SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

SI 568 - Introduction to Applied Data Science

Ethics and Communication in Data Science

Learning Objectives

Ethics

- Current state of ethics and data science
- Understand potential harms of data collection, aggregation, and analysis

Communication

- Introduction to technical reports
- Data visualization basics
- Verbal communication

Data Science

What is Data Science

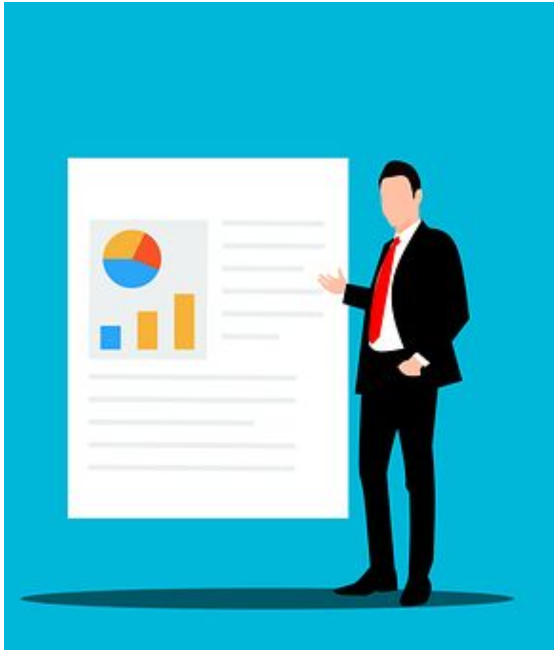
Introduction

- "Data science combines the scientific method, math and statistics, specialized programming, advanced analytics, AI, and even storytelling to uncover and explain the business insights buried in data." - IBM
- "This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, traditional, and inter-disciplinary applications" - David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766. DOI: 10.1080/10618600.2017.1384734

Data Science includes...

Programs, Information and People

- Data
  - gathering
  - cleaning
  - modeling
- Statistical analysis
- Interpretation of results
- Communication of results
  - visualizations
  - reports
  - speaking



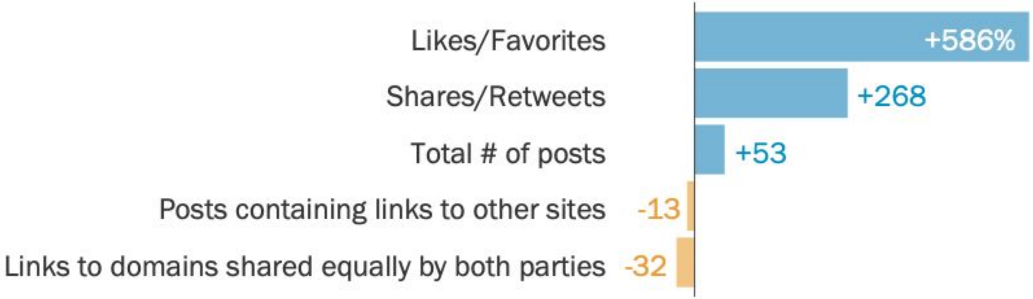
Data Science Uses

Analysis to research trends in public opinion and demographics

Pew Research Center

Lawmaker social media during 2020 election vs. 2016: More posts and audience engagement, but fewer links to sites shared equally by both parties

% change in ____ from 2016 election to 2020 election on lawmaker Facebook & Twitter accounts



Reference: https://www.pewresearch.org/politics/2021/09/30/charting-congress-on-social-media-in-the-2016-and-2020-elections/pdl_09-30-21_social-media_-congress0/

What does a data scientist do?

A data scientist might do the following tasks on a day-to-day basis:

- Find patterns and trends in datasets to uncover insights
- Create algorithms and data models to forecast outcomes
- Use machine learning techniques to improve quality of data or product offerings
- Communicate recommendations to other teams and senior staff
- Deploy data tools such as Python, R, SAS, or SQL in data analysis
- Stay on top of innovations in the data science field

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Ready to be a data scientist?

Ethics

What is Ethics

Definition

"ethics, also called moral philosophy, the discipline concerned with what is **morally good and bad and morally right and wrong**. The term is also applied to any system or theory of moral values or principles."

- Britannica

Morals: "a person's standards of behavior or beliefs concerning what is and is not acceptable for them to do." - Oxford Languages

What is Ethics in Data Science?

"Data ethics is a new 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) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values)."

-Luciano Floridi and Mariarosaria Taddeo

Reference: https://royalsocietypublishing.org/doi/10.1098/rsta.2016.0360#

Codes of Ethics

Where does Data Science meet Ethics?
Why you can't just analyze the data you get

A code of ethics sets out an organization's ethical guidelines and best practices to follow for honesty, integrity, and professionalism. For members of an organization, violating the code of ethics can result in sanctions including termination.

Reference: https://www.investopedia.com/terms/c/code-of-ethics.asp

Types of Codes of Ethics

The main types of codes of ethics include:

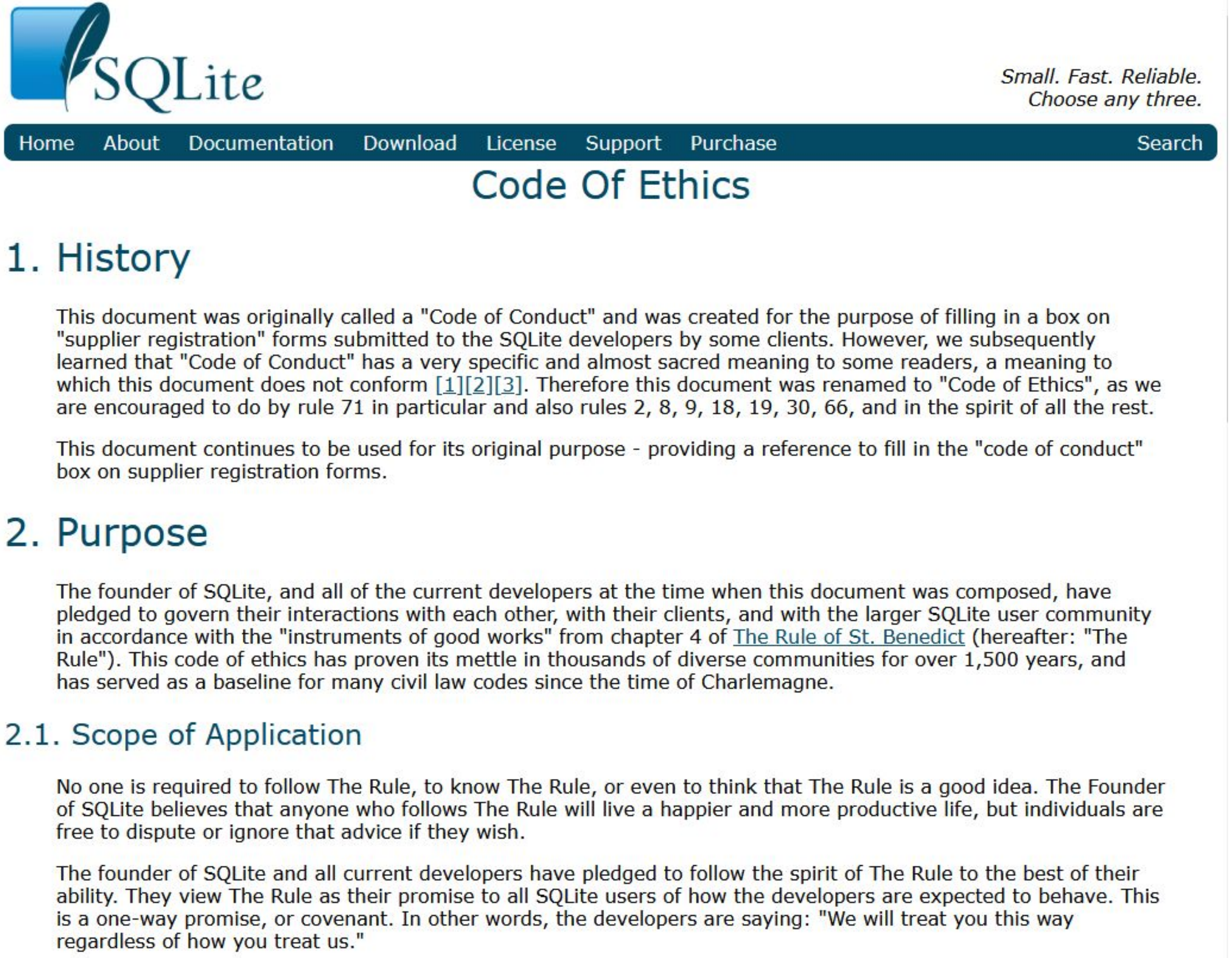
- a compliance-based code of ethics,
- a value-based code of ethics,
- and a code of ethics among professionals.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Examples of Code of Ethics related to data and technology

SQLite code of Ethics

Parts 1 and 2 of 3 parts
Note 2.1 about The Rule



Code Of Ethics

1. History

This document was originally called a "Code of Conduct" and was created for the purpose of filling in a box on "supplier registration" forms submitted to the SQLite developers by some clients. However, we subsequently learned that "Code of Conduct" has a very specific and almost sacred meaning to some readers, a meaning to which this document does not conform [1][2][3]. Therefore this document was renamed to "Code of Ethics", as we are encouraged to do by rule 71 in particular and also rules 2, 8, 9, 18, 19, 30, 66, and in the spirit of all the rest.

This document continues to be used for its original purpose - providing a reference to fill in the "code of conduct" box on supplier registration forms.

2. Purpose

The founder of SQLite, and all of the current developers at the time when this document was composed, have pledged to govern their interactions with each other, with their clients, and with the larger SQLite user community in accordance with the "instruments of good works" from chapter 4 of The Rule of St. Benedict (hereafter: "The Rule"). This code of ethics has proven its mettle in thousands of diverse communities for over 1,500 years, and has served as a baseline for many civil law codes since the time of Charlemagne.

2.1. Scope of Application

No one is required to follow The Rule, to know The Rule, or even to think that The Rule is a good idea. The Founder of SQLite believes that anyone who follows The Rule will live a happier and more productive life, but individuals are free to dispute or ignore that advice if they wish.

The founder of SQLite and all current developers have pledged to follow the spirit of The Rule to the best of their ability. They view The Rule as their promise to all SQLite users of how the developers are expected to behave. This is a one-way promise, or covenant. In other words, the developers are saying: "We will treat you this way regardless of how you treat us."

SQLite code of Ethics

The Rule
(items 50-72)

3. The Rule

50. When wrongful thoughts come into your heart, dash them against Christ immediately.
51. Disclose wrongful thoughts to your spiritual mentor.
52. Guard your tongue against evil and depraved speech.
53. Do not love much talking.
54. Speak no useless words or words that move to laughter.
55. Do not love much or boisterous laughter.
56. Listen willingly to holy reading.
57. Devote yourself frequently to prayer.
58. Daily in your prayers, with tears and sighs, confess your past sins to God, and amend them for the future.
59. Fulfill not the desires of the flesh; hate your own will.
60. Obey in all things the commands of those whom God has placed in authority over you even though they (which God forbid) should act otherwise, mindful of the Lord's precept, "Do what they say, but not what they do."
61. Do not wish to be called holy before one is holy; but first to be holy, that you may be truly so called.
62. Fulfill God's commandments daily in your deeds.
63. Love chastity.
64. Hate no one.
65. Be not jealous, nor harbor envy.
66. Do not love quarreling.
67. Shun arrogance.
68. Respect your seniors.
69. Love your juniors.
70. Pray for your enemies in the love of Christ.
71. Make peace with your adversary before the sun sets.
72. Never despair of God's mercy.

Reference: https://sqlite.org/codeofethics.html

ACM code of Ethics

Preamble excerpts

The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct.

The Code is not an algorithm for solving ethical problems; rather it serves as a basis for ethical decision-making.

When thinking through a particular issue, a computing professional may find that multiple principles should be taken into account, and that different principles will have different relevance to the issue.

Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration.

The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders.

Open discussions about ethical issues promote this accountability and transparency.

ACM code of Ethics

Section header excerpts

1. GENERAL ETHICAL PRINCIPLES.
1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
1.2 Avoid harm.
1.3 Be honest and trustworthy.

2. PROFESSIONAL RESPONSIBILITIES.
2.1 Strive to achieve high quality in both the processes and products of professional work.
2.2 Maintain high standards of professional competence, conduct, and ethical practice.
2.3 Know and respect existing rules pertaining to professional work.

3. PROFESSIONAL LEADERSHIP PRINCIPLES.
3.1 Ensure that the public good is the central concern during all professional computing work.
3.2 Articulate, encourage acceptance of, and evaluate fulfillment of social responsibilities by members of the organization or group.
3.3 Manage personnel and resources to enhance the quality of working life.

4. COMPLIANCE WITH THE CODE.
4.1 Uphold, promote, and respect the principles of the Code.
4.2 Treat violations of the Code as inconsistent with membership in the ACM

Reference: https://www.acm.org/code-of-ethics

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

What?
Why?

Ethical Concerns in Data Science

Areas to consider:

- Privacy
- Data Ownership
- Bias
- Transparency
- Accountability

Privacy

Data privacy generally means the ability of a person to determine for themselves when, how, and to what extent personal information about them is shared with or communicated to others.

Reference: https://www.cloudflare.com/learning/privacy/what-is-data-privacy/

Privacy

Any of these look familiar?

General Data Protection Regulation (GDPR)

Family Educational Rights and Privacy Act (FERPA)

Health Insurance Portability and Accountability Act of 1996 (HIPAA)

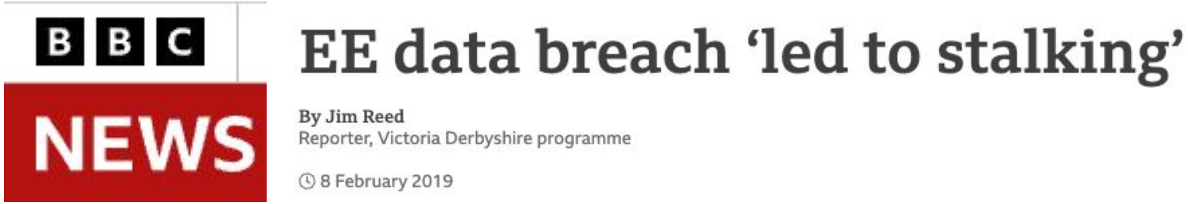
Children's Online Privacy Protection Rule (COPPA)

California Consumer Privacy Act (CCPA)

Health Information Technology for Economic and Clinical Health (HITECH) Act

Cases of Breach of Data Privacy

Privacy


BBC NEWS

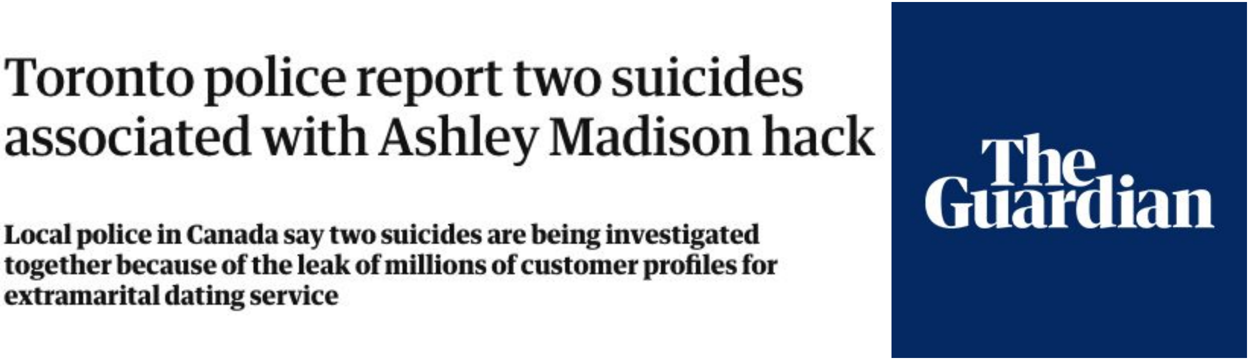
EE data breach ‘led to stalking’

By Jim Reed
Reporter, Victoria Derbyshire programme

8 February 2019

EE data breach

- The scenario: In 2019, **An EE customer has said she was stalked by an ex-partner who worked at the firm, after he accessed her personal data without permission.**


Toronto police report two suicides associated with Ashley Madison hack

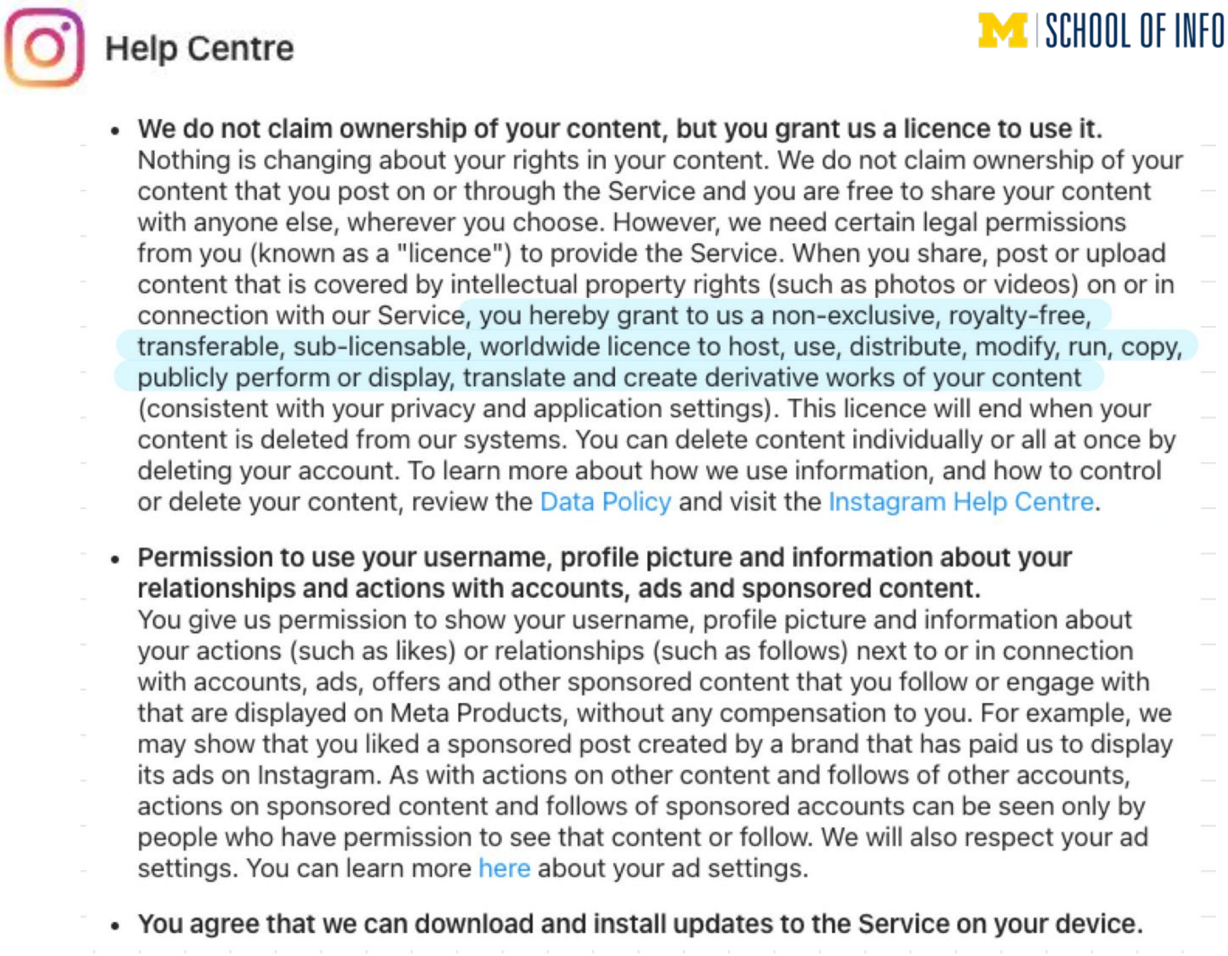
Local police in Canada say two suicides are being investigated together because of the leak of millions of customer profiles for extramarital dating service

The Guardian

Reference: https://www.bbc.com/news/technology-46896329

Do you really “own” your data on Social Media?

Data Ownership



Help Centre

- **We do not claim ownership of your content, but you grant us a licence to use it.** Nothing is changing about your rights in your content. We do not claim ownership of your content that you post on or through the Service and you are free to share your content with anyone else, wherever you choose. However, we need certain legal permissions from you (known as a "licence") to provide the Service. When you share, post or upload content that is covered by intellectual property rights (such as photos or videos) on or in connection with our Service, you hereby grant to us a non-exclusive, royalty-free, transferable, sub-licensable, worldwide licence to host, use, distribute, modify, run, copy, publicly perform or display, translate and create derivative works of your content (consistent with your privacy and application settings). This licence will end when your content is deleted from our systems. You can delete content individually or all at once by deleting your account. To learn more about how we use information, and how to control or delete your content, review the Data Policy and visit the Instagram Help Centre.
- **Permission to use your username, profile picture and information about your relationships and actions with accounts, ads and sponsored content.** You give us permission to show your username, profile picture and information about your actions (such as likes) or relationships (such as follows) next to or in connection with accounts, ads, offers and other sponsored content that you follow or engage with that are displayed on Meta Products, without any compensation to you. For example, we may show that you liked a sponsored post created by a brand that has paid us to display its ads on Instagram. As with actions on other content and follows of other accounts, actions on sponsored content and follows of sponsored accounts can be seen only by people who have permission to see that content or follow. We will also respect your ad settings. You can learn more here about your ad settings.
- **You agree that we can download and install updates to the Service on your device.**

Reference: https://help.instagram.com/581066165581870

Privacy

Actions to take

Steps to take

- Opt-in to sharing information, not opt out
- Get consent
- Anonymized collected data
- Limit the amount of time the data is stored
- Limit access to data (physically, technologically, and through organization access standards)

Bias

Inaccurate representation in the dataset, deviation from expectation in the data, error in the data…

NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT SURVIVORSHIP BIAS.

Image source: xkcd.com

What is Bias?

Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair.

Reference: https://en.wikipedia.org/wiki/Bias

What is Bias in Data Science?

Andrew Gelman said: “The most important aspect of a statistical analysis is not what you do with the data, it’s what data you use”

Data bias occurs due to structural characteristics of the systems that produce the data.

Reference: https://statmodeling.stat.columbia.edu/2018/08/07/important-aspect-statistical-analysis-not-data-data-use-survey-adjustment-edition/

Some Common Types of Data Bias

- Historical Bias
- Representation Bias
- Measurement Bias
- Evaluation Bias
- Aggregation Bias
- Population Bias

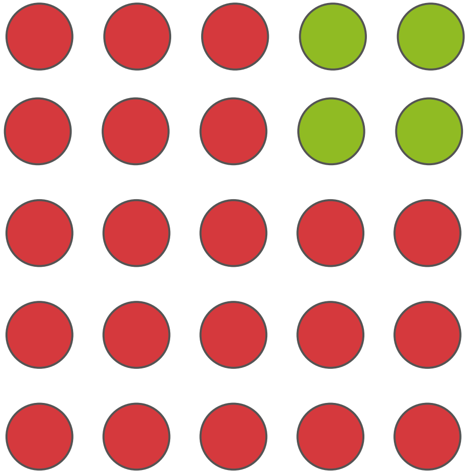
- Sampling Bias
- Observer Bias
- Social Bias
- Algorithmic Bias
- Confirmation Bias

Reference: https://arxiv.org/pdf/1908.09635.pdf

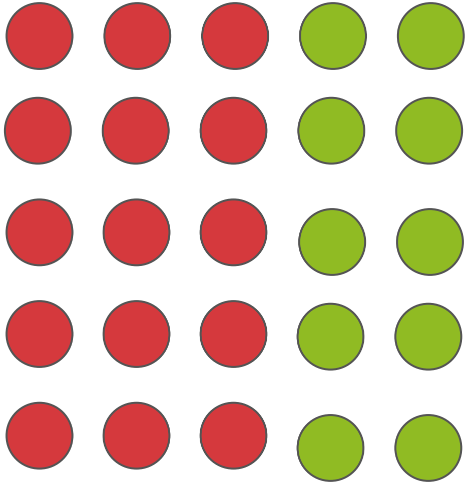
Bias

Representation Bias

Representation: 16% of circles are green

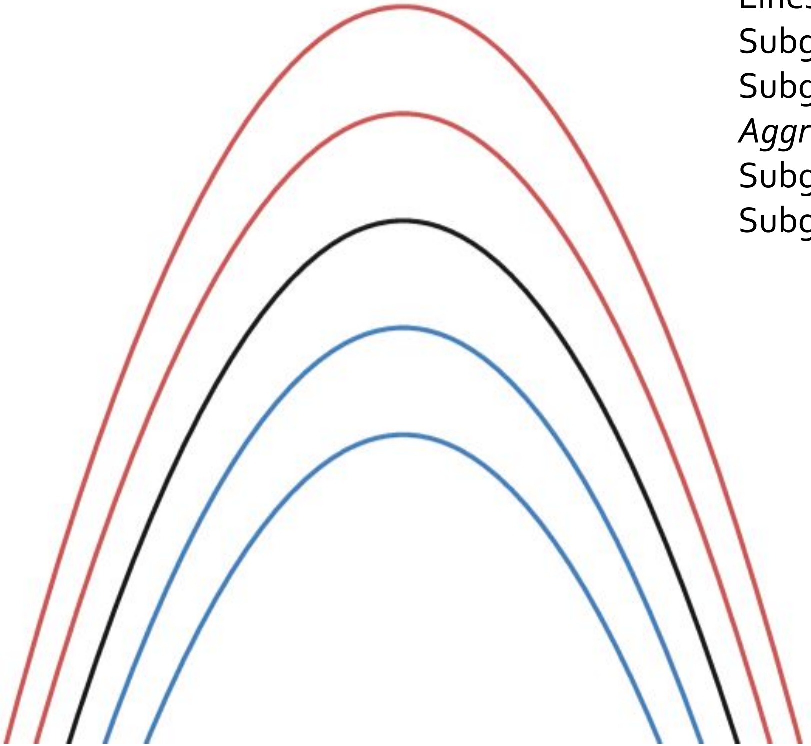


Actual: 40% of circles are green



Bias

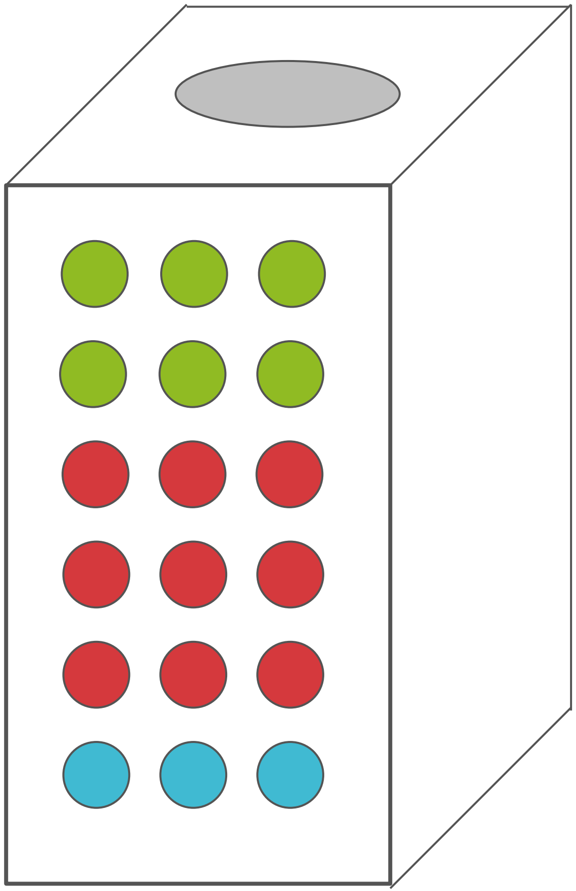
Aggregation Bias



Lines in order:
Subgroup A
Subgroup B
Aggregated
Subgroup C
Subgroup D

Bias

Sampling Bias



Confirmation bias

Social media algorithms take advantage of people's natural confirmation biases. By promoting and amplifying content that confirms what visitors already believe, social media platforms reinforce visitors' prior beliefs, keeping them engaged with the platform. Users see what they already believe, and leave feeling more convinced that their views are supported in reality.

Selection bias

A startup wants to know whether reducing the price of their product would result in higher overall revenues. They decide to test their new pricing for a week, but only test with visitors from the US. When they roll-out the pricing to the rest of the world, they're surprised to find that the broader audience behaves differently than their sample.

Historical bias

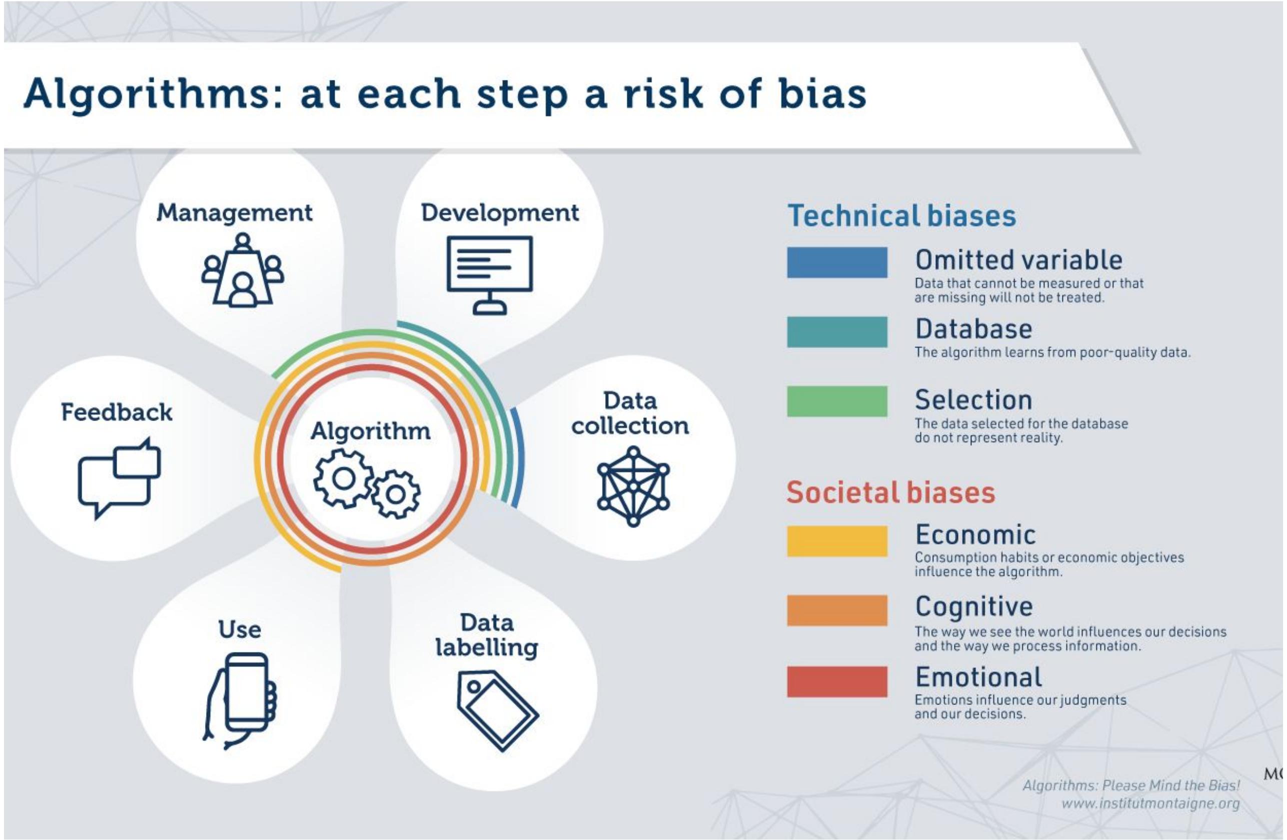
In 2013, neural network models transformed the way machines understand written words. This technology allows computers to encode the semantic meaning of words, by learning from giant sets of written text, like Wikipedia, Google News, or Reddit. However, we've seen several examples where text sourced from existing datasets has produced models that mirror and amplify the existing biases contained in those datasets. For example, a machine learning model trained on Wikipedia produced gender-biased analogies like: man : doctor :: woman : nurse, or man : commander :: woman : school teacher. The model inherited the historical biases of society by learning from the huge corpora of text, and produced work further reinforcing those biases.

Availability bias

A breakthrough new technology is taking the world by storm. You're seeing it on every billboard, news article, and hearing about it from your colleagues nonstop. When you encounter a problem that this technology could help you with, it's the first thing on your mind, and you jump right into implementing it on your new project. After a few weeks, your project isn't going as well as you'd hoped, and you realized that an older, more proven technology might have been a better solution. But because the breakthrough was easily available in your memory, you didn't fully investigate, and ended up needing to rethink your work.

Bias

Algorithmic Bias



Algorithms: at each step a risk of bias

Technical biases

- 🟦 **Omitted variable** Data that cannot be measured or that are missing will not be treated.
- 🟦 **Database** The algorithm learns from poor-quality data.
- 🟩 **Selection** The data selected for the database do not represent reality.

Societal biases

- 🟨 **Economic** Consumption habits or economic objectives influence the algorithm.
- 🟧 **Cognitive** The way we see the world influences our decisions and the way we process information.
- 🟥 **Emotional** Emotions influence our judgments and our decisions.

Algorithms: Please Mind the Bias!
www.institutmontaigne.org

MC

Reference: https://www.institutmontaigne.org/en/publications/algorithms-please-mind-bias

Case studies

Bias (race)

MIC

The Reason This "Racist Soap Dispenser" Doesn't Work on Black Skin

Case: An African-American guest of the Dragon Con sci-fi and fantasy convention visited a bathroom in the event's host hotel and discovered the soap dispenser, from a British company called Technical Concepts, wouldn't sense his hands. When his friend, a white man named Larry, tried after him, out came the soap.

Issue: Racial bias emerging from how testing was done while technology was in development, representative sampling issues

The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Reference: https://www.mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin#.XeuPqZmzH and https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Case study

Bias (gender)

REUTERS

Amazon scraps secret AI recruiting tool that showed bias against women

2018

Case: Amazon recruiting tool for automated resume reviews showed bias against women

Intended result: Input resumes, tool rates them all on a 1-5 scale and narrows down which ones were the best for hiring

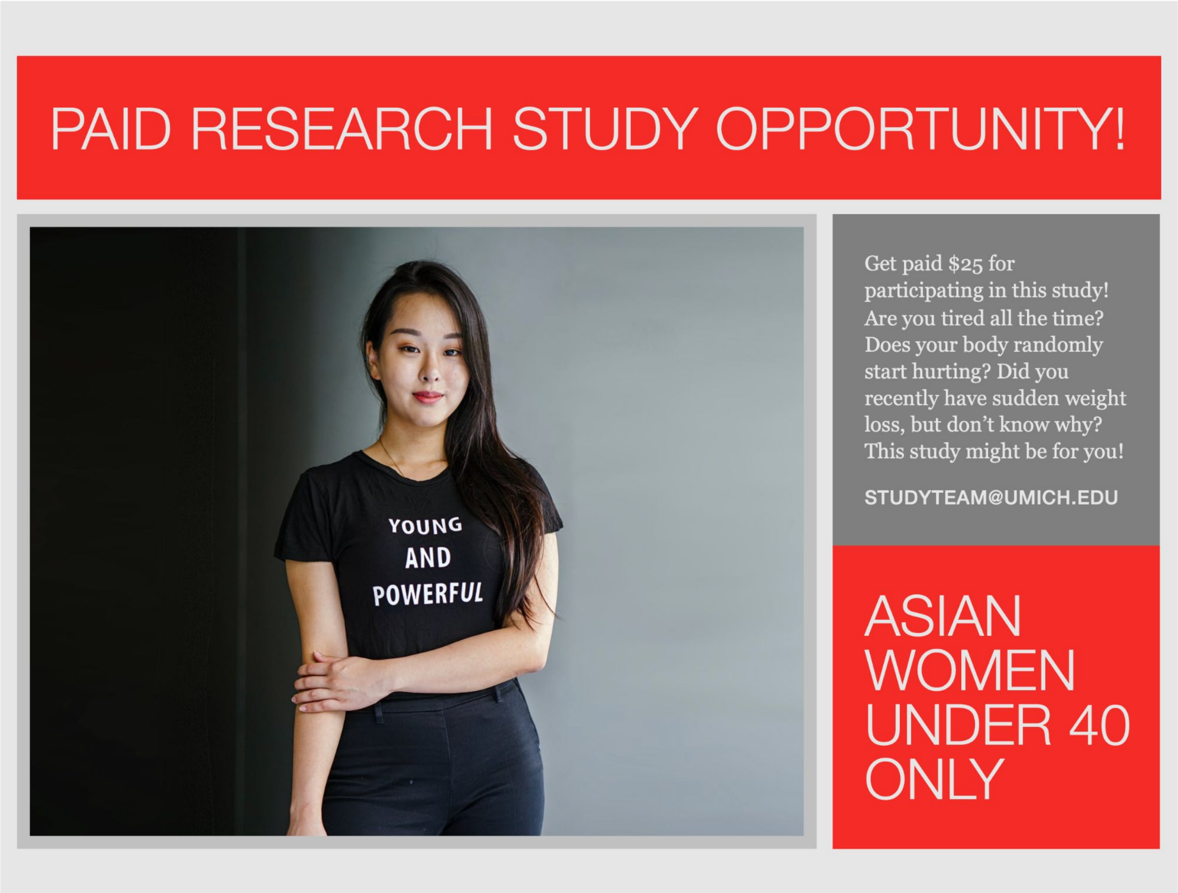
Actual result:

- Amazon's system taught itself that male candidates were preferable.
- Penalized resumes that included the word "women's," as in "women's chess club captain."
- Downgraded graduates of two all-women's colleges, according to people familiar with the matter.
- Bias of AI and gender

Reasons: Machine learning trained on historic data (resumes), which mainly came from men.

Reference: Amazon scraps secret AI recruiting tool that showed bias against women

Case Study



PAID RESEARCH STUDY OPPORTUNITY!

YOUNG
AND
POWERFUL

Get paid $25 for participating in this study! Are you tired all the time? Does your body randomly start hurting? Did you recently have sudden weight loss, but don't know why? This study might be for you!

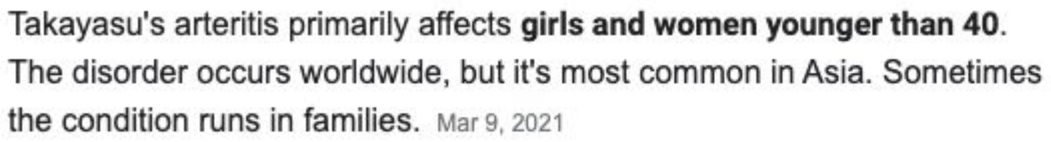
STUDYTEAM@UMICH.EDU

ASIAN WOMEN UNDER 40 ONLY

Not Biased!

Context is important

What is I told you this research was for interventions of Takayasu's Arteritis?



Takayasu's arteritis primarily affects **girls and women younger than 40**. The disorder occurs worldwide, but it's most common in Asia. Sometimes the condition runs in families. Mar 9, 2021

Reference: https://www.mayoclinic.org/diseases-conditions/takayasus-arteritis/symptoms-causes/syc-20351335#

Questions to think about

These are some questions to think about when it comes to Data Ethics.

- What is Data Science Ethics?
- Is Privacy-Respecting Data Science Even Possible?
- Is it Inherently Discriminatory?
- Can large, automated systems be effectively controlled?
- Does data science always leave something out?
- When are YOU responsible?

The way data scientists build models can have real implications for justice, health, and opportunity in people's lives. And we have an obligation to consider the ethics of our discipline each and every day. **When built correctly, algorithms can have massive power to do good in the world**.

Data Ethics Checklist

Based on authors of the first reading of this week: Patil, Mason and Loukides

https://www.oreilly.com/radar/of-oaths-and-checklists/

❏ Have we listed how this technology can be attacked or abused? [SECURITY]

❏ Have we tested our training data to ensure it is fair and representative? [FAIRNESS]

❏ Have we studied and understood possible sources of bias in our data? [FAIRNESS]

❏ Does our team reflect diversity of opinions, backgrounds, and kinds of thought? [FAIRNESS]

❏ What kind of user consent do we need to collect to use the data? [PRIVACY/TRANSPARENCY]

❏ Do we have a mechanism for gathering consent from users? [TRANSPARENCY]

❏ Have we explained clearly what users are consenting to? [TRANSPARENCY]

❏ Do we have a mechanism for redress if people are harmed by the results? [TRANSPARENCY]

❏ Can we shut down this software in production if it is behaving badly?

❏ Have we tested for fairness with respect to different user groups? [FAIRNESS]

❏ Have we tested for disparate error rates among different user groups? [FAIRNESS]

❏ Do we test and monitor for model drift to ensure our software remains fair over time? [FAIRNESS]

❏ Do we have a plan to protect and secure user data? [SECURITY] ethicalds

Communication

Communication Can be of Different Types

Technical

- Other team members
- Tech Manager
- Collaboration with other teams

Non-Technical

- General Manager
- C-suite
- General Public

Visual/Written

- Technical Reports
- Articles
- Flyers

Verbal

- Oral reports
- Stand-Up meetings
- Presentations

Questions to ask before formulating your “communication”

- What is the scope of this project?
- Who are the stakeholders? Who’s your audience?
- What are the overall requirements?
- What are the objectives of this piece? How does it fit into the overall goals?

Technical Reports

Technical Reports

What is it?

- Professional writing used for reporting and explaining your data analysis project
- Typical structure
  - Introduction: What is this about, why are you doing this
  - Methodology: How you did what you did
  - Results: What did you find out from what you did
  - Recommendations: Based on what you found out, what should your client do and why?
  - Conclusions: Summarize everything
- Not everyone in your audience will read the full report

Technical Reports

Key points

- On topic
- Clear
- Consistent
- Specific
- Avoid first person
- passive vs active voice
- Appropriate use of tense
  - Present tense for the Objective, Background, Results and Conclusions section and any time you state general rules or truths: “The relationship between uniaxial stress and strain is $\sigma = \varepsilon E$.”
  - Past tense for the Procedure/Experiment section. Tell what was done and what happened in your particular case.

Technical Reports

Example scenario

- You received a request to analyze how often various softwares were used at UM campus computing sites.
- Your client, UM ITS, wants to know which kinds of softwares they should focus on providing.
- You write a technical report and also give a presentation on your findings to the purchasing department.
- Data looks like below

Software name	type	computer #	date used	total time used
Adobe Illustrator	illustration	4RTD3S	2021.11.20	5.5
Zoom	communication	MTTNGS	2021.11.21	2
R	analysis	AAAAAA	2021.11.23	4

Technical Reports

Introduction

Bad: This is about the most popular softwares used at the U. I got the data from the SQL database ITS gave me. ITS wants to know what to focus on keeping and what to get rid of.

Good: This report summarizes the types of software with the highest used hours at the University of Michigan.

Technical Reports

Methodology

Bad: We like Jupyter notebooks and python here at UMSI, so that's what I used. Everything got loaded into a pandas dataframe for data exploration. It was all daily logs. Most of the variables was about computer specs, which was all the same anyways. I added up all the times spent.

Good: The data was loaded from a SQL database managed by ITS Procurement into a Jupyter notebook. The data consisted of 100 rows and 5 columns, being Name, type, computer #, date used, and total time used on that date. The groupby function was used to find the sum of total time used by software type.

Technical Reports

Results

Bad: There was a month's worth of data, so counting up all the hours from the rows, people here used communication software a lot, more than the other types, which were illustration, communication, analysis, presentation, and development. Hours spent was 225, 100, 75, 60, and 40.

Good: The exploratory analysis revealed that there were 5 types of softwares offered at U-M. These types were illustration, communication, analysis, presentation, and development. The dataset covered 30 days. The total hours of software usage was 500. Usage was measured in percentage of the total hours of software use. The percentages were:

- communication software at 45%
- analysis software at 20%
- development software at 15%
- presentation software at 12%
- illustration software at 8%

Technical Reports

Recommendations

Bad: Zoom's the best, we need to make sure it never has an outage or else we're all doomed. Only the UX people use the photoshop and we also have a lot of photoshop alternatives so we should get rid of all the extras.

Good: Based on the results of the exploratory data analysis, communication software is highly used at U-M. ITS should focus on ensuring this type of software is available. Illustration software was the smallest percentage of hours used. As such, ITS can further investigate if there are specific softwares within this category where it may be appropriate to discontinue enterprise service.

Technical Reports

Conclusion

Bad: Funds should be directed towards communication software.

Good: Exploratory analysis of 30 days worth of software usage data shows that communication software is the most heavily used type of software at U-M.

Technical Reports

Example scenario

- You received a request to analyze how often various softwares were used at UM campus computing sites.
- Your client, UM ITS, wants to know which kinds of softwares they should focus on providing.
- You write a technical report and also give a presentation on your findings to the purchasing department.
- Data looks like below

Software name	type	computer #	date used	total time used
Adobe Illustrator	illustration	4RTD3S	2021.11.20	5.5
Zoom	communication	MTTNGS	2021.11.21	2
R	analysis	AAAAAA	2021.11.23	4

Technical Reports

Key points

- On topic
- Clear
- Consistent
- Specific
- Avoid first person
- passive vs active voice
- Appropriate use of tense
    - Present tense for the Objective, Background, Results and Conclusions section and any time you state general rules or truths: “The relationship between uniaxial stress and strain is $\sigma = \varepsilon E$.”
    - Past tense for the Procedure/Experiment section. Tell what was done and what happened in your particular case.

Visualizations

Visualization

A good data visualization **tells a story to the audience**, usually in images, graphs, or charts using language and ideas that they understand. Good data visualizations, as the name implies, have **good data**, it has a **good choice** of data visualization, the **color or information are simple and explicit**, the data are **accurately represented**, and there is **consistency in scales**.

Visualization

What makes them good?

- Data is accurate
- Format is appropriate for what you want to convey
- Color choice and usage follows common expectations
- Data is represented accurately
- Visualization has consistency

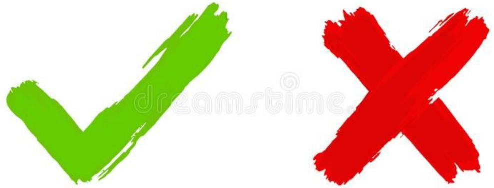
Pitfalls of Data Visualisation

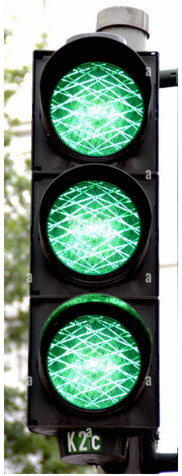
- Colour Abuse
- Bad Choice of Type of Chart
- Visual Clutter
- Scale Inconsistencies
- Bad Data

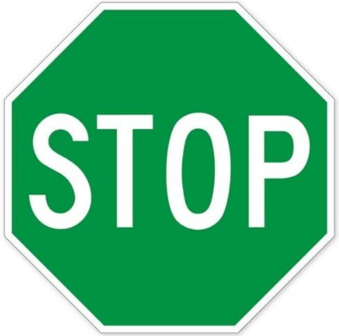
#1 Colour Abuse

Colour is one of the essential components in helping distinguish between different data points. But it is essential to remember not to overdo it.

The wrong choice of colour can lead to confusion, or even worse, misinterpretation.







Colour

Some common issues that arise when incorporating color into your visualizations include:

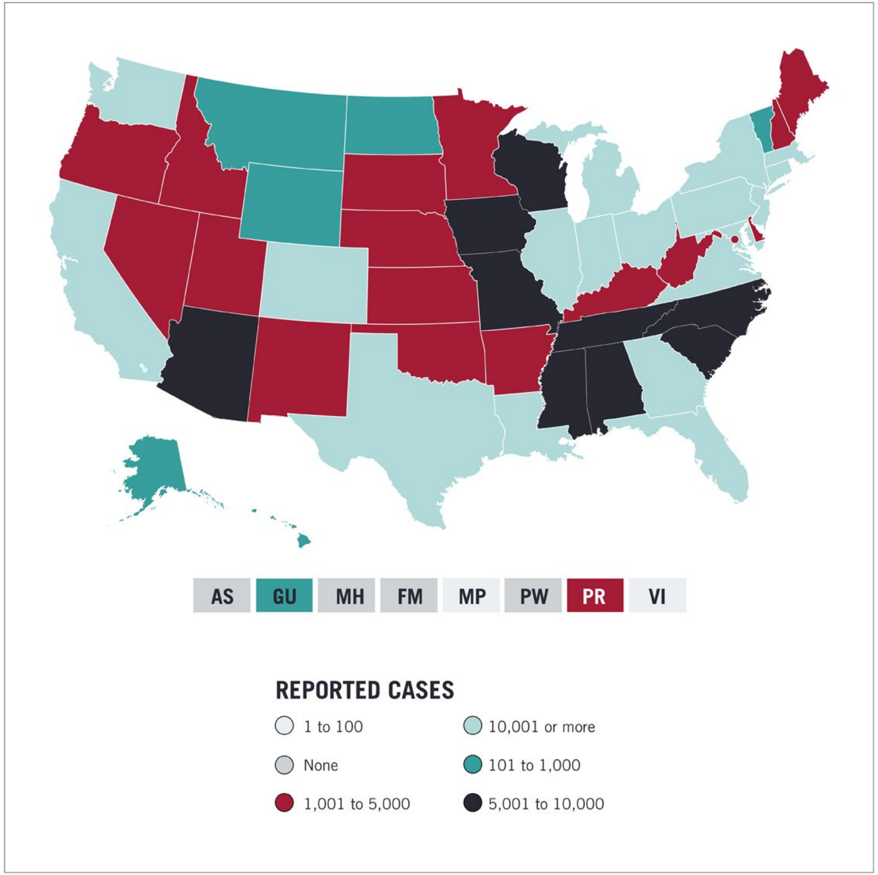
Using too many colors, making it difficult for the reader to quickly understand what they're looking at.

Using familiar colors (for example, red and green) in surprising ways.

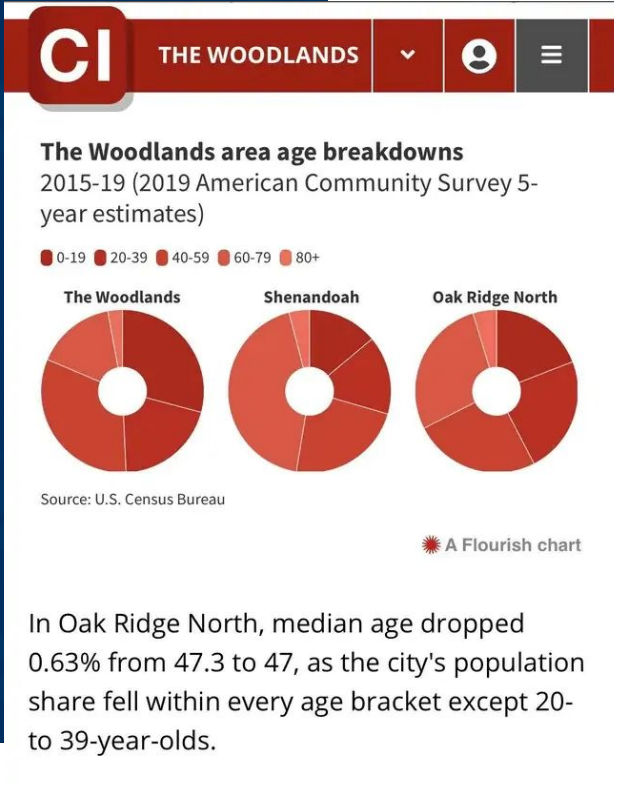
Using colors with little contrast.

Not accounting for viewers who may be colorblind

For example, blue/orange is a common colorblind-friendly palette. Blue/red or blue/brown would also work.


AS
GU
MH
FM
MP
PW
PR
VI
REPORTED CASES
○ 1 to 100
○ 10,001 or more
○ None
● 101 to 1,000
● 1,001 to 5,000
● 5,001 to 10,000

Bad vs Good



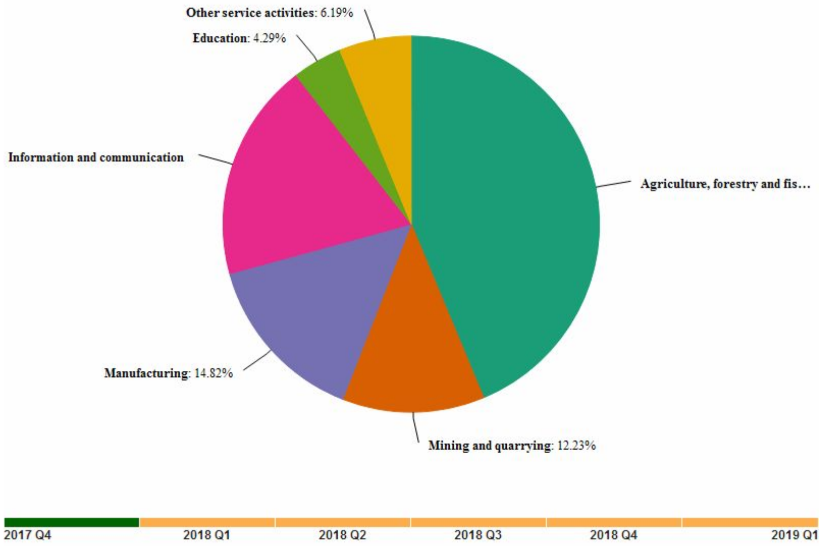
The Woodlands area age breakdowns

2015-19 (2019 American Community Survey 5-year estimates)

Source: U.S. Census Bureau

A Flourish chart

In Oak Ridge North, median age dropped 0.63% from 47.3 to 47, as the city's population share fell within every age bracket except 20- to 39-year-olds.

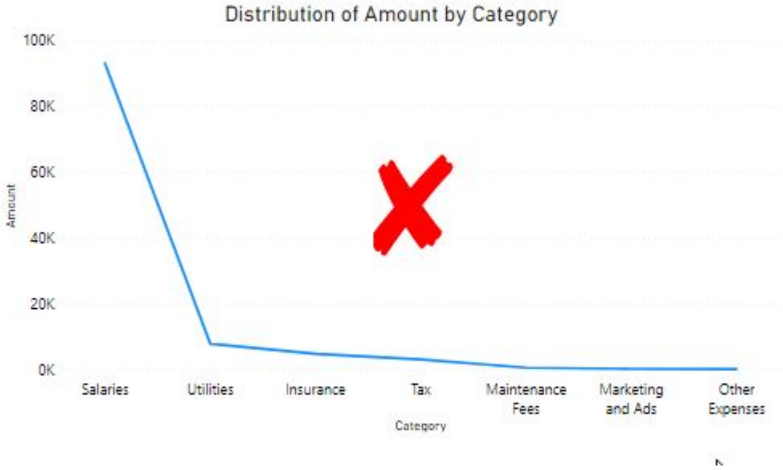


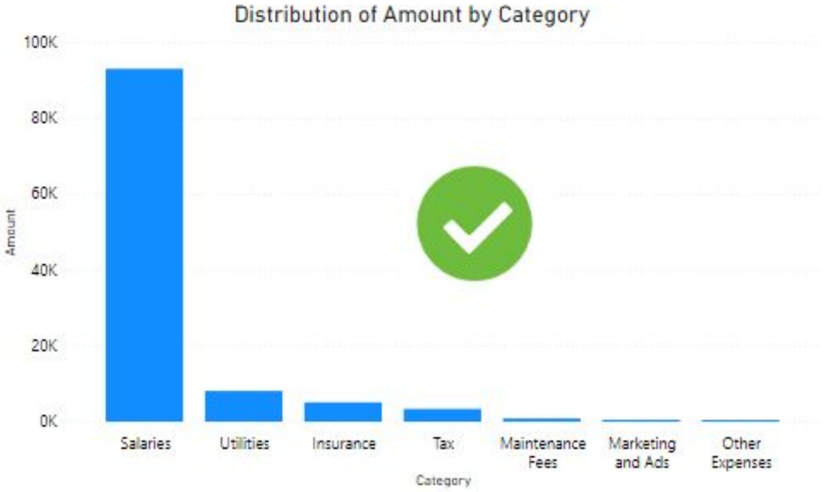
#2 Inappropriate type of chart

Some graphs and charts work well for communicating specific types of information, but not others. Problems can arise when you try visualizing data using an unsuitable format.

The nature of your data usually dictates the format of your visualization. The most important characteristic is whether the data is qualitative (it describes or categorizes) or quantitative (meaning, it's measurable). Qualitative data tends to be better suited to bar graphs and pie charts, while quantitative data is best represented in formats like charts and histograms.

Bad v/s Good


Distribution of Amount by Category
100K
80K
60K
40K
20K
0K
Amount
Salaries
Utilities
Insurance
Tax
Maintenance Fees
Marketing and Ads
Other Expenses
Category


Distribution of Amount by Category
100K
80K
60K
40K
20K
0K
Amount
Salaries
Utilities
Insurance
Tax
Maintenance Fees
Marketing and Ads
Other Expenses
Category

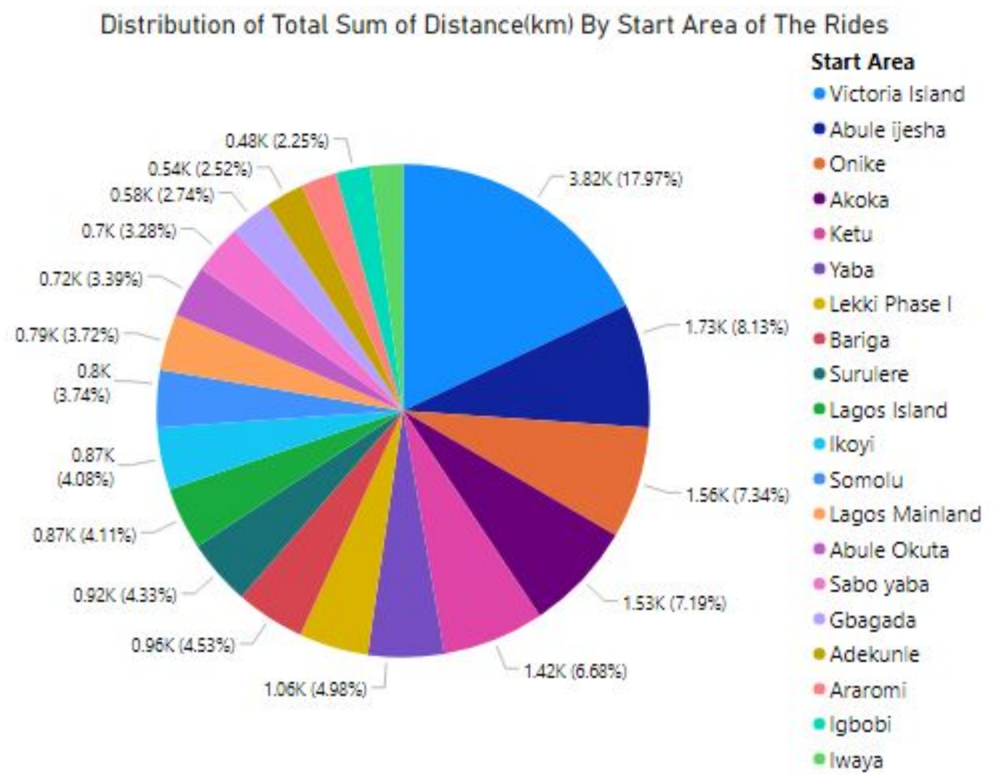
#3 VISUAL CLUTTER

The point of generating a data visualisation is to tell a story. As such, it’s your job to include as much relevant information as possible—while excluding irrelevant or unnecessary details. Doing so ensures your audience pays attention to the most important data. Too much information defeats the purpose of clarity.

For this reason, in conceptualising your data visualisation, you should first seek to identify the necessary variables. The number of variables you select will then inform your visualisation’s format. Ask yourself: Which format will help communicate the data in the clearest manner possible?

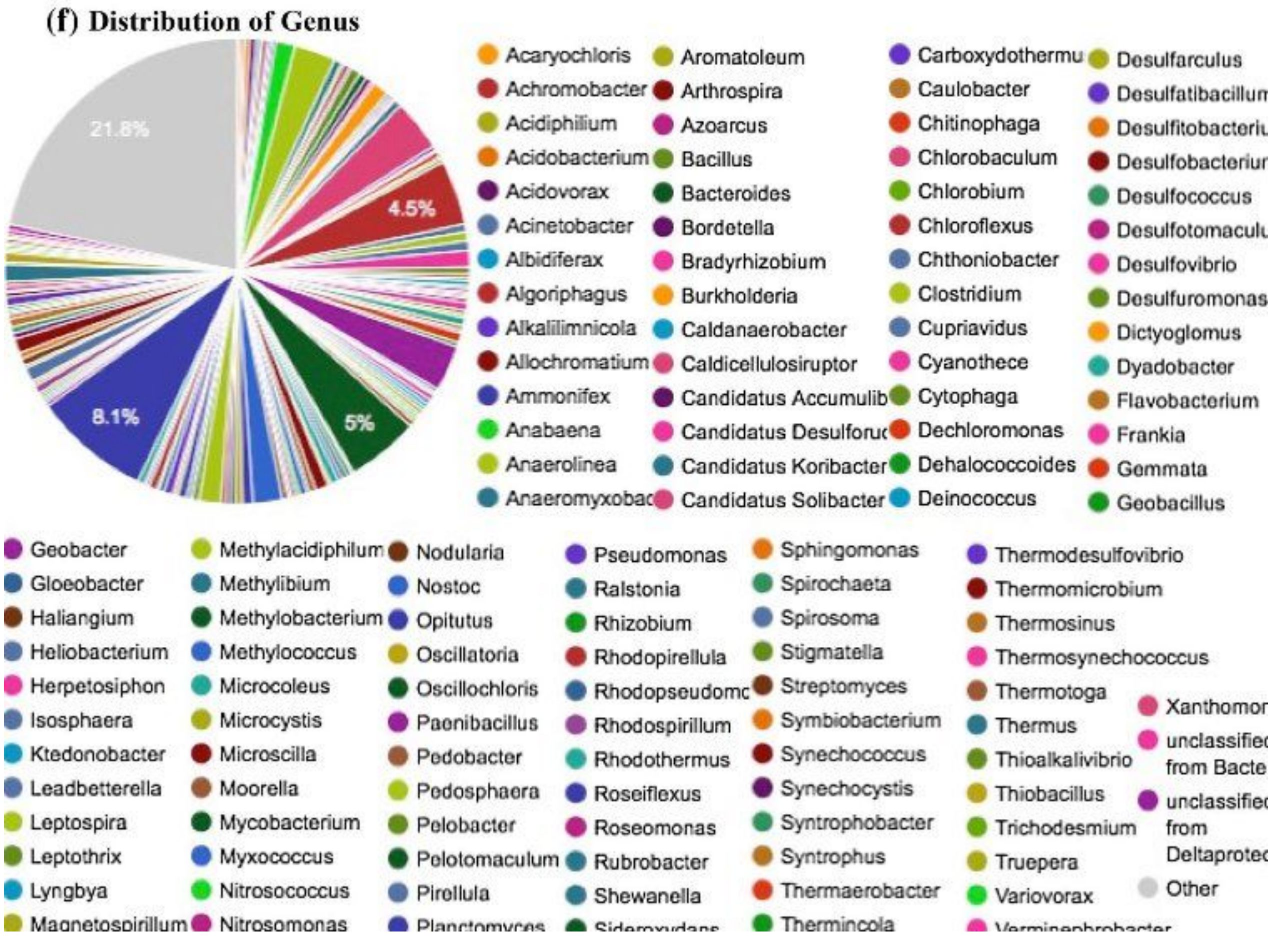
Visualization

Clutter



Visualization

Clutter

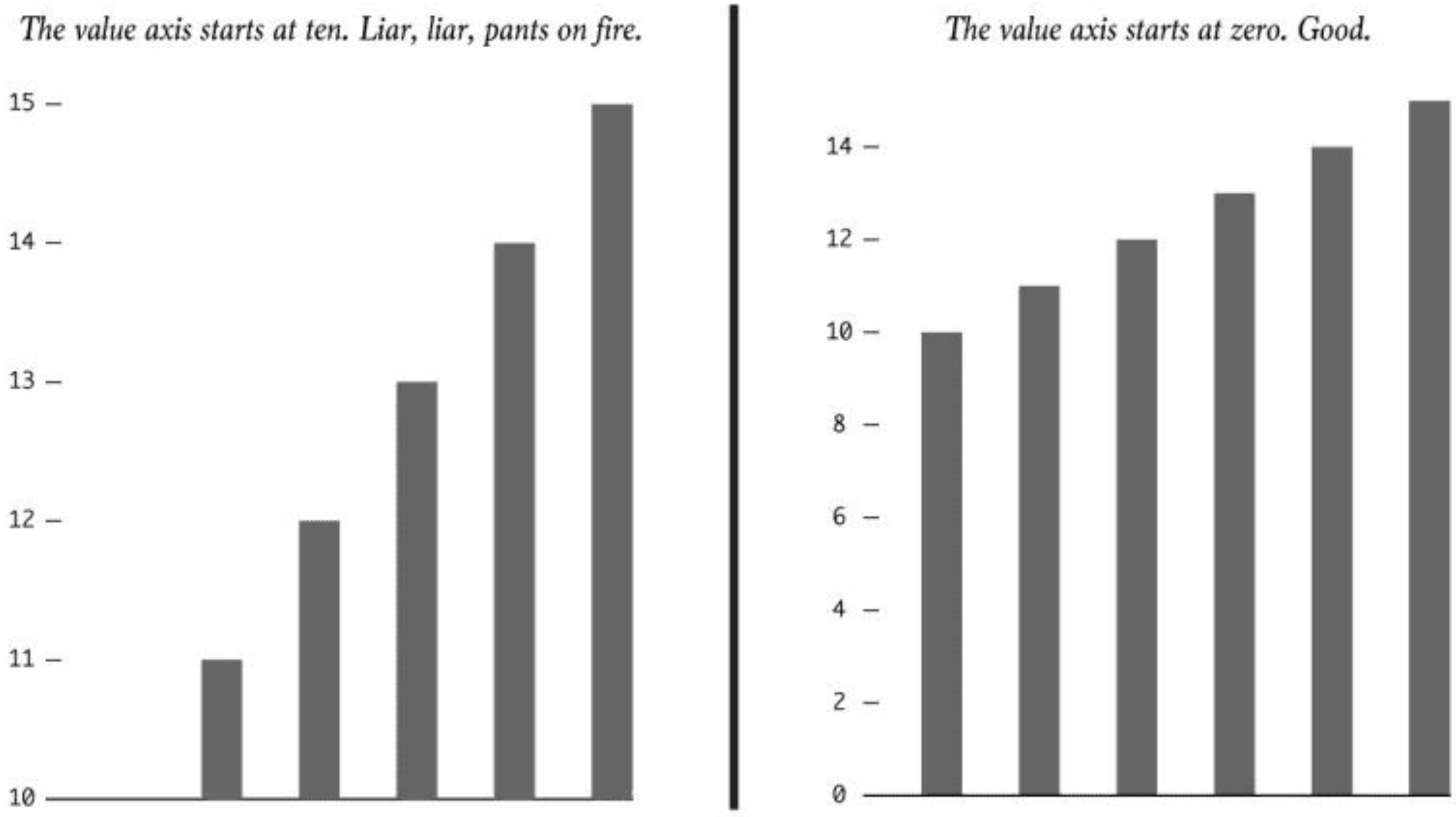


(f) Distribution of Genus

#4 Inconsistencies in Scale

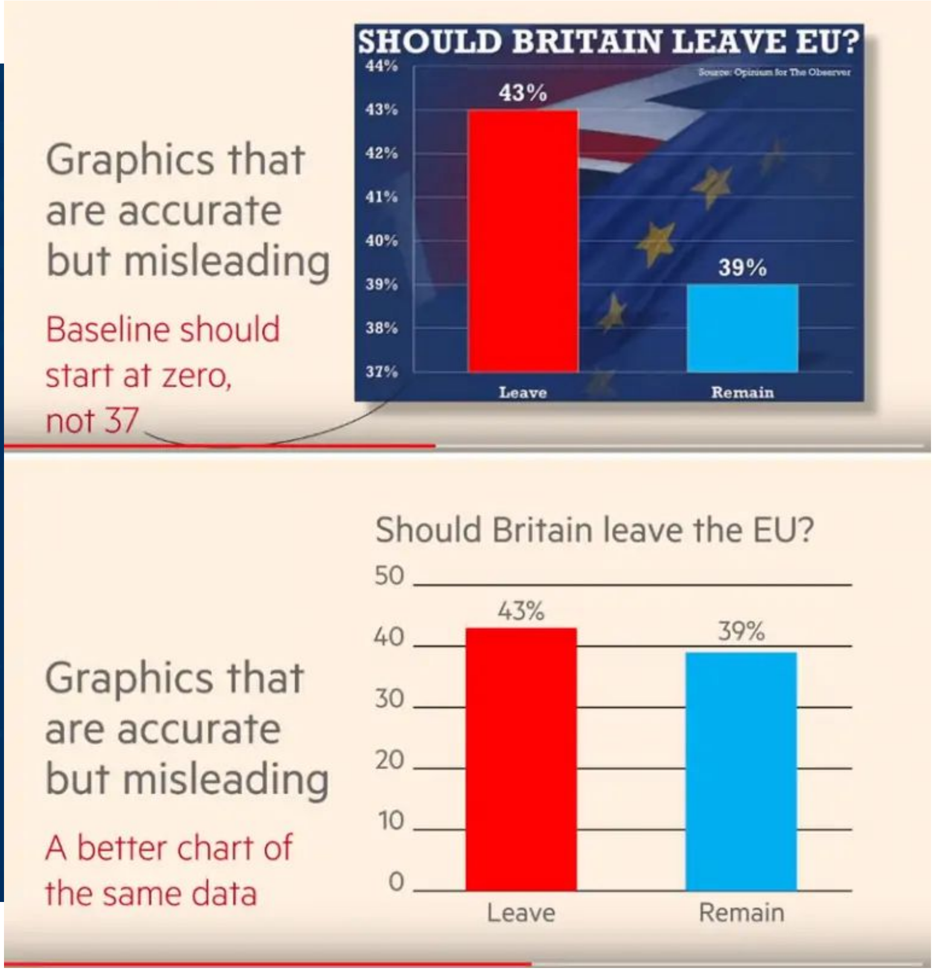
If your chart or graph is meant to show the difference between data points, your scale must remain consistent. If your visualisation's scale is inconsistent, it can cause significant confusion for the viewer.

TRUNCATED AXIS



The value axis starts at ten. Liar, liar, pants on fire.

The value axis starts at zero. Good.



Graphics that are accurate but misleading

Baseline should start at zero, not 37

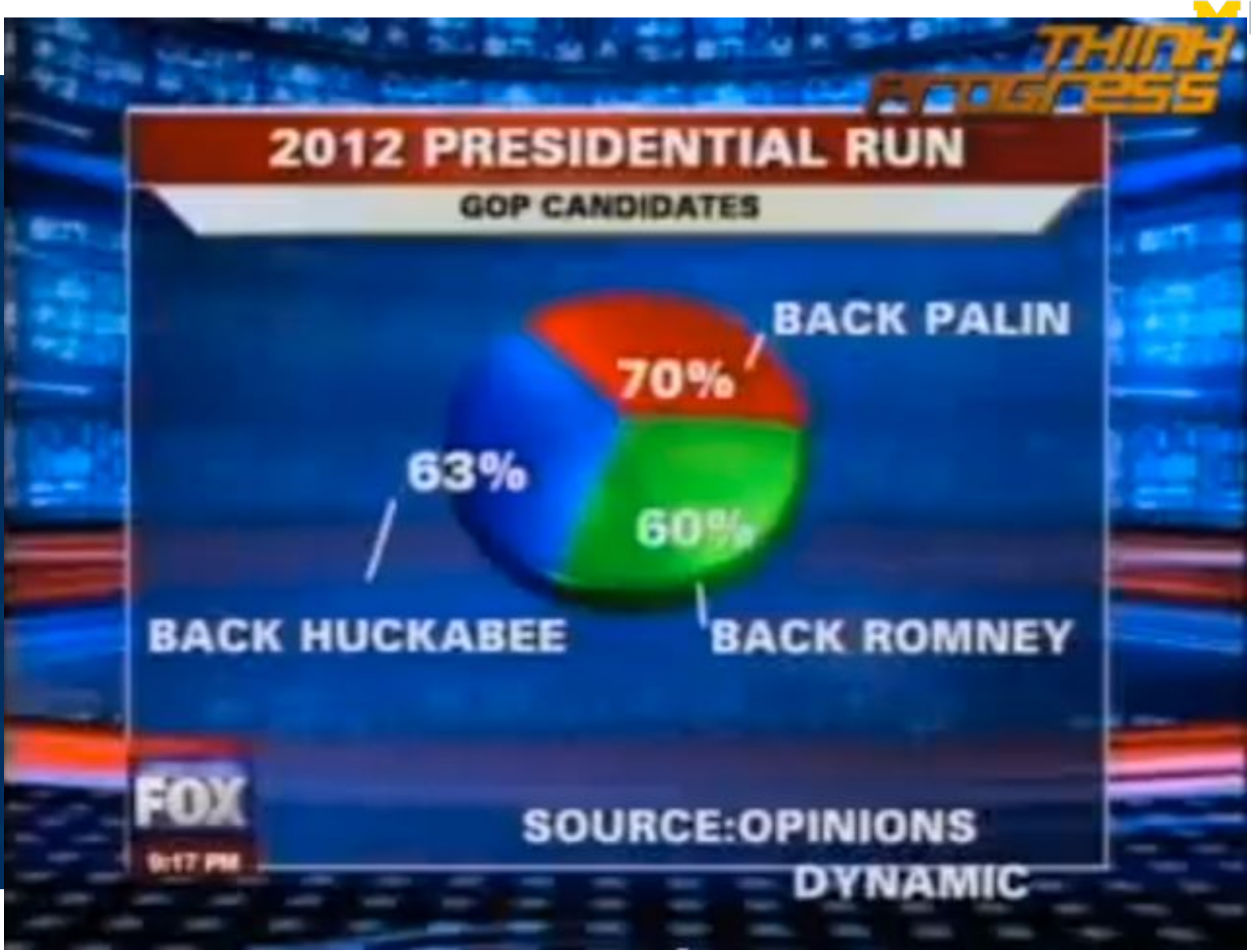
Graphics that are accurate but misleading

A better chart of the same data

Should Britain leave the EU?

#5 Bad Data

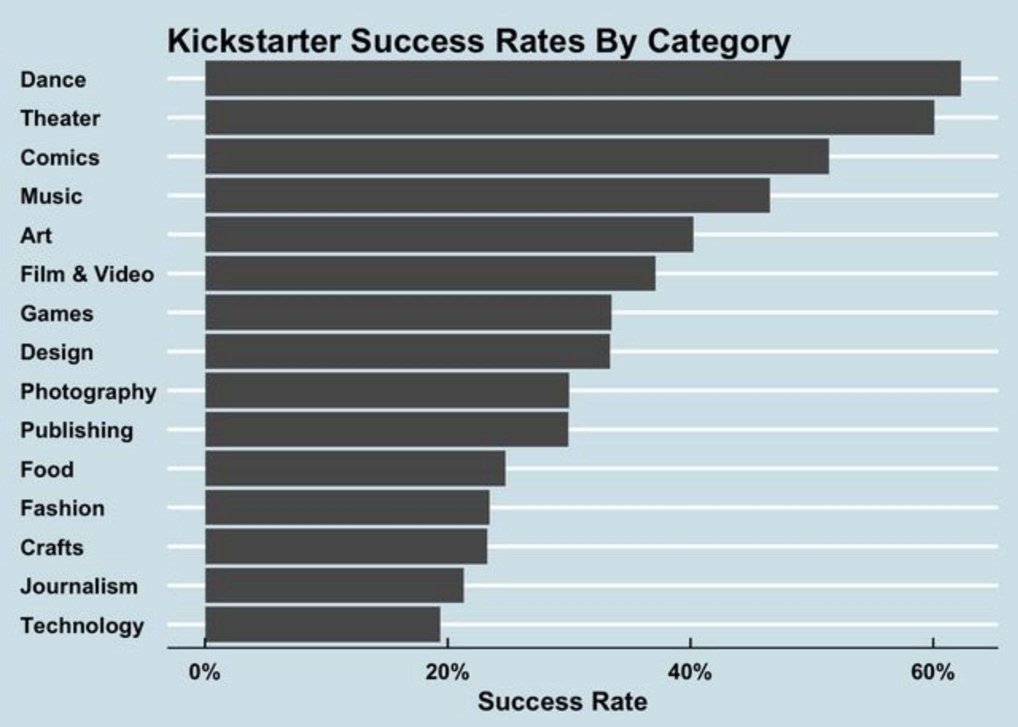
Good visualisations start with good data; Bad data will lead to bad visualizations. Start with the basics: is your data clean? Use checks at every stage the data goes through — collection, sourcing, cleaning, and compiling — before it is visualized. Common errors include data duplication, missed data, NA values not marked, and so on.


THINK PROGRESS
2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
9:17 PM
SOURCE:OPINIONS
DYNAMIC

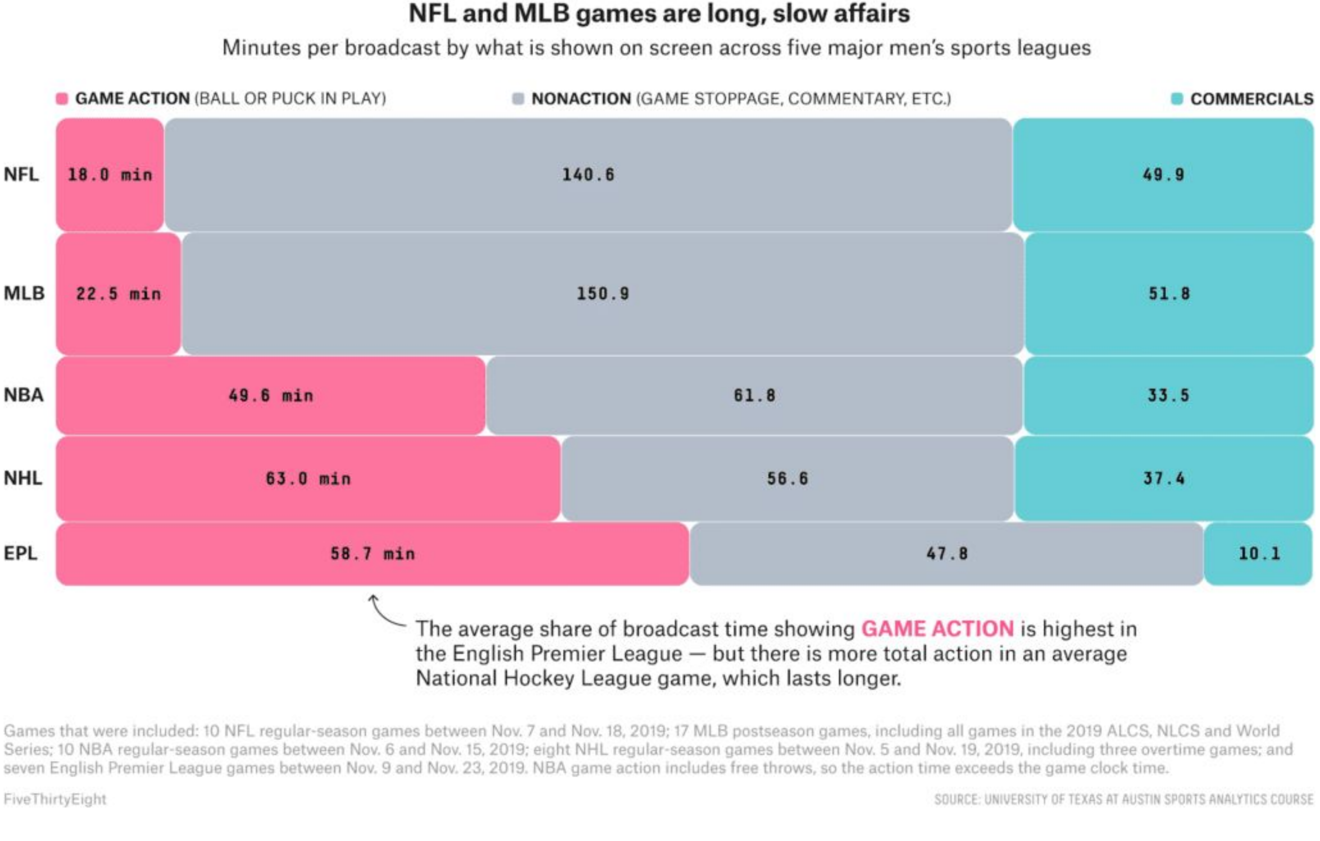
THE IMPACT OF POOR DATA VISUALIZATION

Good data Visualisations

Good visualizations



Good visualizations



NFL and MLB games are long, slow affairs

Minutes per broadcast by what is shown on screen across five major men's sports leagues

The average share of broadcast time showing **GAME ACTION** is highest in the English Premier League — but there is more total action in an average National Hockey League game, which lasts longer.

Games that were included: 10 NFL regular-season games between Nov. 7 and Nov. 18, 2019; 17 MLB postseason games, including all games in the 2019 ALCS, NLCS and World Series; 10 NBA regular-season games between Nov. 6 and Nov. 15, 2019; eight NHL regular-season games between Nov. 5 and Nov. 19, 2019, including three overtime games; and seven English Premier League games between Nov. 9 and Nov. 23, 2019. NBA game action includes free throws, so the action time exceeds the game clock time.

FiveThirtyEight

SOURCE: UNIVERSITY OF TEXAS AT AUSTIN SPORTS ANALYTICS COURSE

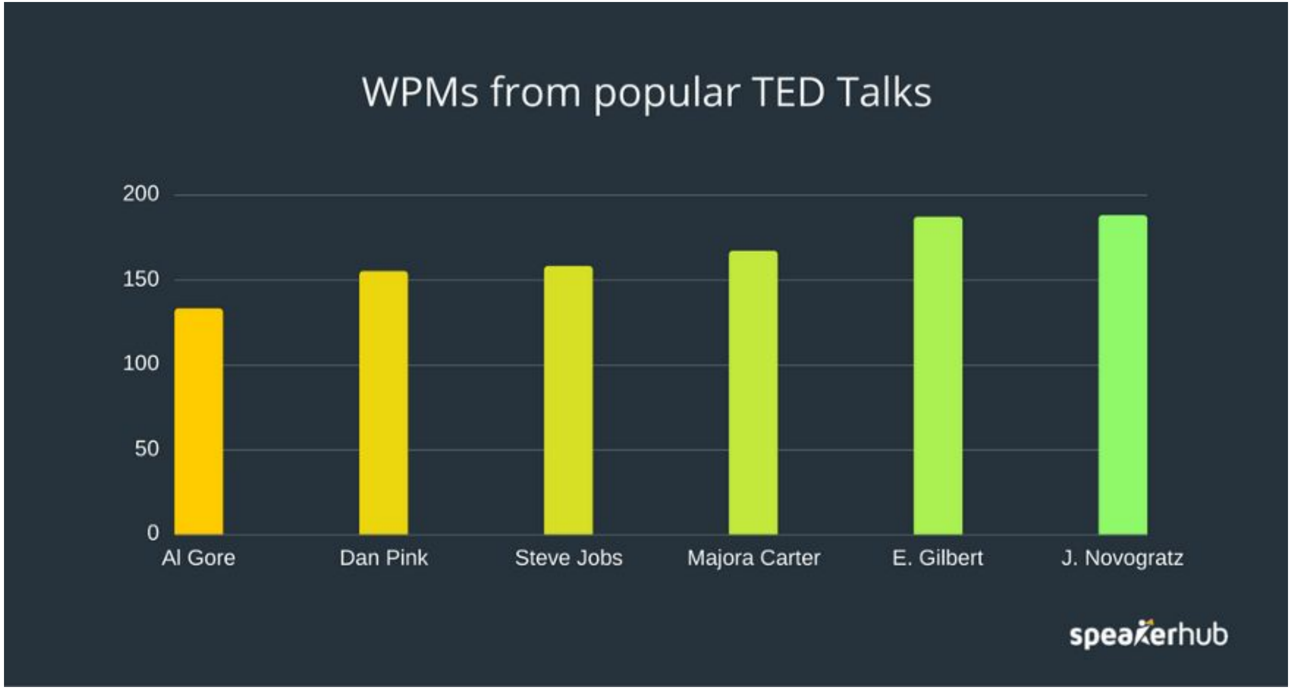
Presenting your findings

What is the single thing that makes presentations go well.

What holds _you_ back from a good presentation?

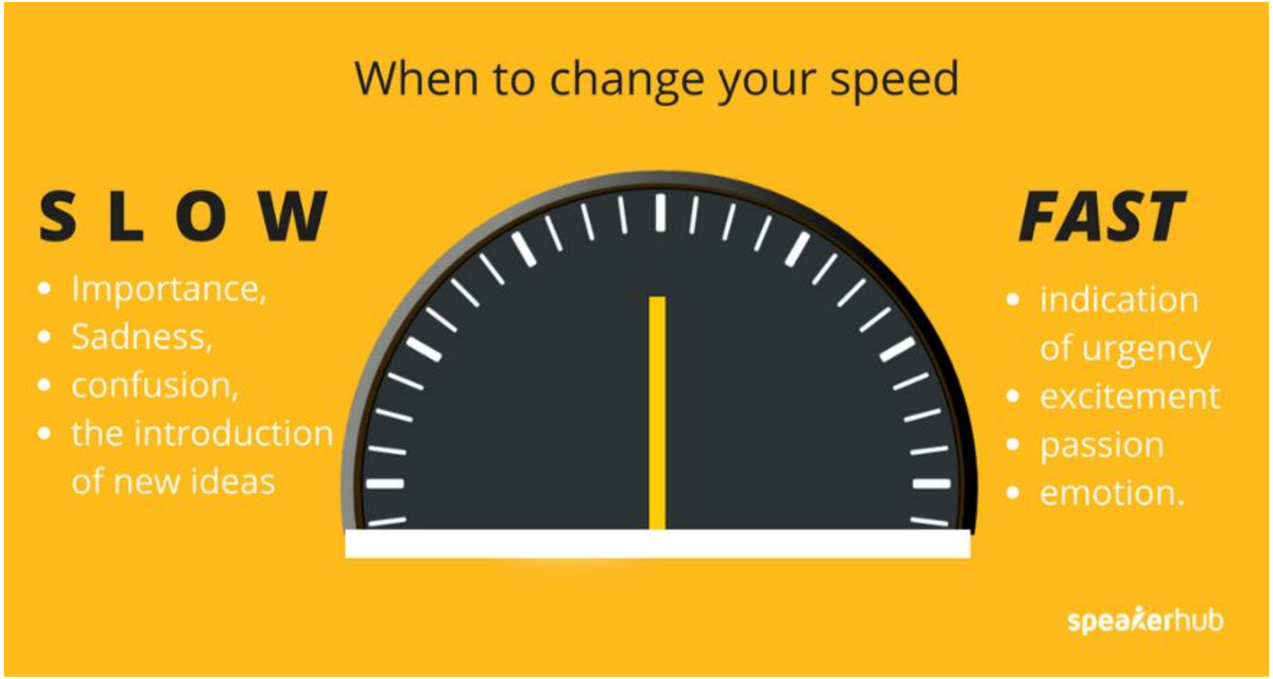
Presenting

Speed



Presenting

Speed



When to change your speed

SLOW

- Importance,
- Sadness,
- confusion,
- the introduction of new ideas

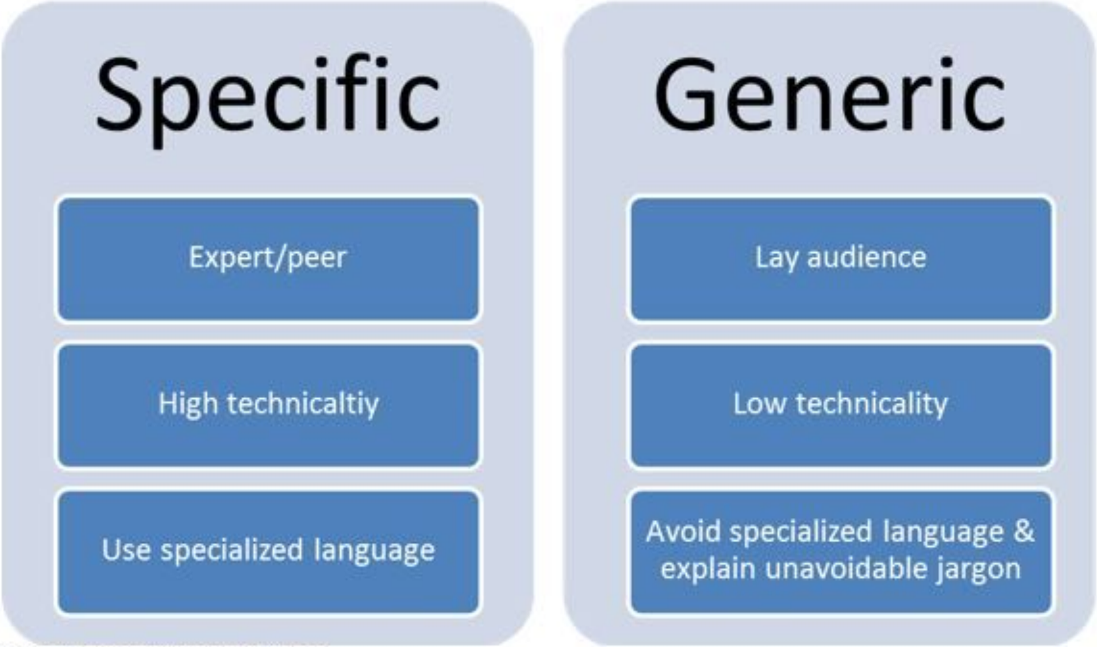
FAST

- indication of urgency
- excitement
- passion
- emotion.

speakerhub

Presenting

Listener's background



Specific

Expert/peer

High technicaltiy

Use specialized language

Generic

Lay audience

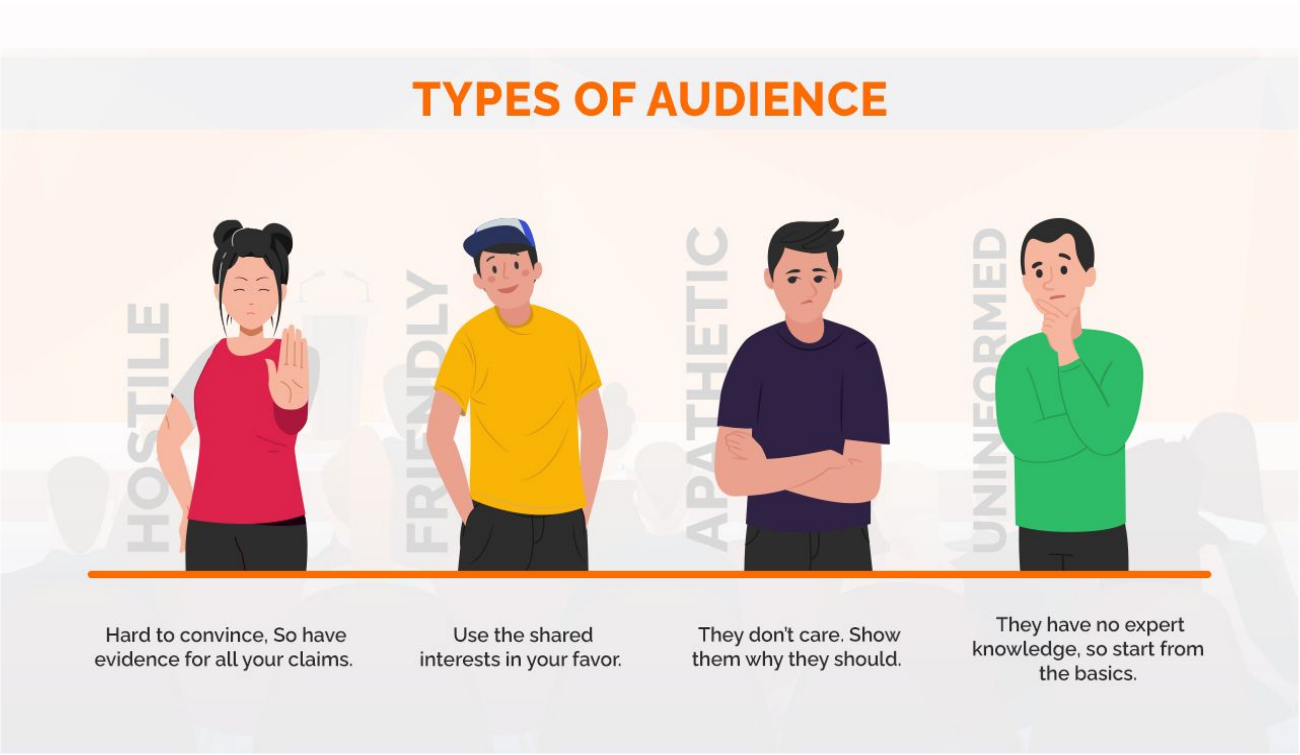
Low technicality

Avoid specialized language & explain unavoidable jargon

Figure 1 Difference between audiences

Presenting

Listener's background



TYPES OF AUDIENCE

Hard to convince, So have evidence for all your claims.

Use the shared interests in your favor.

They don't care. Show them why they should.

They have no expert knowledge, so start from the basics.

Make a good intro

Breathe

You are the expert. You are the expert.

Come Early

You may have questions you don’t know the answer to.

Practice, but not too much

Your body language is important.

You are on “stage” .

Use your slides, not your notes

Assume people can read.
(unless they can’t)

Interaction of some sorts.

Thank people

When you can, tell a story.

References

Links

- https://towardsdatascience.com/what-is-data-science-and-what-is-it-not-c6a09d735f02
- https://www.pewresearch.org/politics/2021/09/30/charting-congress-on-social-media-in-the-2016-and-2020-elections/
- https://builtin.com/data-science/data-science-applications-examples