# UTSA CS 4593: CS-CURE

## Course-based Undergraduate Research Experience in CS

**Amanda Fernandez, Ph.D**
UTSA Department of Computer Science

**Spring 2024**

# Week 7: **Data Analysis**

# UTSA CS-CURE

Week 7

**Reminder about Mid Term Grades:**

Mid semester grades are posted to ASAP for undergraduates.

You must contact your academic advisor if you are below a C-

- Objectives:
    - Identify the data & sources useful to a research initiative
    - Understand ethical considerations for research execution & analysis

- Deliverables:
    - Activity 5: Data Analysis & Visualization (in-class on Thursday)
    - Research Outline - *Canvas > Modules > Research Project - due* ***next week****!*

# Data Analysis
## Fundamentals

# Data Analysis vs Analytics

- **Data analytics** = field of study using data and tools for business applications
- **Data analysis** = specific actions/techniques used for data analytics

- Goals:
    - Find trends
    - Predict actions, events, or triggers
    - Make decisions

# Data Analysis

*General components*

- Collection
- Cleaning & preparation
- Data analysis - initial (IDA) & exploratory (EDA)
- Data transformations
- Modeling & analysis

# The Data Science Lifecycle

*"Data science pipeline"*

1. Acquisition of data
2. Preparation & maintenance
3. Preprocessing
4. Analysis
5. Communication

# Data Acquisition

*Where do we get the data?*

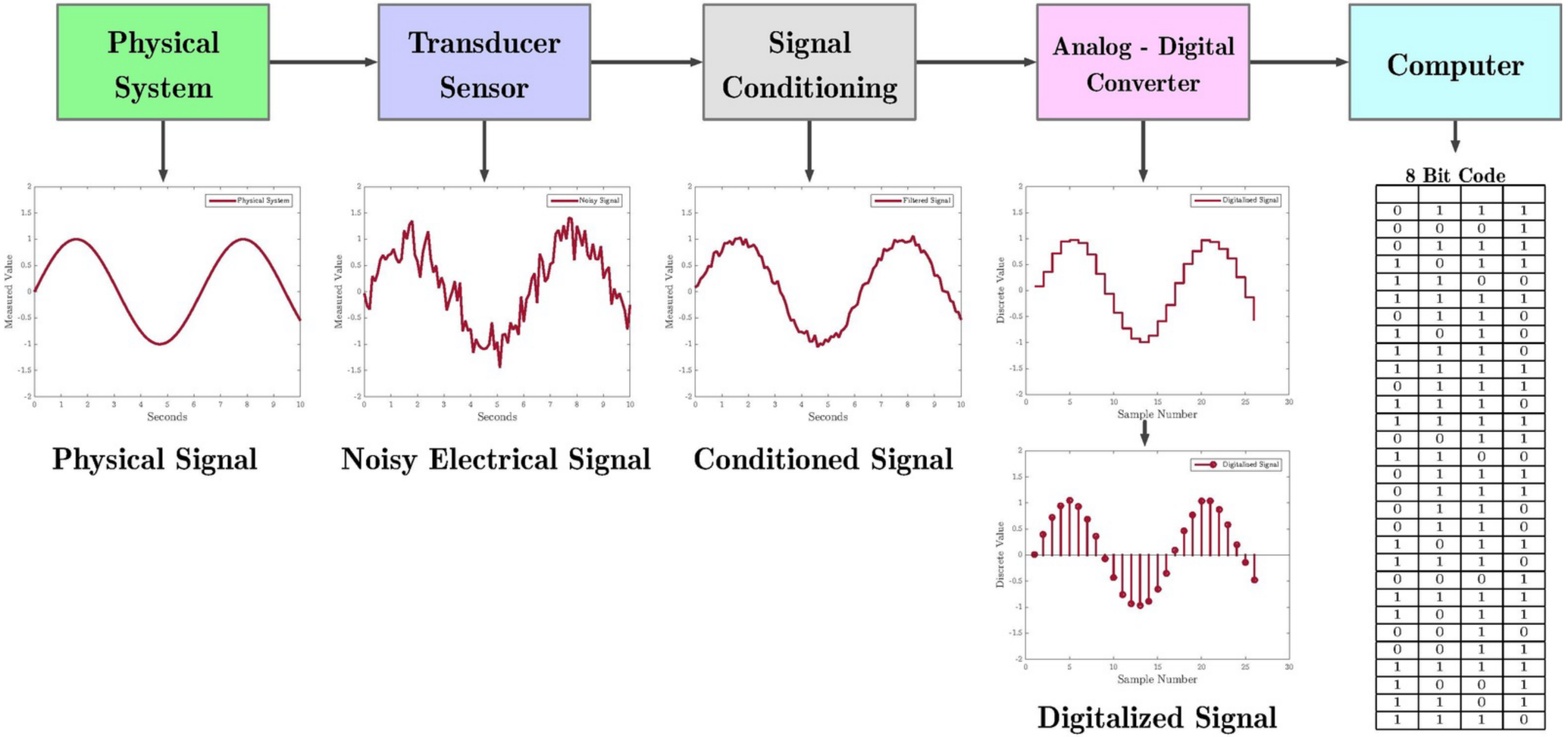

- Consider the source(s)
- Examples of data sources:
    - APIs
    - Web-scraping
    - Open source repositories
    - Data loggers or acquisition systems


**UC Irvine Machine Learning Repository**


https://new.nsf.gov/focus-areas/artificial-intelligence/nairr

# Data Acquisition

*Data acquisition systems (DAS, DAQ, DAU)*




| | | | |
|---|---|---|---|
| 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 |
| 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 |

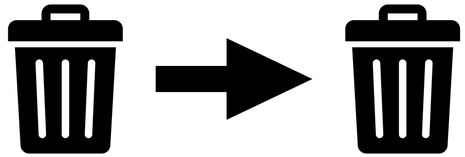


*Image source: Wikipedia*

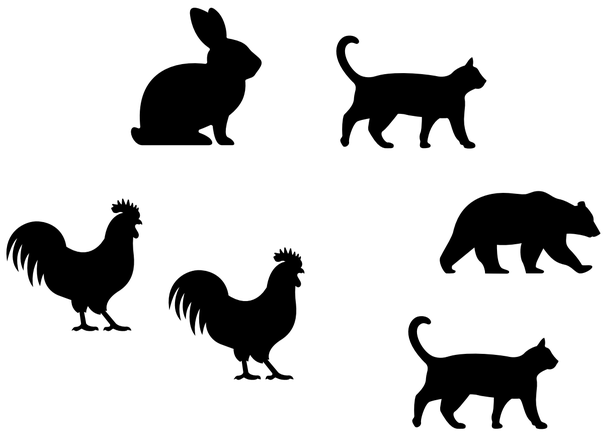
# Data Preparation

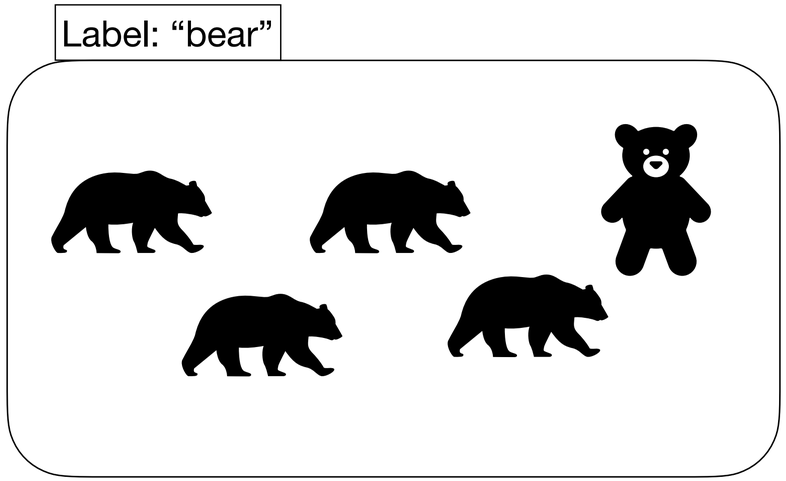
*Cleaning*

- **Better data** is preferable to any “fancy” method/algorithm.



- Question any **outliers** in your data.





# Data Preparation

*Cleaning*

- Reduce any **structural problems** in your data.
    - In text, fix typographical errors, capitalization, & any inconsistencies.
    - In images, you may decide to use one orientation (i.e. portrait, or landscape), or one image file type (i.e. JPEG)

- It is important to reduce duplicates and any **irrelevant information**.
    - Duplicate records/samples are typical when combining data from different sources.
    - Data is “irrelevant” if it doesn’t fit the specific problem you’re trying to solve!

# Data Analysis
## Error & Confidence

# Error & Confidence

In data analytics

- Data in the “real world” can be **uncertain**
    - *Uncertainty is the estimation of error present in our data*

- *Precision* and *accuracy* identify **error**
- *Mean* and *standard deviation* describe **confidence**

# Error

In data analytics

- **Systematic error**
  - *Also referred to as **bias**, or statistical bias*
- **Random error**
  - *Also random variation*

Let **T** be a statistic used to estimate parameter **X** and
let **E(T)** denote expected value of **T**

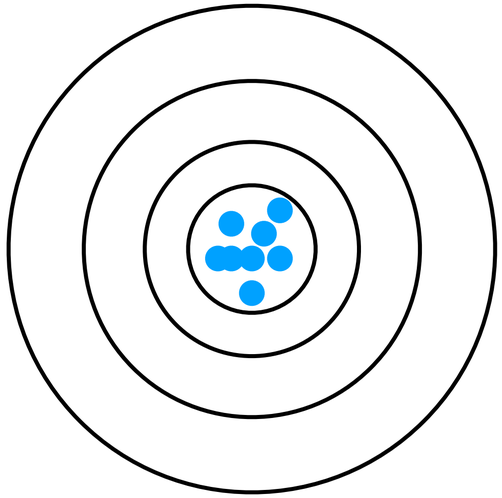
```
bias(T,X) = bias(T)

bias(T) = E(T) - X
```

If **`bias(T,X) = 0`**, then **T** is unbiased

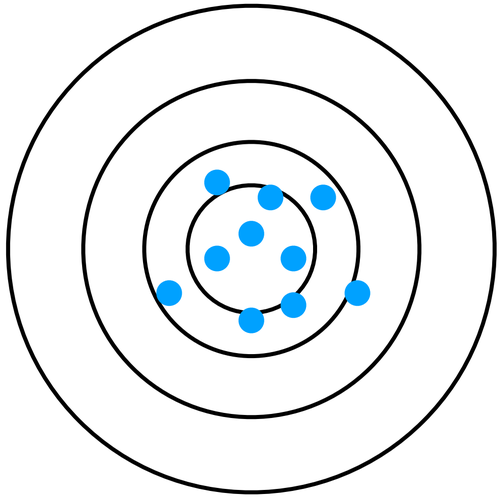
# Error & Confidence

## Precision vs. Accuracy



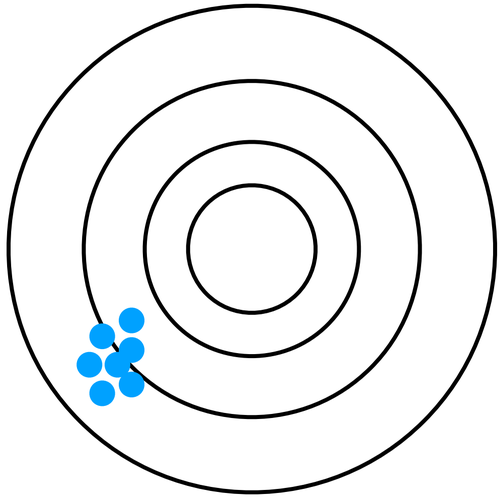
Accurate

Precise



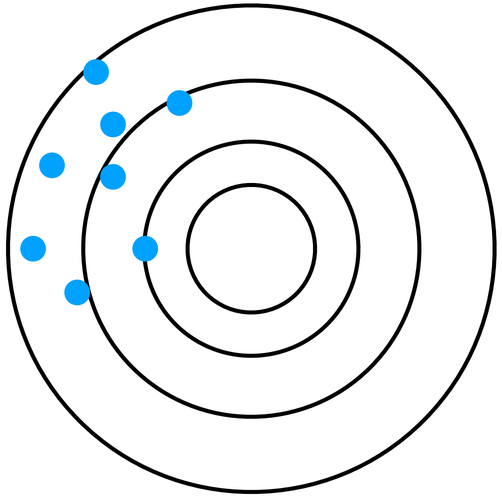
Accurate

Precise



Accurate

Precise



Accurate

Precise

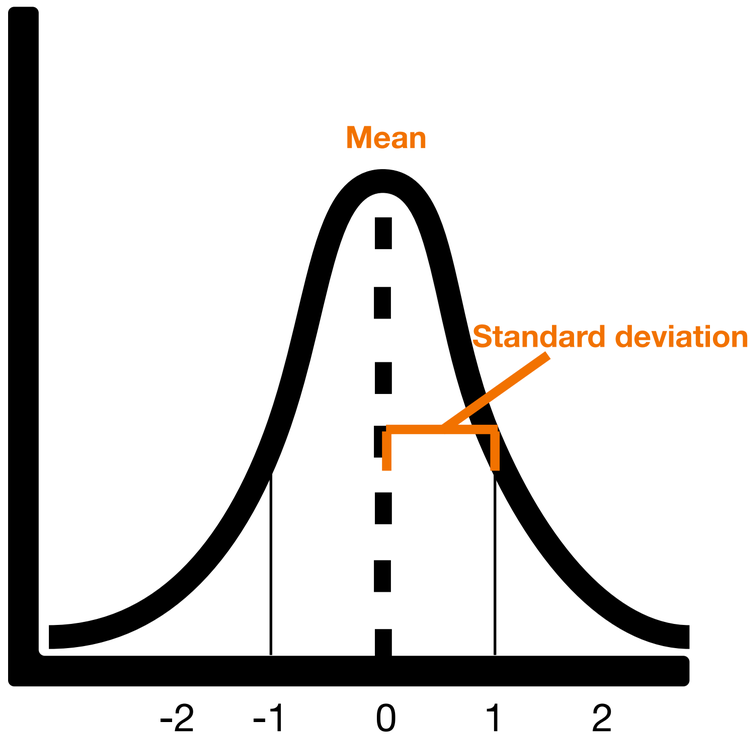
# Error & Confidence

*Uncertainty in data*

- **Confidence intervals** provide a range of estimates for an unknown parameter
- **Mean**

$$\frac{x1+x2+\ldots+xn}{N}$$

- **Standard Deviation**



# Data Visualization
## Techniques

# Data Visualization

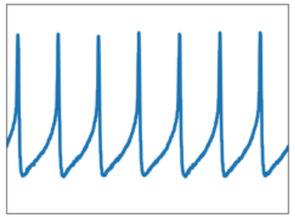
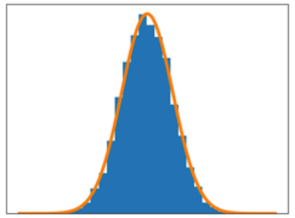
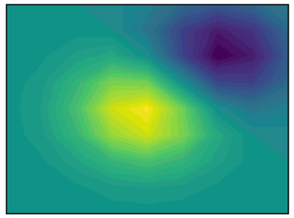
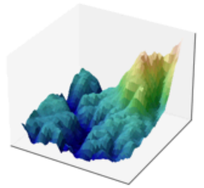
*Techniques & considerations*

- **Clarity & accuracy**
    - Ensure data representation is truthful & readily interpretable
- **Choose the right chart**
    - Different chart types (bar charts, line graphs, scatter plots, pie charts, etc.) are suitable for different data types & research questions
- **Aesthetics & design**
    - Importance of color, labels, & layout - *avoid clutter!*

# Matplotlib

*Python for data science*

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Matplotlib makes easy things easy and hard things possible.

### Create

- Develop publication quality plots with just a few lines of code
- Use interactive figures that can zoom, pan, update...

### Customize

- Take full control of line styles, font properties, axes properties...
- Export and embed to a number of file formats and interactive environments

### Extend

- Explore tailored functionality provided by third party packages
- Learn more about Matplotlib through the many external learning resources

https://matplotlib.org

information is beautiful

# A Matter of Fact?

Fact checking major documentaries scene-by-scene

TRUE | TRUE-ISH | MISLEADING | FALSE-ISH | FALSE | UNKNOWN

## 48.7% What the Health? 2017

SHOW ONLY TRUE TRUE-ISH MISLEADING FALSE-ISH FALSE UNKNOWN Info

Previous 00:50:50 Next


**Movie**

A surgeon cancels an interview for fear it might lose them paying patients.

**Reality** UNKNOWN

It seems like that's what's happening from the footage, but the message doesn't come from the surgeon directly and we don't know the full context of the cancellation.

https://informationisbeautiful.net/visualizations/what-the-health-netflix-documentary-fact-checked-debunked/

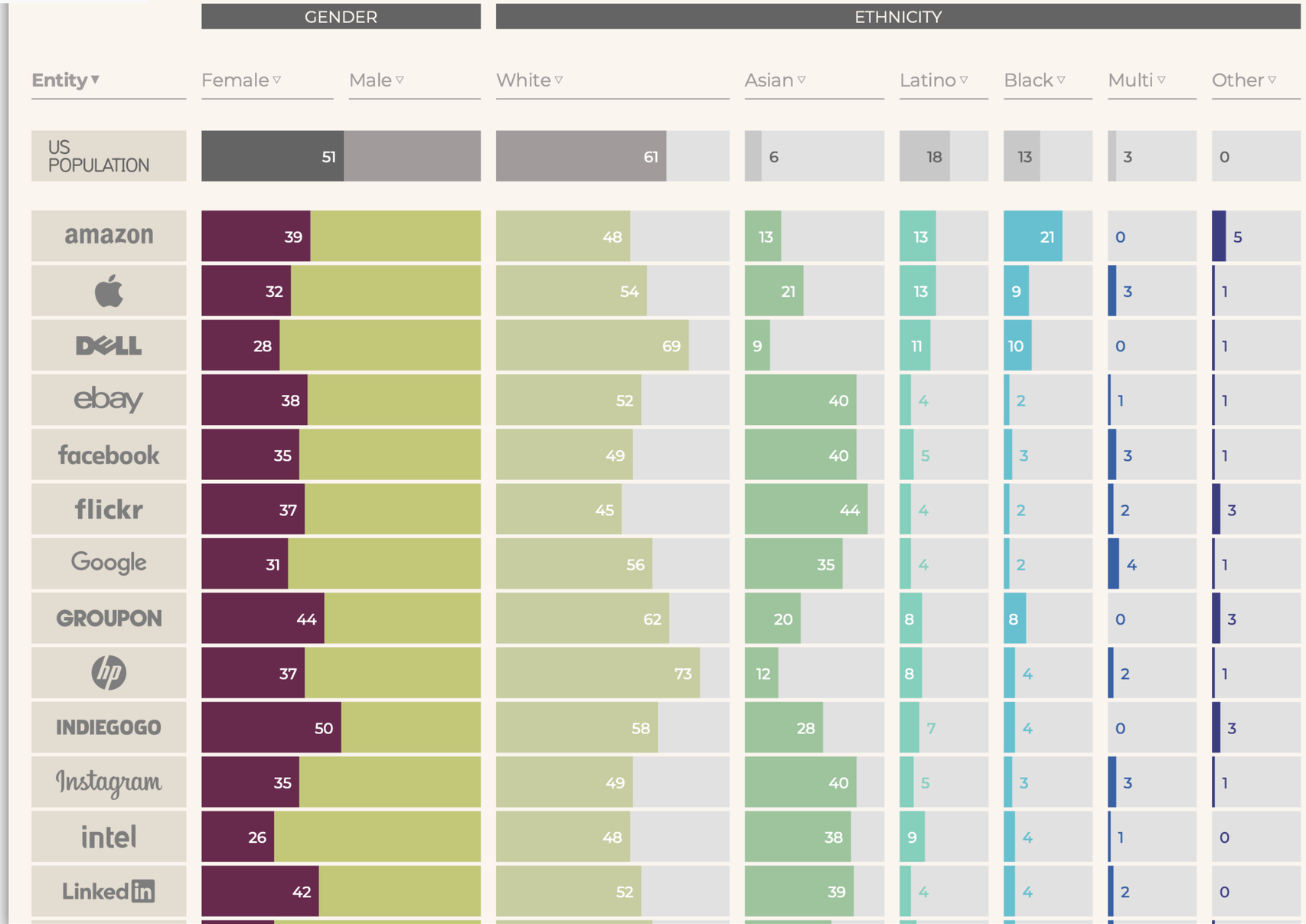
information is beautiful

# Diversity in Tech

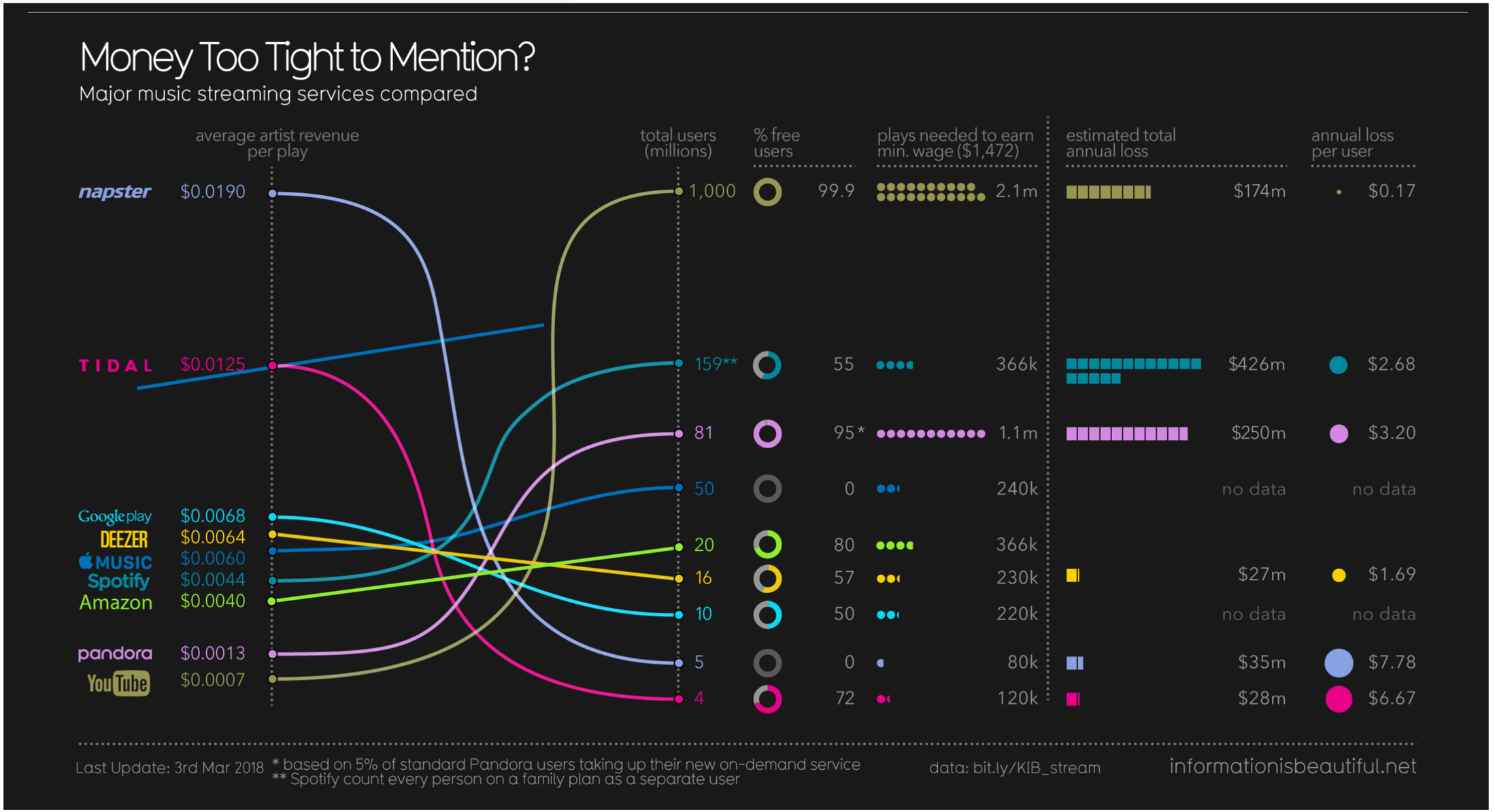
## Employee breakdown of key technology companies

year on year change

YEAR: 2014 2015 2016 **2017**



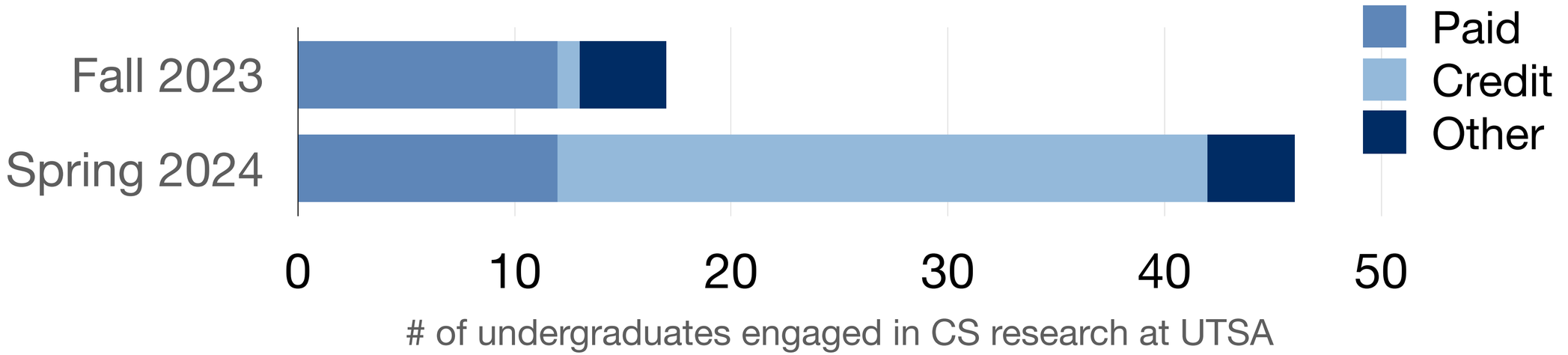
information is beautiful




https://informationisbeautiful.net/visualizations/spotify-apple-music-tidal-music-streaming-services-royalty-rates-compared/

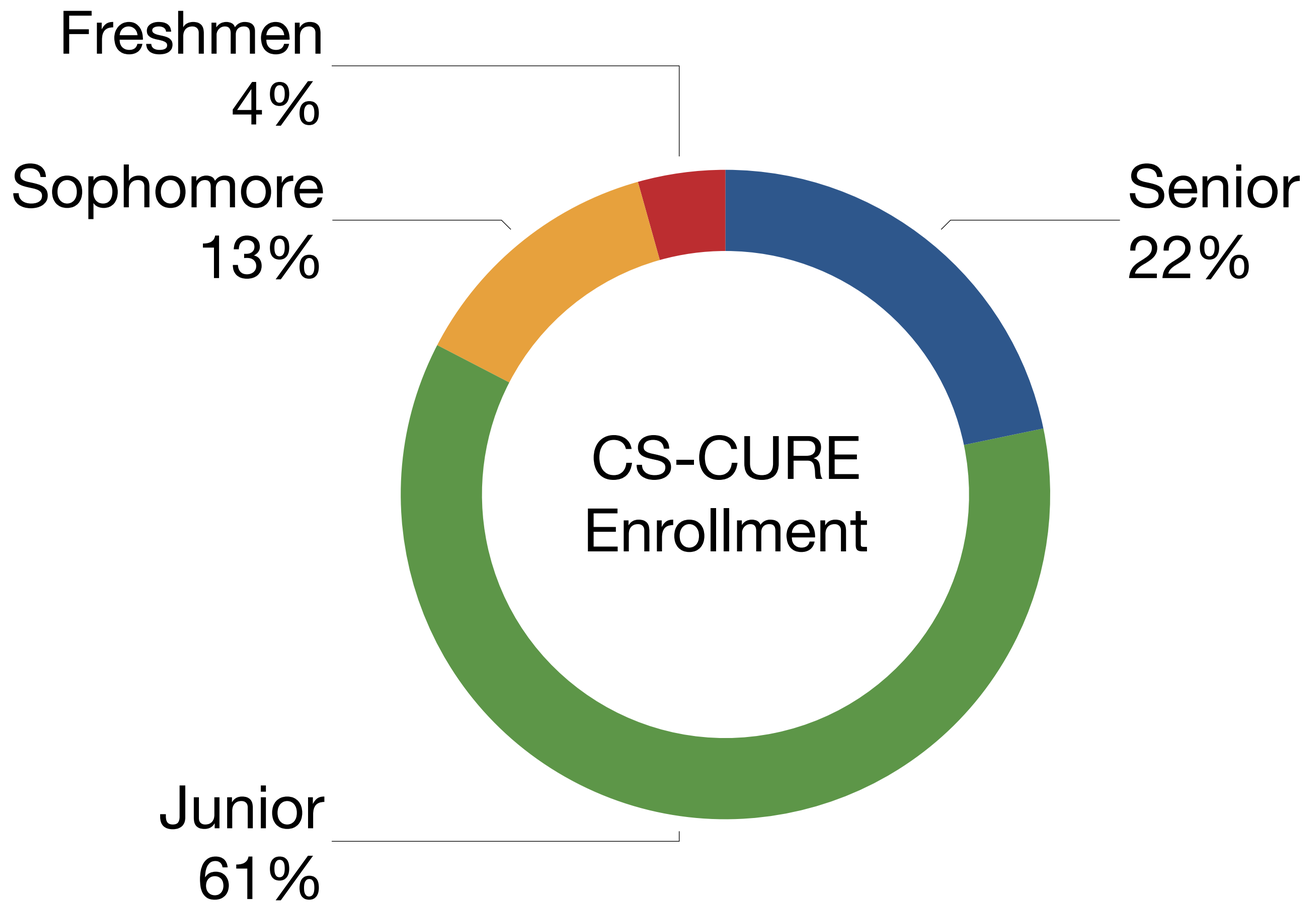
# Data Visualization

*Live examples!*



# Data Visualization

*Live examples!*



# Data Analysis
## Ethics

# Ethics & Considerations

*Data analysis*

- Ethical considerations protect both research participants and the integrity of research findings.

- Consider, for example:
  - **Privacy:** *protecting the anonymity & confidentiality of participants' data*
  - **Accuracy:** *data is complete & free from errors/biases*
  - **Transparency:** *about data collection, analysis methods, & limitations*

# Ethics & Considerations

*Data in research*

- **Consent**: Ensure data collection complies with informed consent procedures.
- **Data anonymization**: If possible, anonymize data before analysis to minimize risks.
- **Data security**: Implement secure storage and access protocols to protect data privacy.
- **Bias awareness**: Be aware of potential biases in data collection and analysis methods, and strive to mitigate them.
- **Honest reporting**: Report results truthfully and accurately, acknowledging limitations and uncertainties.

# Data Analysis - IRB

*Internal Review Board approval*

- Even with the best intentions, research can pose risks to participants.

- IRBs help ensure that:
    - potential risks are minimized
    - benefits are maximized
    - individuals participate voluntarily with informed consent.

# Data Analysis - IRB

*Internal Review Board approval*

**“**The UTSA IRB is the **standing committee that reviews & approves human subjects research** for the purpose of protecting the rights and welfare of participants.”

- Independent committee of experts
- Meet regularly to review applications
- Approve or deny based on established criteria

# Human Subjects Research–IRB

**What is the Institutional Review Board (IRB)?** +

**Who should submit research to the IRB?** −

UTSA faculty, students, or staff conducting human subject research as a part of their position at UTSA Investigators who wish to recruit UTSA students, faculty, or staff using non-publicly available information or who wish access to those individuals. Investigators conducting human subjects research using UTSA facilities (when UTSA is engaged in research – consult the IRB Office for further information regarding engagement in research)

**What is meant by "research"?** +

**What is meant by "human subject"?** +

**How do I apply for IRB approval?** +

**When should I submit my study to the IRB for approval?** +

**What is minimal risk?** +

**Is there a time period for IRB approval of a research study?** +

**How do I obtain approval for a change to my research?** +

**I want to submit a study with researchers from other institutions. How do I add them to my study?** +

**I'm a student/faculty member at another university. How do I get IRB approval to conduct research at UTSA?** +

# Key Elements of an IRB Review

*Data ethics & considerations*

- Does the research have a **valid scientific purpose**?
- Are the **risks** to participants reasonable compared to potential benefits?
- Is there a clear **informed consent** process for participants?
- Are participants' **data protected** and kept confidential?
- Will **vulnerable populations** (children, prisoners, etc.) be adequately protected if involved in the research?

# How do I do work with an IRB?

*Data ethics & considerations*

1. Develop your research protocol (clear description of your methods & procedures).
2. Complete any required IRB training.
3. Submit your protocol for review by the IRB.
4. Address any questions or revisions requested by the IRB.
5. Receive IRB approval *before starting your research*.

# Wrap-Up

*Tuesday*

- Identify the data & sources useful to a research initiative
- Understand ethical considerations for research execution & analysis

- To Do:
    - Activity 5: Data Analysis & Visualization (*in-class on Thursday*)

*See you Thursday!*

# Research Project:
# **Research Outline**

# UTSA CS-CURE

*Research Project*

- ☑ **Research Proposal** [week 3] - *identifying a research problem*
- ☐ **Research Outline** [week 8] - *organizing your research*
- ☐ **Research Draft** [week 12] - *telling a good story*
- ☐ **Research Paper** [week 16] - *depth of research in a field*

# Activity 5:
# **Data Analysis & Visualization**

# Wrap-Up

*Thursday*

- Identify the data & sources useful to a research initiative
- Understand ethical considerations for research execution & analysis

- To Do:
    - Activity 5: Data Analysis & Visualization (in-class on Thursday)

*See you next week!*