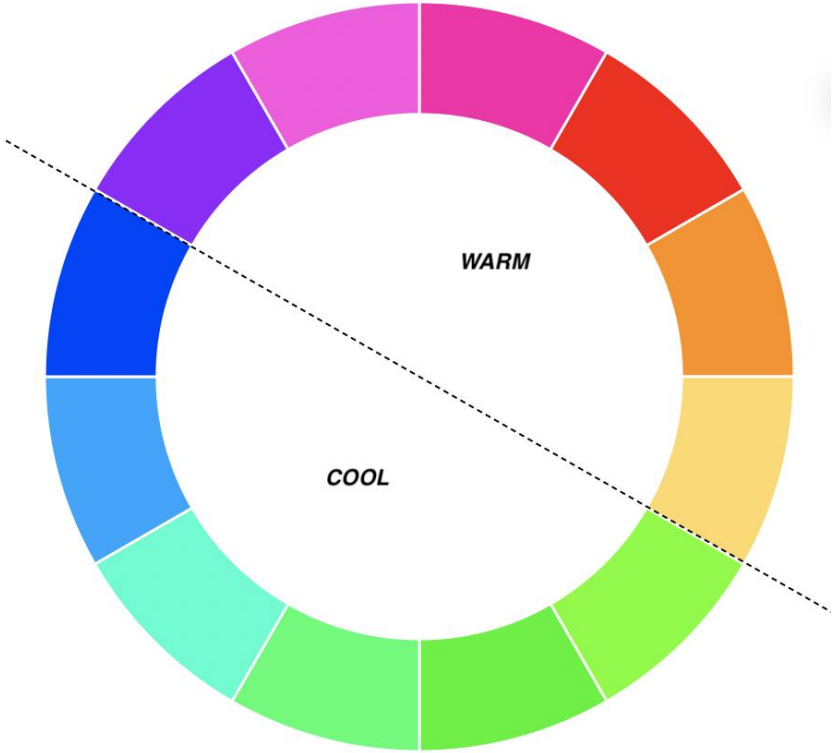
Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	1	5	2	4	5	9	3	6	10	7	8	11
B	6	5	8	1	2	7	5	4	3	9	10	11
C	1	6	9	7	3	8	6	5	2	4	9	10
D	8	6	2	9	9	11	5	1	4	3	7	10
E	7	6	3	2	1	5	4	6	9	8	7	10
F	12	11	10	5	8	9	1	6	3	2	4	7
G	4	5	2	5	6	9	5	3	8	1	7	10
H	9	4	2	5	1	6	6	7	8	3	5	10
I	7	8	6	4	6	4	5	3	2	1	1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

Tip 8

Use color to evoke emotion

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

Warm colors	represent energy
Cool colors	represent calmness



Tip 9

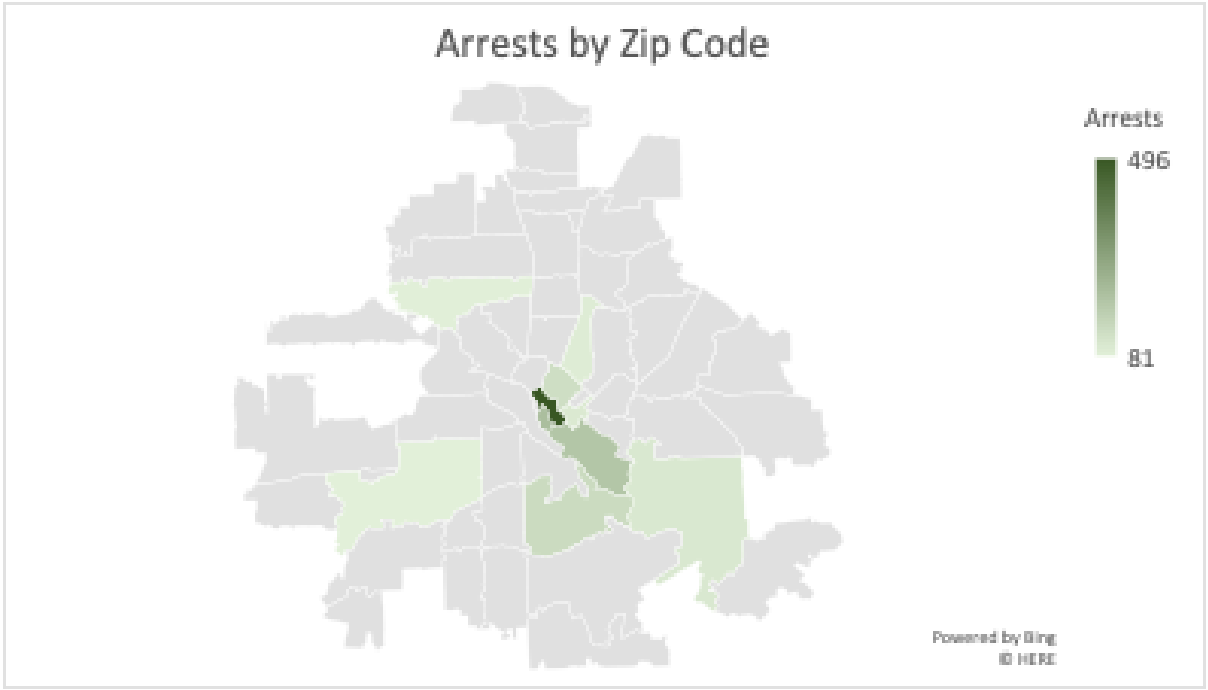
Encode data with color

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

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

Sequential color schemes

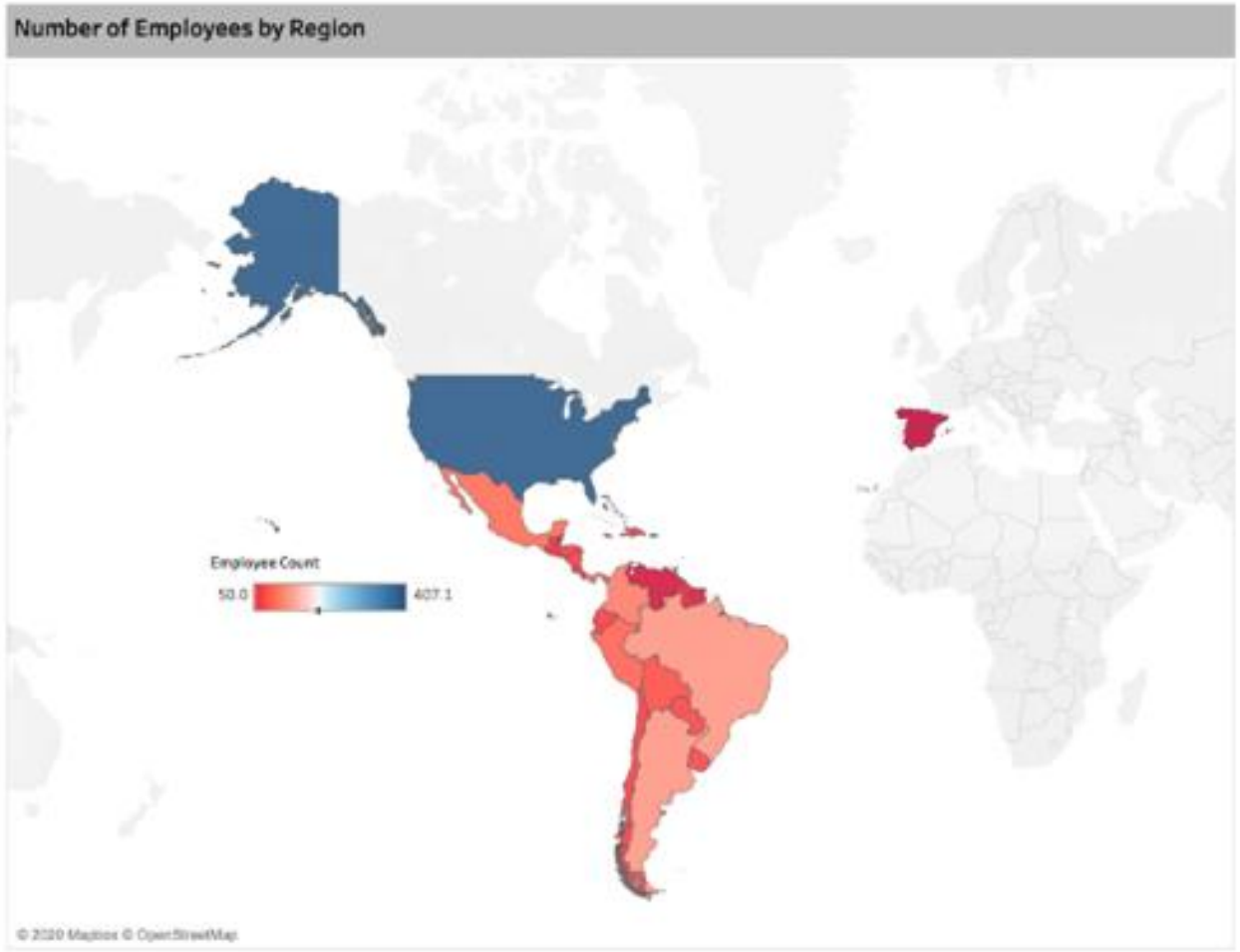
- Use a sequential color scheme when the order matters.
- These schemes range between two colors—usually a lighter shade to a darker one—by varying one or more parameters such as saturation.



Diverging color schemes

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

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



Categorical color schemes

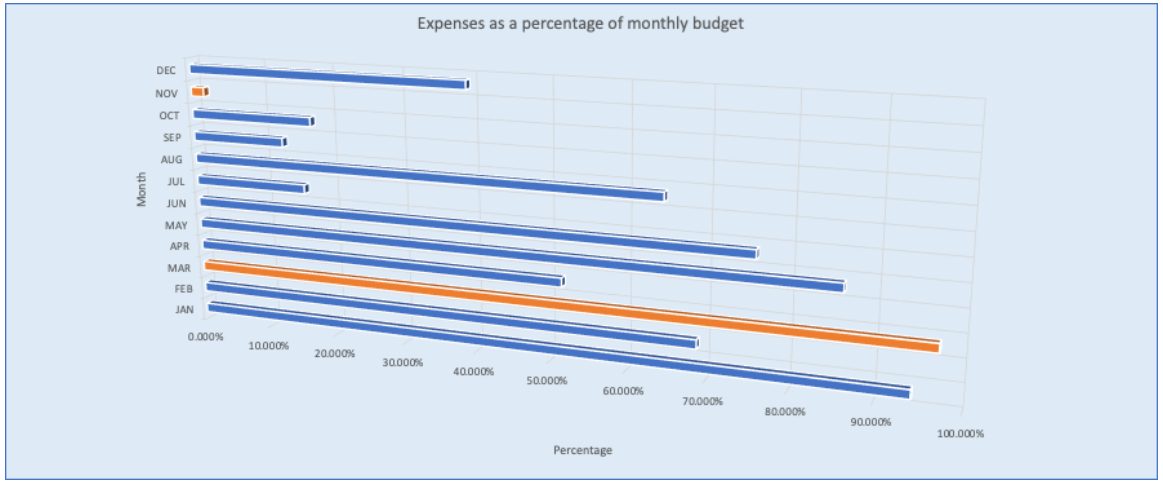
- Use a categorical color scheme for discrete data values representing distinct categories.
- These schemes use different hues with consistent steps in lightness and saturation.

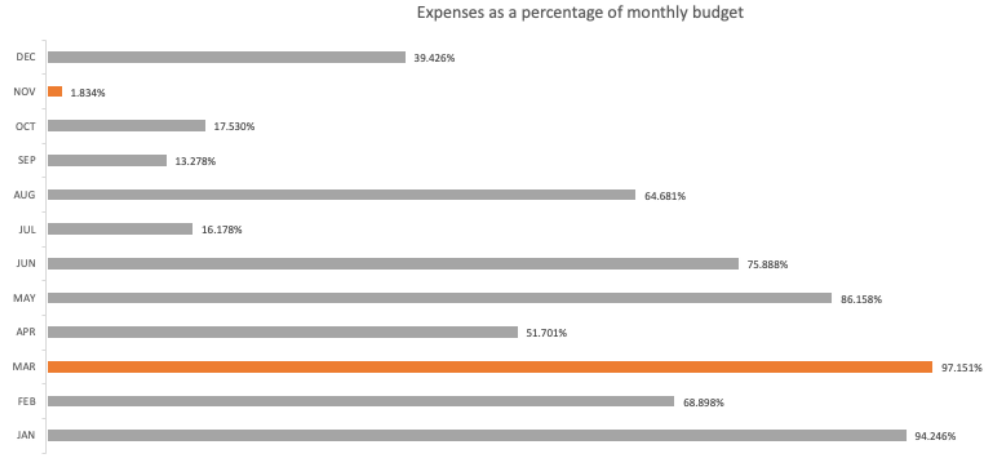
Tip 10

Reduce chart clutter

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

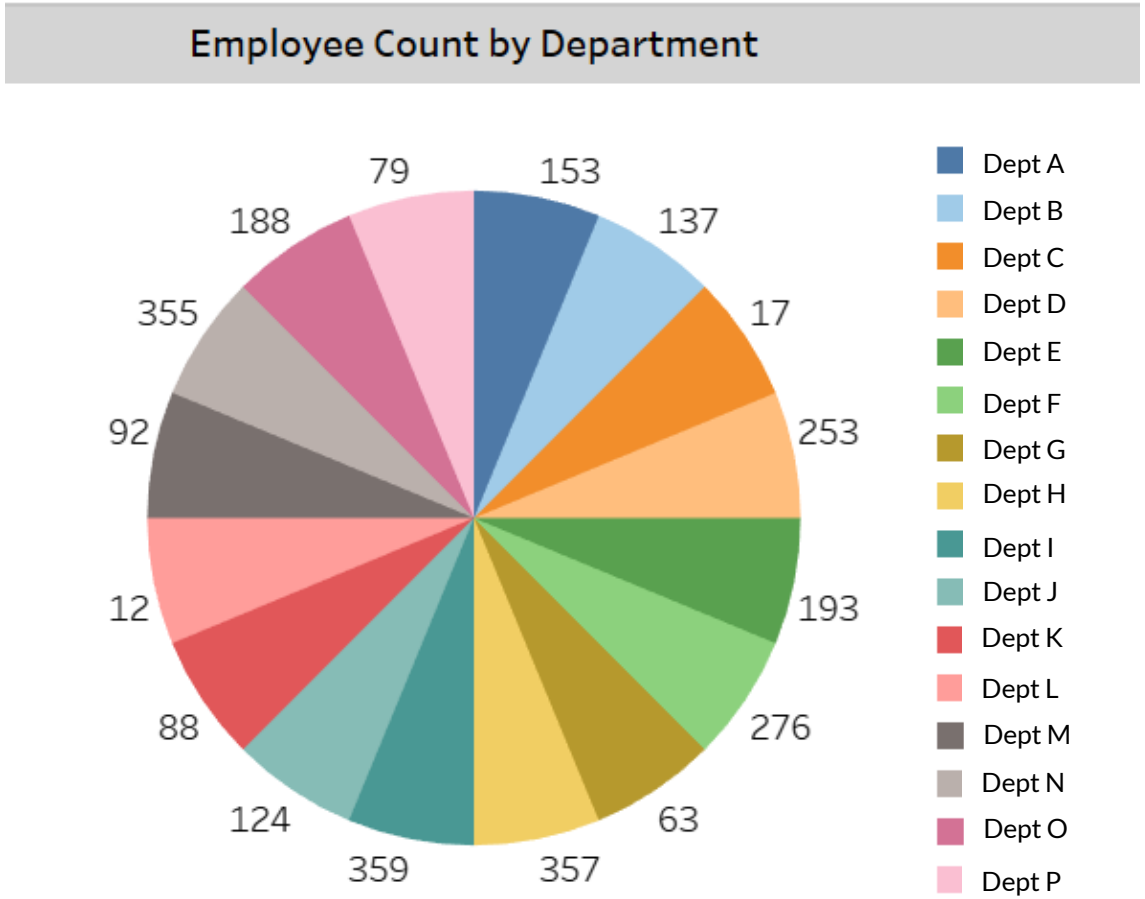
1. Remove special effects
2. Lighten the background
3. Remove chart borders
4. Remove gridlines
5. Direct label
6. Clean up axis titles and labels
7. Use consistent colors





Common mistakes

Poll question



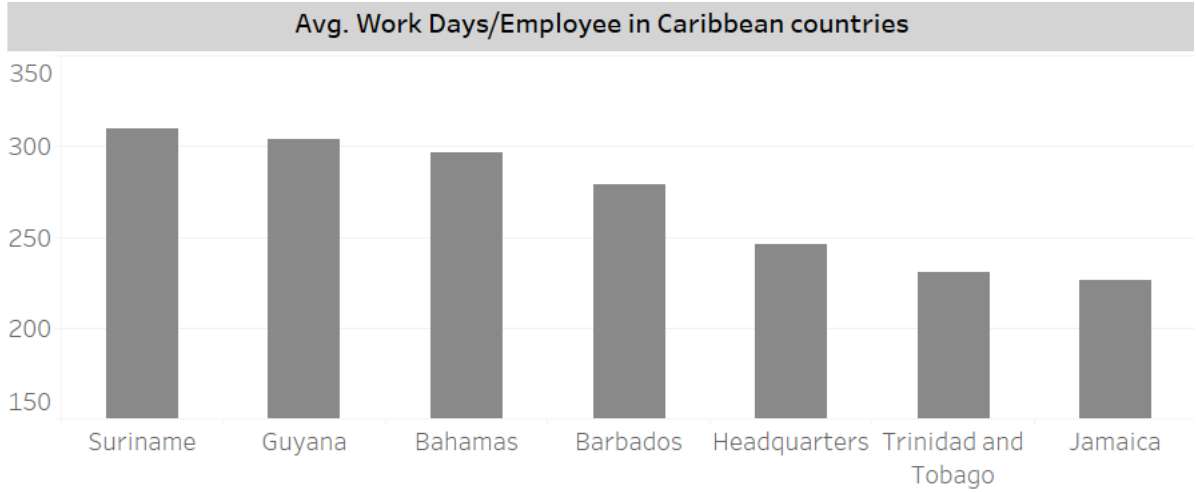
What type of visual might work better for this data?

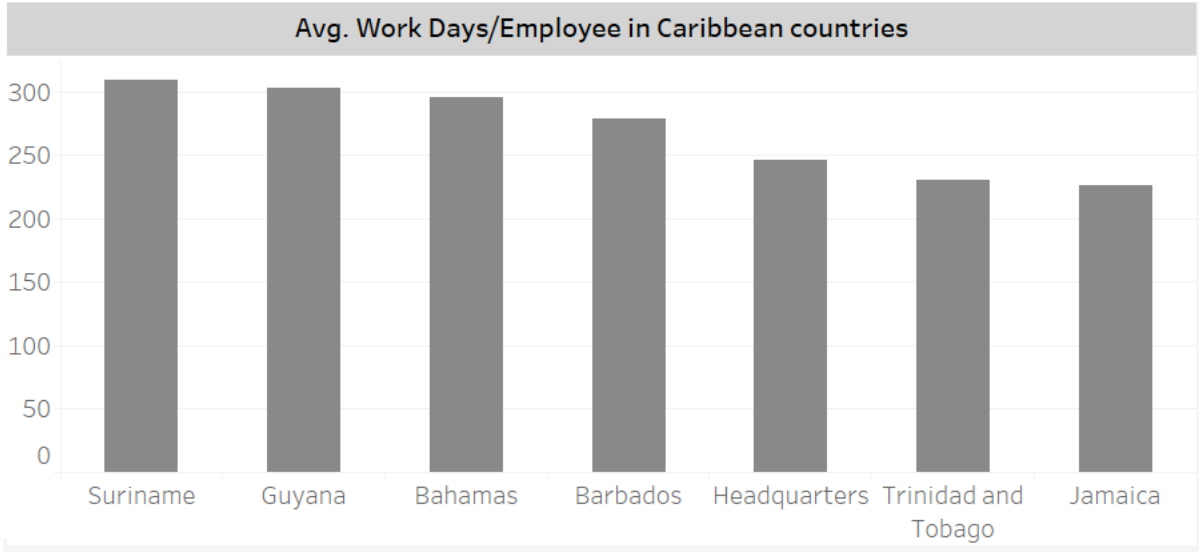
Mistakes to avoid

- Poor data visualizations confuse the viewer or, worse, mislead them and cause more harm than good.
- Poor visuals can cause a loss of time and effort and might delay the decision-making process in deadline-driven projects.
- Let's talk about a few common mistakes…

Misleading data

- Sometimes charts look presentable but could be misleading.
- Unreliable data comparisons erode credibility and eventually dissuade viewers from using your analysis.
- Look at the top graph. At first, Jamaica seems to have half the average workdays per employee that Suriname does. In reality, the difference is much less.



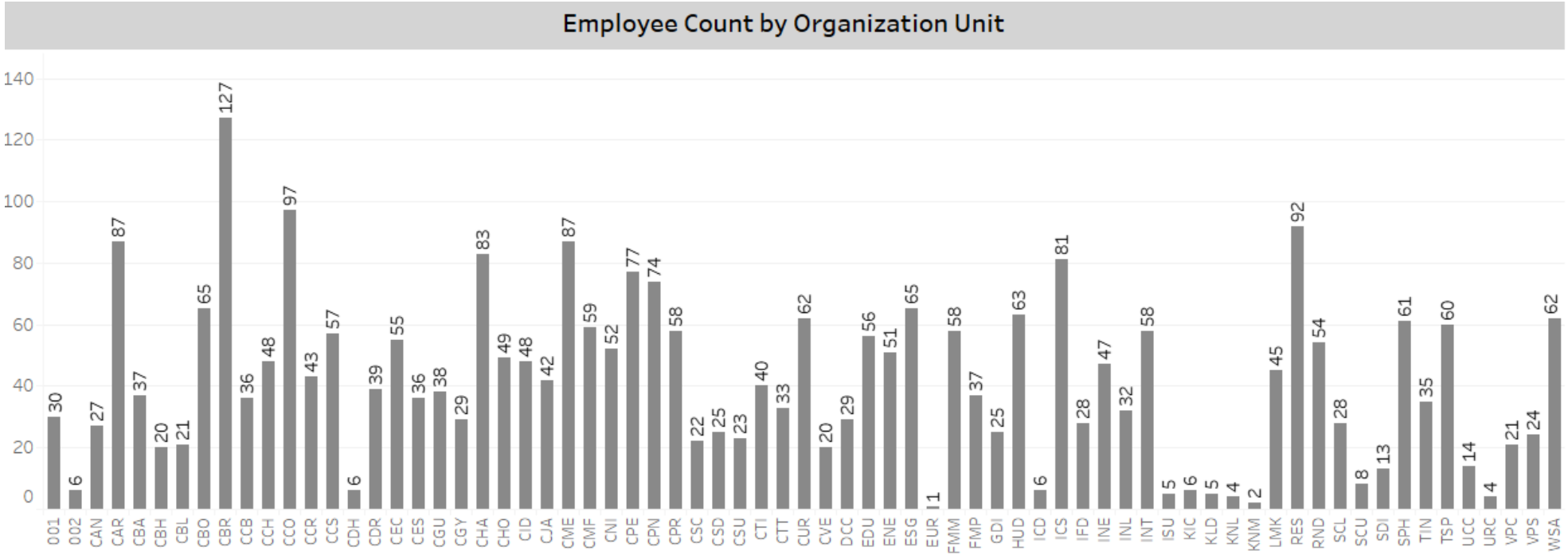


Making difficult comparisons

- Sometimes we like using bright and colorful, graphic visuals to brighten up a mundane report. However, aesthetics should never supersede substance.
- A plain, less creative column chart would be better to represent this information:

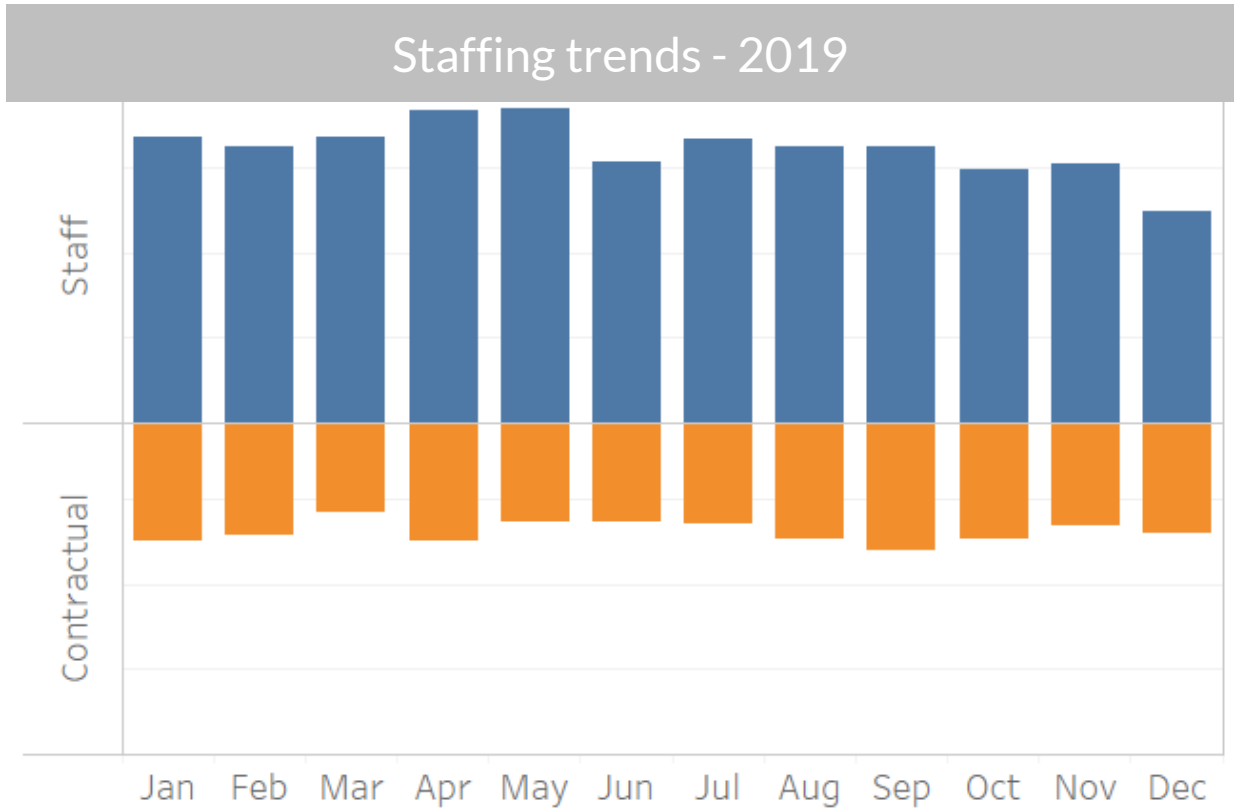
Plotting too much data

- When there are many data points, visualizing data for the sake of visualization is counterproductive and unhelpful to the viewer.
- The visualization below fails to properly compare employee count by location, and in cramming in too much information, is simply overwhelming.



Not following conventions

- Usually, a graph moves up and to the right.
- *Look at the graph below. What do you think happened?*



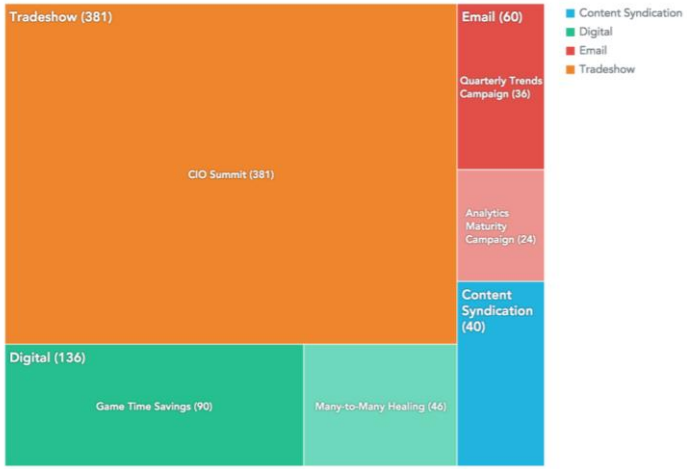
Other common mistakes

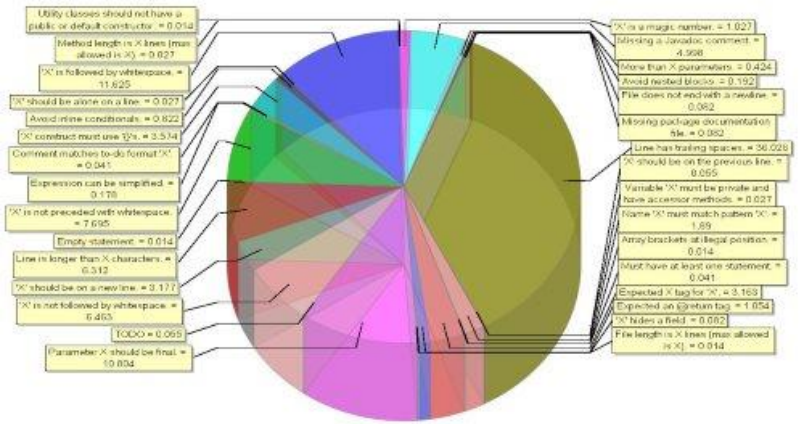
- Other common mistakes include:
  - Using the wrong chart type
  - Presenting numbers that don't add up
  - Not annotating
  - Using 3D charts poorly

Activity: assess visualizations

- Turn to page 13 of your participant guide to find the **Analyzing visualizations** activity.

- You will be asked to assess 4 visualizations. Write down your notes.





Poll questions

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

- Change colors
- Remove extra information
- Add more information

Which chart is the best?

Visualization tools

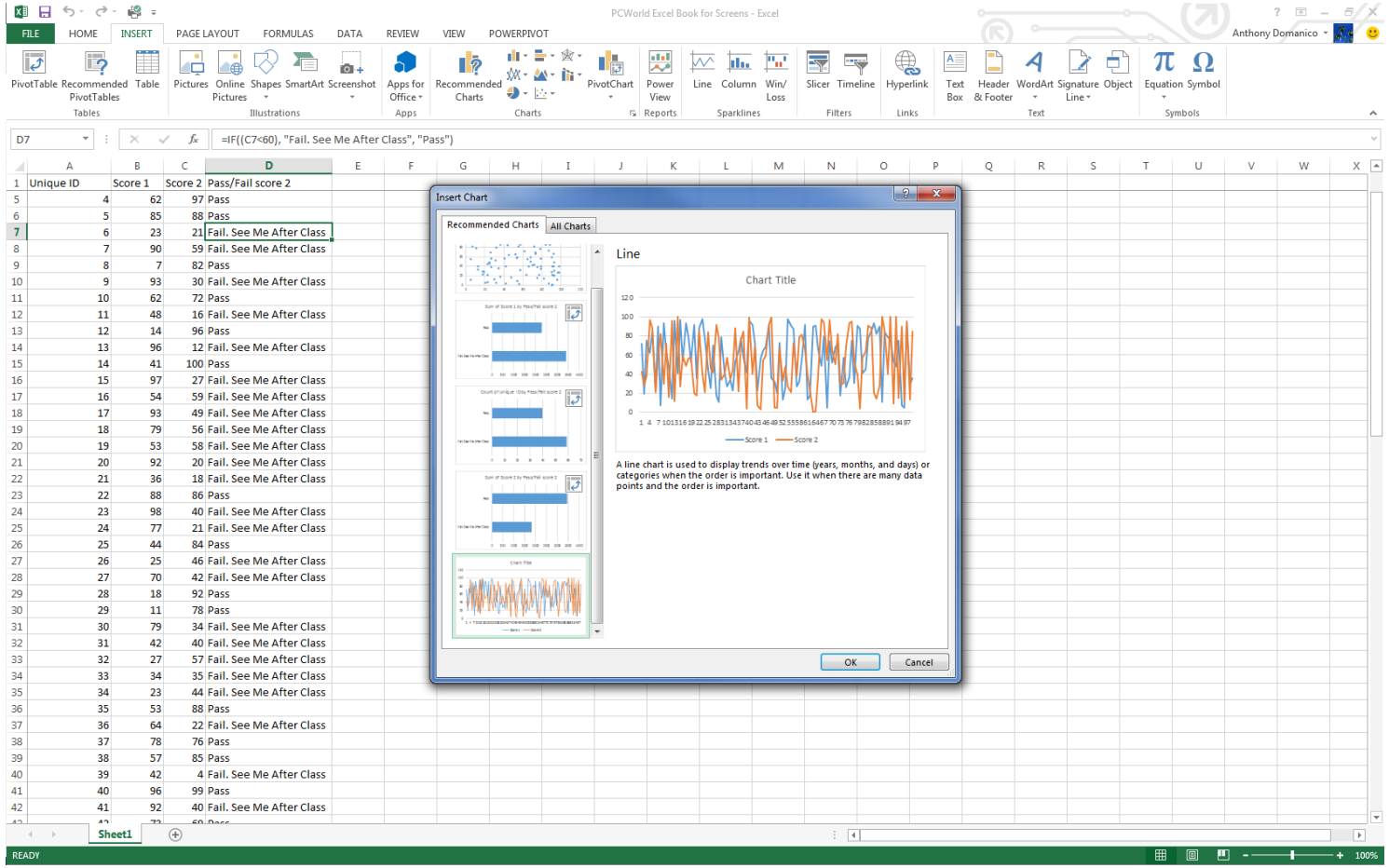
Poll question

What tools have you used to visualize data?

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

Excel

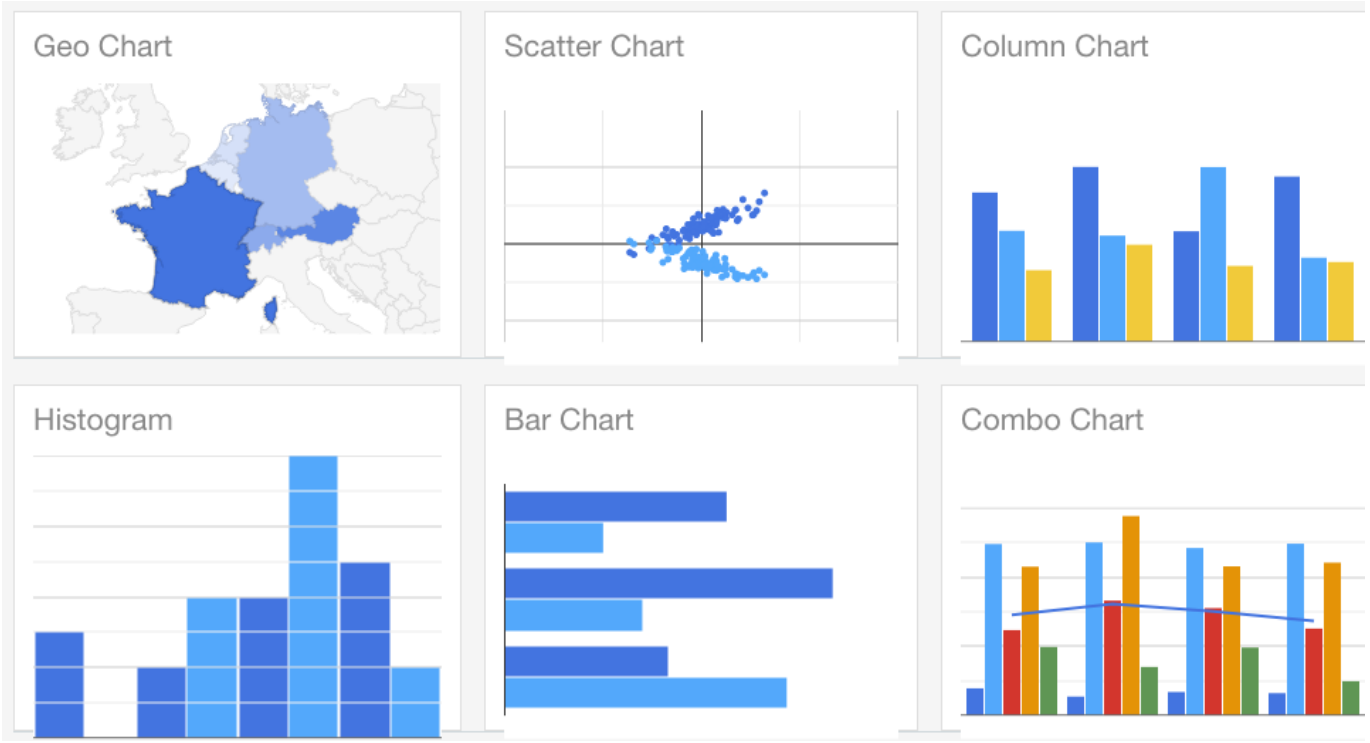
- Create basic chart types such as pie, line, bar, scatter, and more.
- Charts created in Excel can easily be ported to PowerPoint and Word.



Google Charts

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

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

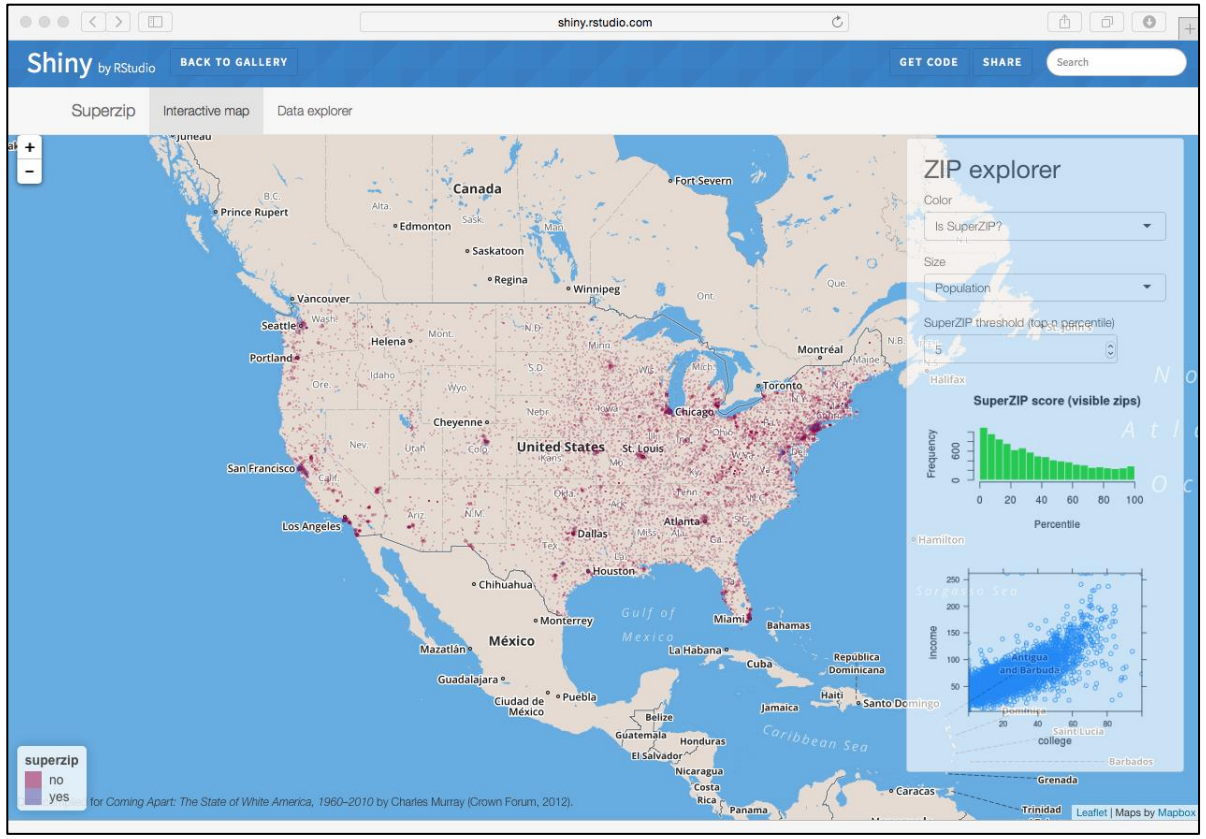


Tableau

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

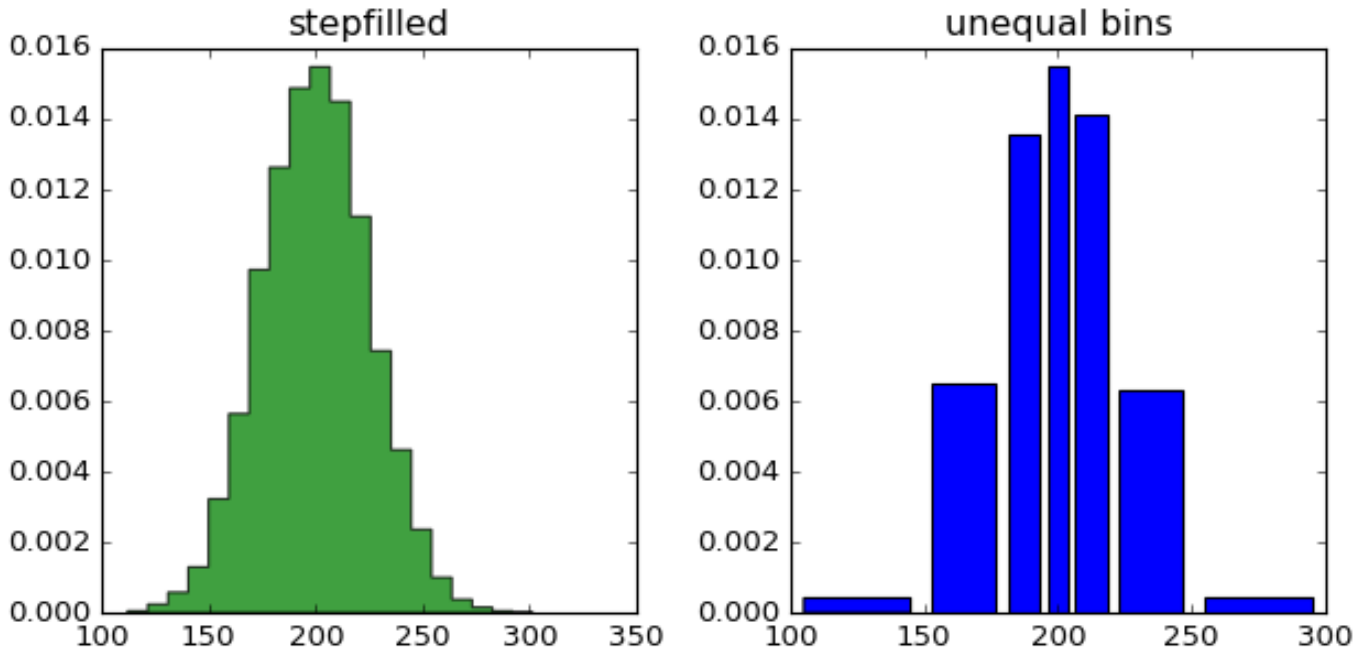
R and RStudio

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



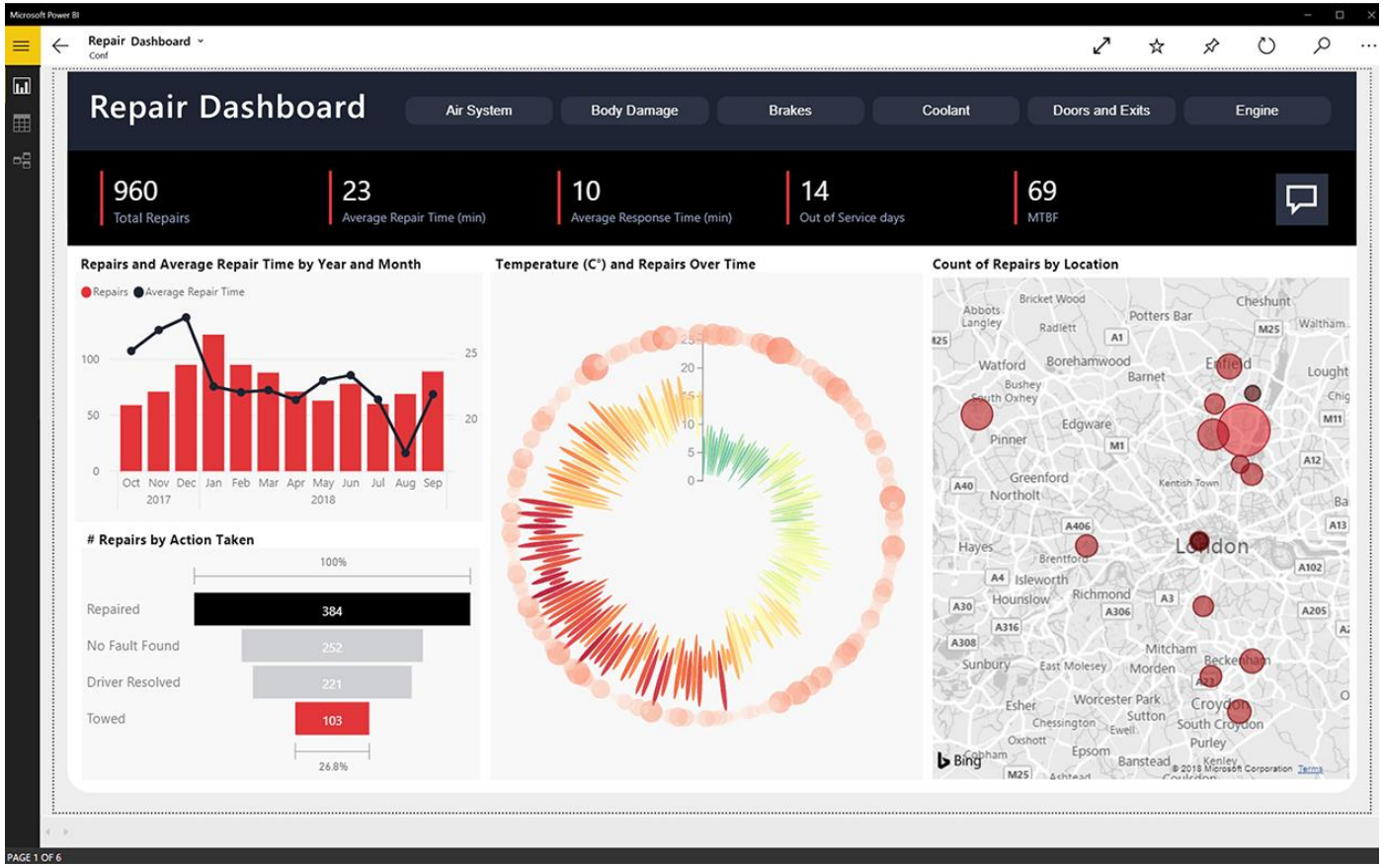
Python

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

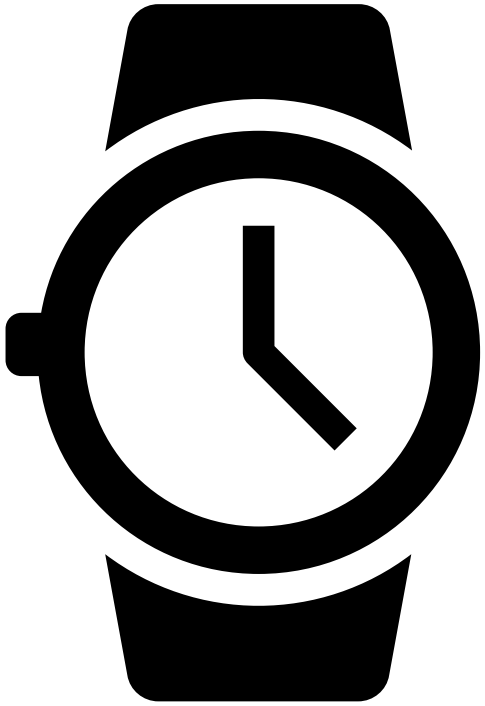


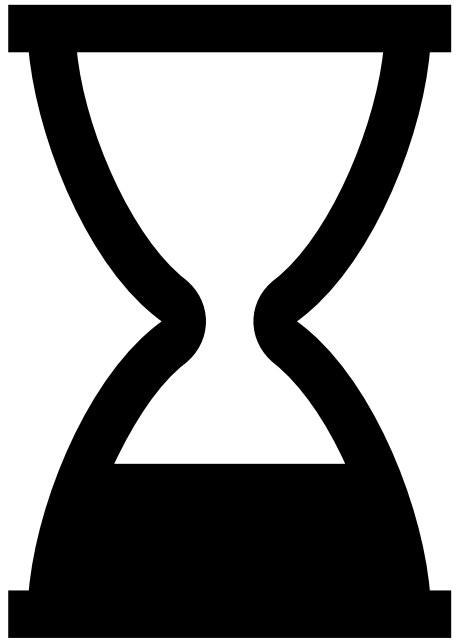
Power BI

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



Break





Agenda

Day 4

- Data visualization
- Data storytelling

- How do I tell a story with data to inspire action?

What is data storytelling?

1. You focus on an **insight** and
2. persuade an **audience**
3. that the **outcome** of your analysis
4. demands a course of **action**
5. through narrative and visual **communication.**

Data stories and data visualizations

- A **single** data story may make use of **multiple** data visualizations.
- Data stories arrange visualizations into the **linear sequence** of storytelling: a beginning, a middle, and an end.
- Data story formats will likely incorporate other elements to explain and contextualize the visualizations:
  - prose text, either written or spoken
  - annotations, callouts, and labels
  - icons or graphics
  - images or photographs

Can’t I just use a chart?

1. Narratives are super effective, “**sticky**” content delivery mechanisms.

2. Not everyone is a statistician, but they still want to make **evidence-based** decisions.

3. Stories let you overview key findings **quickly.**

4. Stories tap into both the **logical** and the **emotional** aspects of persuasion.

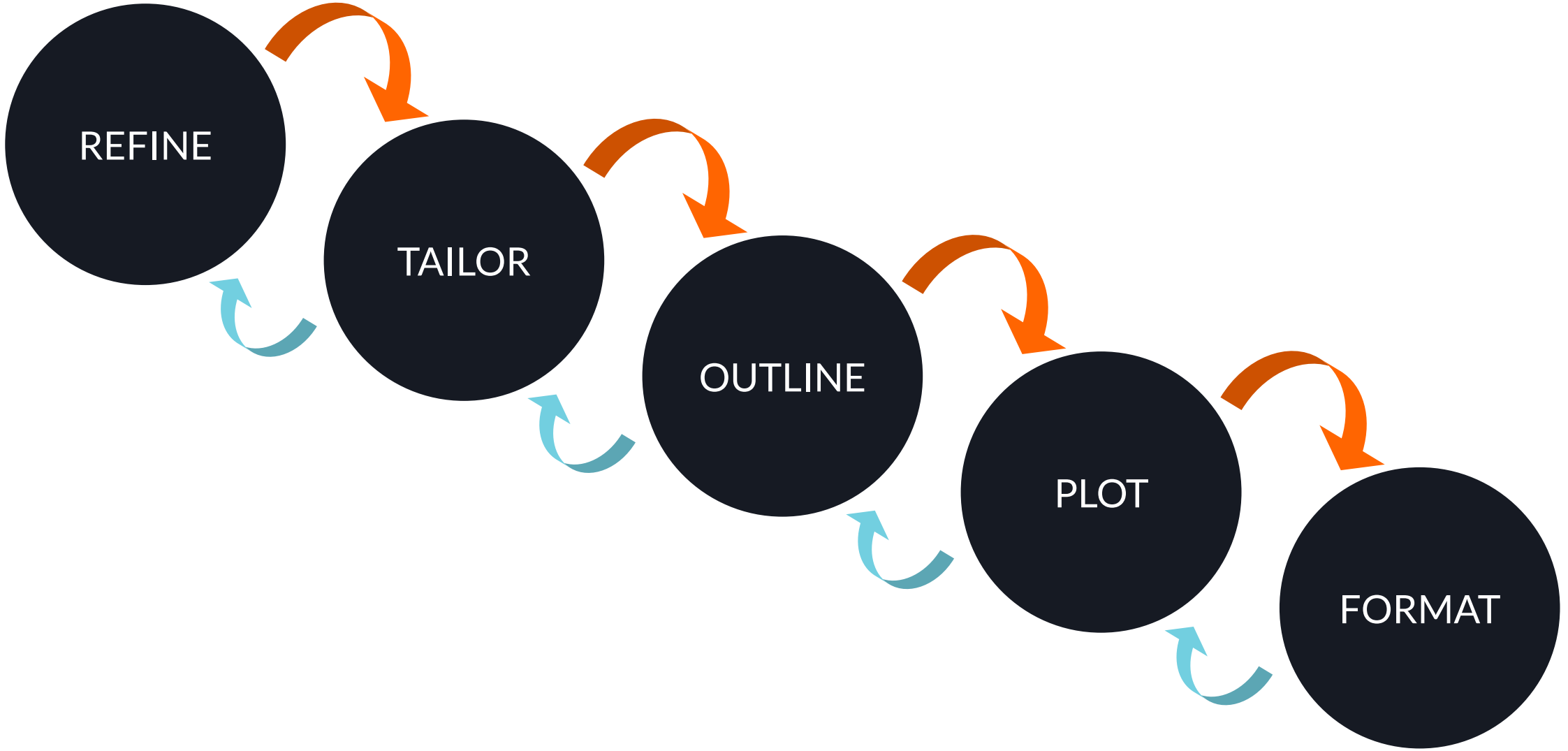
Why choose story?

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

Why choose story?, ctd.

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

How do I craft a data story?



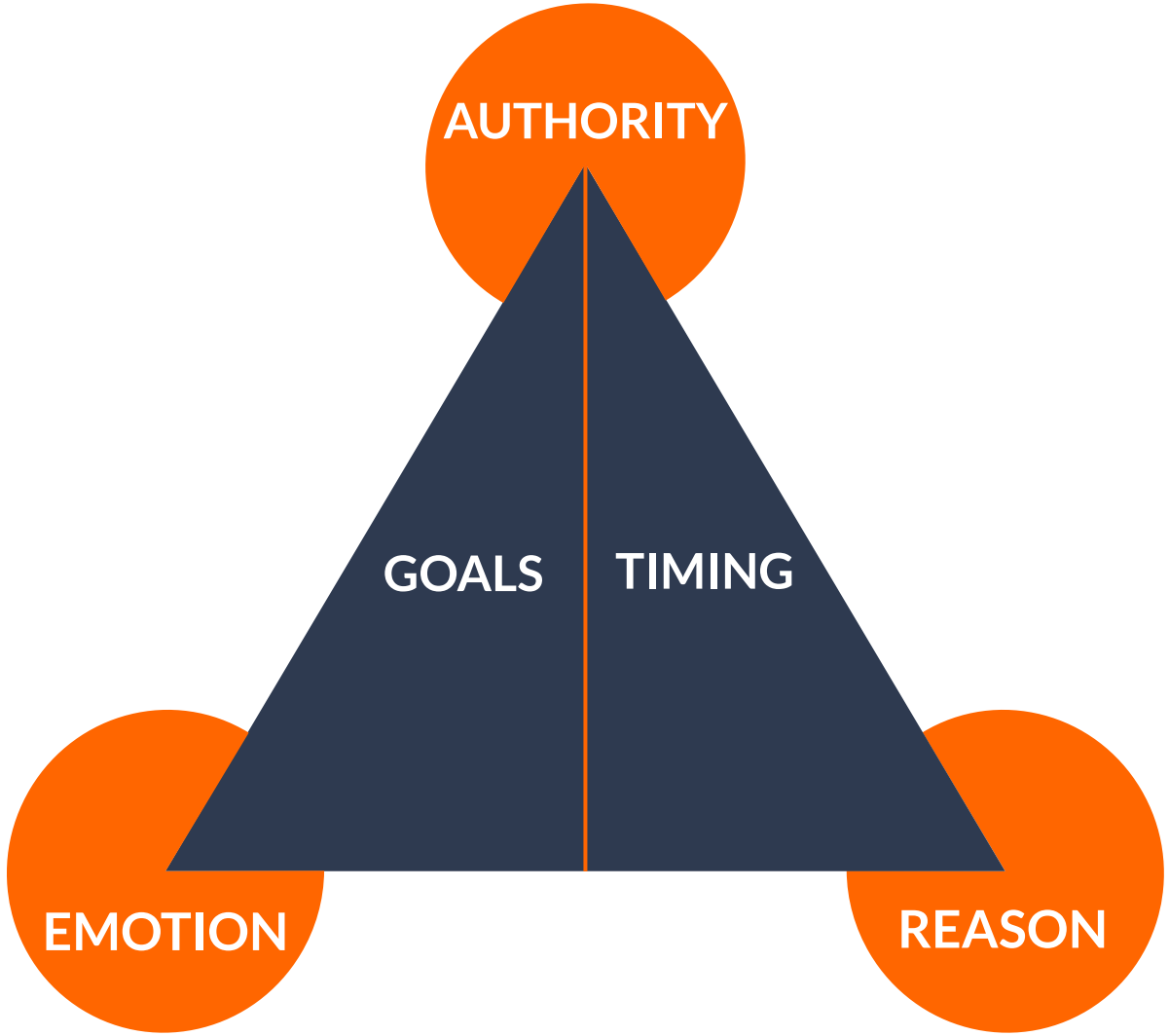
REFINE

Refining your insight

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

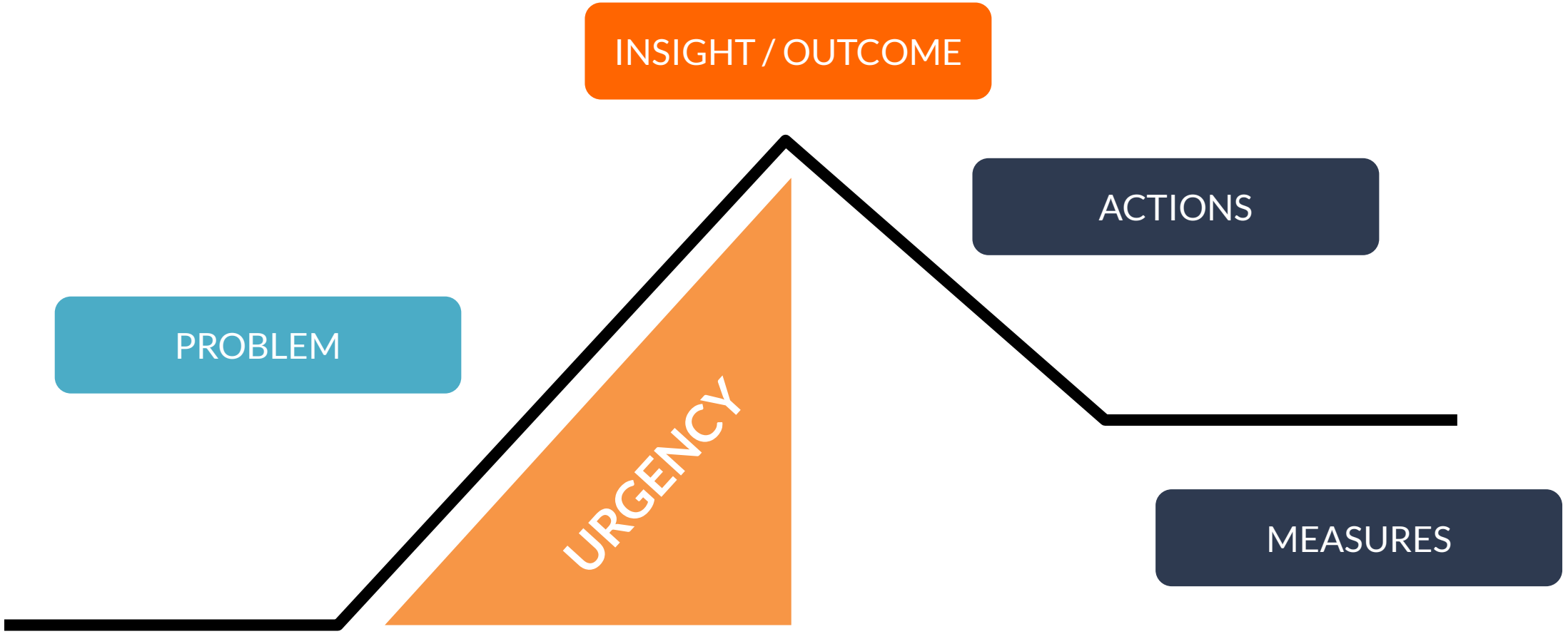
TAILOR

Tailoring to your audience



OUTLINE

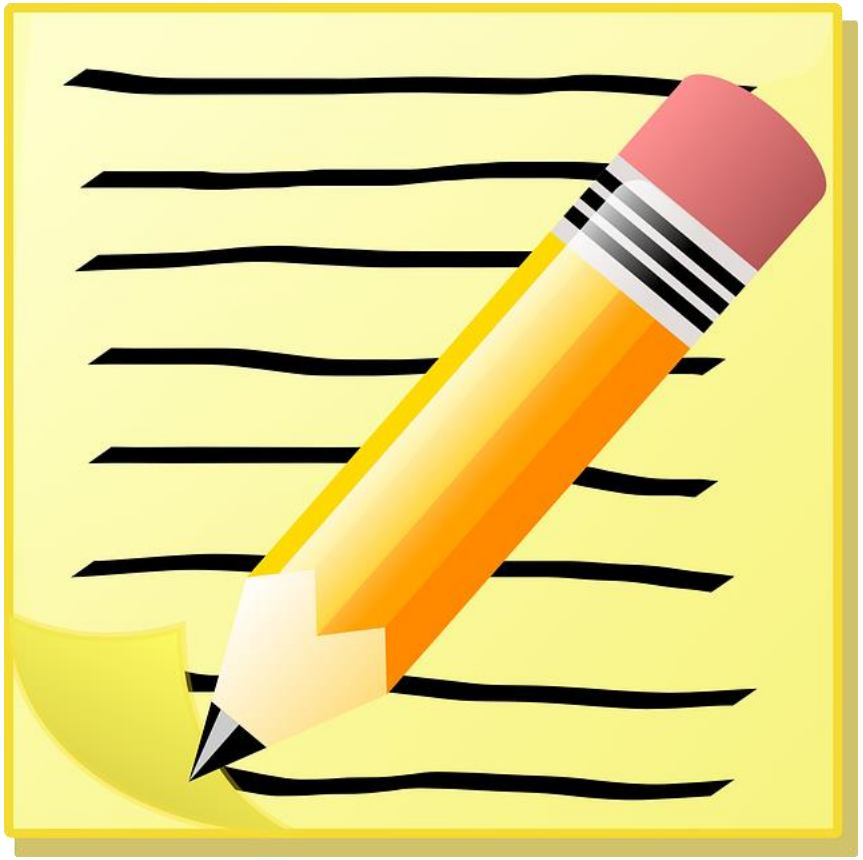
Outlining



PLOT

Plotting with a storyboard

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



FORMAT

Formatting for delivery

- You may find yourself needing to alter the way you tell your data story based on the affordances of the format.

- Sometimes the format is a given, but other times, it will depend upon your input and the use case.

- As with visualizations, the simplest storytelling format is often the best.

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

The Joy of Stats

DATA SOCIETY®

Thank you!