Visualization (Vis)

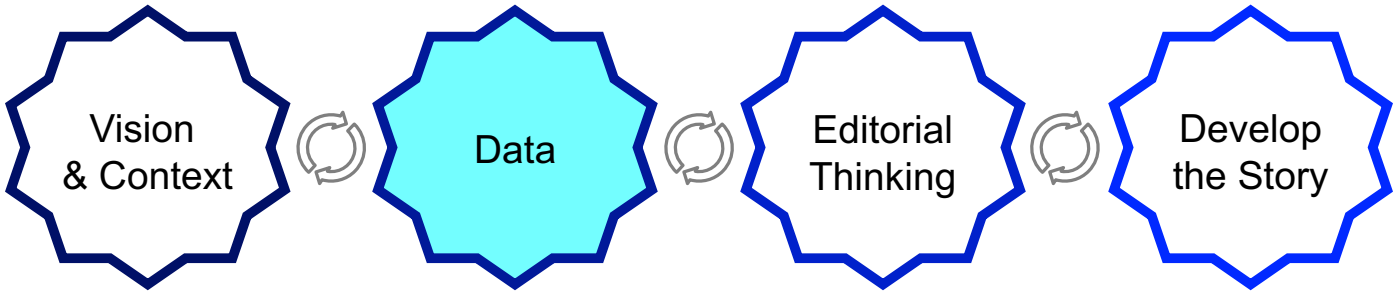
Storytelling with Interactive Data Visualizations

Lecture 3

–

Data

The Vis Process: Data



- Data Acquisition: Source the data to visualize
- Data Examination: Look at the data
- Data Transformation: Cleaning, Creating, Consolidating
- Data Exploration: See the data, broaden the viewpoint via EDA

Data

1. Data Acquisition
2. Data Examination
3. Data Transformation
4. Data Exploration

Importance of Structured Data Acquisition

- Quality Data = Foundation for Visualization
  - Accurate and reliable data: essential for creating meaningful visualizations
  - Quality of insights depends on the quality of the underlying data
- Understanding the Data Source
  - Need to evaluate credibility and trustworthiness of each source
  - Need to evaluate bias / hidden agenda of each source
  - Different sources can provide varied perspectives on the same issue
  - Combining data from different sources can lead to unexpected insights

Steps in Data Acquisition

1. Identify Data Requirements

- Define Objectives based on your vision / mission / intended effect / 3-min story
  - Clearly state what you need to achieve with the data
  - Determine the key questions you want to answer with the data
- Scope of Data Needed
  - Identify content, type, granularity, and time frame of the data required

2. Find Data Sources

- Types of Data Sources
  - Primary Data Sources
    - Data collected directly from experiments, surveys, or observations
    - Examples: field research, lab experiments, direct observations
  - Secondary Data Sources
    - Data collected by someone else, typically for a different purpose
    - Examples: government reports, published research, historical data

- Possible Data Sources
  - Supplied by stakeholder: usually for commissioned projects
  - System Download
    - Within the own organisation, e.g., ERP Systems, CRM Systems
    - Public Databases and Repositories, e.g., Kaggle, Data.gov, World Bank Open Data
  - APIs: access data programmatically, often streams of data, e.g., air quality, traffic disruptions
  - Web scraping: extract data from websites
  - Primary Collection: gather primary data
  - Data foraging: combine small, manually collected amounts of data

3. Documentation and Metadata

- Maintain Clear Documentation
  - Keep records of data sources, collection methods, and preprocessing steps
  - Document the collection method used
  - Include notes on reliability and limitations of the data source / collection method → domain knowledge is critical!
- Use Metadata
  - Metadata provides context and helps in understanding the data better
  - Examples: data description, data type, data collection date

Data

1. Data Acquisition
2. Data Examination
3. Data Transformation
4. Data Exploration

Data-Inherent Data Types (1)

Data-Inherent Data Types = Data Type, the data inherently has

- TNOIR classification: Textual, Nominal, Ordinal, Interval, and Ratio
  - Not all kinds of data are compatible with the TNOIR classification!
- Textual: unstructured passages of text
  - Typical visualizations: usually needs to be transformed before visualization (e.g., counting words)
  - Examples: “Any other comments?” in a questionaire; product description in a web shop
- Nominal: Data without a natural order or ranking
  - Characteristics: Labels or names, cannot be quantitatively measured
  - Typical visualizations: bar charts, pie charts
  - Examples: Gender, Eye Color

Data-Inherent Data Types (2)

- Ordinal: Data with a meaningful order, but intervals between values are not consistent
  - Characteristics: Ordered categories, Differences between categories are not uniform
  - Typical Visualization: Bar charts, ordered bar charts
  - Examples: Satisfaction Ratings (e.g., Poor, Fair, Good, Excellent)
- Interval: Data with meaningful intervals between values, but no true zero point
  - Characteristics: Consistent intervals, no true zero point
  - Typical Visualization: Histograms, line graphs
  - Examples: Temperature (Celsius, Fahrenheit), latitude and longitude
- Ratio: Data with meaningful intervals and a true zero point
  - Characteristics: Consistent intervals, true zero point allows for meaningful ratios.
  - Typical Visualization: Histograms, line graphs, scatter plots.
  - Examples: Height, Weight, Age.

Other Inherent Data-Type Distinctions

- Discrete vs. Continuous Data
- Scale of the Data: often linear scale, sometimes logarithmic scale
  - Example for logarithmic scale: strength of sound (decibel), magnitude of earthquake (Richter)
- Not all kinds of data are compatible with the TNOIR classification!
  - Example: time based data – TNOIR classification depends on format + usage

Technical Data Types

- Programming languages like Python offer ”technical” Data Types
  - Derived from the internal workings of a computer
  - Examples: String, Int, Float
- These do not correlate strongly with the data-inherent data types
  - Example 1: satisfaction rating (ordinal) data can be represented as String (“poor”, “fair”, “good”) or Integer (0, 1, 2) or Float (0.0, 10.0, 100.0)
  - Example 2: Discrete Data can be represented as String (“zero”, “one”, “two”) or Integer (0, 1, 2) or Float (0.0, 1.0, 2.0)

➔ Do not infer the inherent data type from the technical data type

➔ Domain knowledge = only way to find out the inherent data type!

Amount and Range of the Data

- Questions to answer
  - Highest and lowest value (for interval and ratio data)?
  - How many decimal places (for numeric data)?
  - How many different values (for nominal and ordinal data)?
  - Min/Max char length and word count (for textual data)?
- Methods to evalute the amount and range
  - Frequency Distributions (shape of the distribution): histogram, KDE plot
  - Frequency Counts: for ordinal/nominal data
  - Measurements of central tendency: mean, median, mode
  - Measurements of spread: min, max, quantiles, variance/std. deviation
- Why is this important
  - Vision/Mission stimulate visualization ideas, but shape/size of data is driving it

Example: OECD better live index

- Design: one flower for each of the 41 countries with 11 petals for the quality of live indicators
- Amount and Range of the data paramount for the design, would not work…
  - if there were 25 quality indicators
  - if there were 3 quality indicators
  - if there were 150 countries

Source: https://www.oecdbetterlifeindex.org/

Quality of the Data

- Undiscovered/Unresolved data quality issues undermine trust in/accuracy of visualization

➔ Discover and address data quality issues, e.g.

- Missing values
- Erroneous values
- Inconsistencies
- Duplicates
- Expired values
- System characters
- Leading/Trailing spaces
- Date issues (format, basis, …)

Representativeness of the Data

- Are the data tuples available representative?

➔ Is it a true random sample of the ground truth?

➔ Is it influenced by the data collection method?

➔ Is there a (hidden) limitation on the availability of the data leading to an obstructed view?

Your Turn !

Exercise 1

Inherent Data Types

Data

1. Data Acquisition
2. Data Examination
3. Data Transformation
4. Data Exploration

Cleaning the Data

- Goal: Increase the Quality of the Data
- Attempt to correct the Data Quality issues identified in the previous step, by for example
  - Removing Duplicates: Identify and eliminate duplicate records
  - Handling Missing Values: imputation, deletion, or using placeholders
  - Correcting Inaccuracies: Fix errors in data entry and measurement
  - Standardizing Formats: Ensure consistency in data formats (e.g., dates, currencies)
  - Outlier Detection: Identify and address outliers that may skew analysis
  - …
- Careful: may introduce bias in the dataset!

Creating new Attributes/Features

- Expand your data via new calculations, new groupings, etc.
- Most substantial and creative Data Transformation task
- Broaden analytical options to explore
- Examples
  - Create percentages based on existing quantities
  - Create rolling totals
  - Convert “start time” and “end time” into duration
  - Convert absolute qunatites for different geographic regions into “per capita” values
  - Create new domain specific categories like “child” if age<18

Creating new Attributes/Features from Textual Data

- Unstructured, textual data is often very valuable but hard to handle
- Examples: transformation of textual data into nominal data
  - Identify keywords
  - Compute summary themes
  - Flag existance of something, e.g. X is mentioned in this text
  - Flag existance of relationships, e.g. A was always mentioned before B
  - Determine sentiment
- Examples: transformation of textual data into ordinal/interval/ratio data
  - Frequency of certain words
  - Total word counts
  - Number of sentences
  - Reading duration
  - Word / Document embeddings

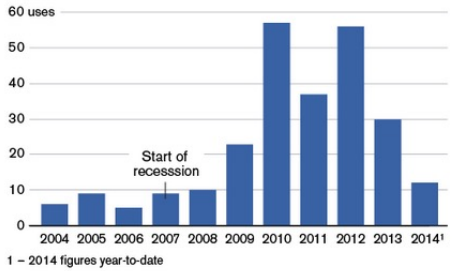
Example for Transformation of Textual Data

Graphic Language: The Curse of the CEO

Chief executive officers' use of profanity in public changes with the economy. A review of thousands of conference calls recorded in the past 10 years shows CEO cursing spiked after the recession in 2009 and waned as the recovery strengthened. The three CEOs who were quoted the most often swearing on calls have all been taking steps to tone it down in recent years. Here's a review of the four most common words of choice: the infamous F-bomb, the scatalogical S-word, the blasphemous G.D., and the derogatory A.H.

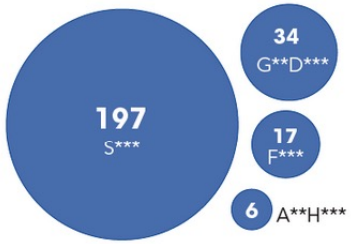
PUBLISHED JULY 16, 2014 RELATED ARTICLE >>

Cursing on conference calls by year *(F, S, A.H. and G.D.)*

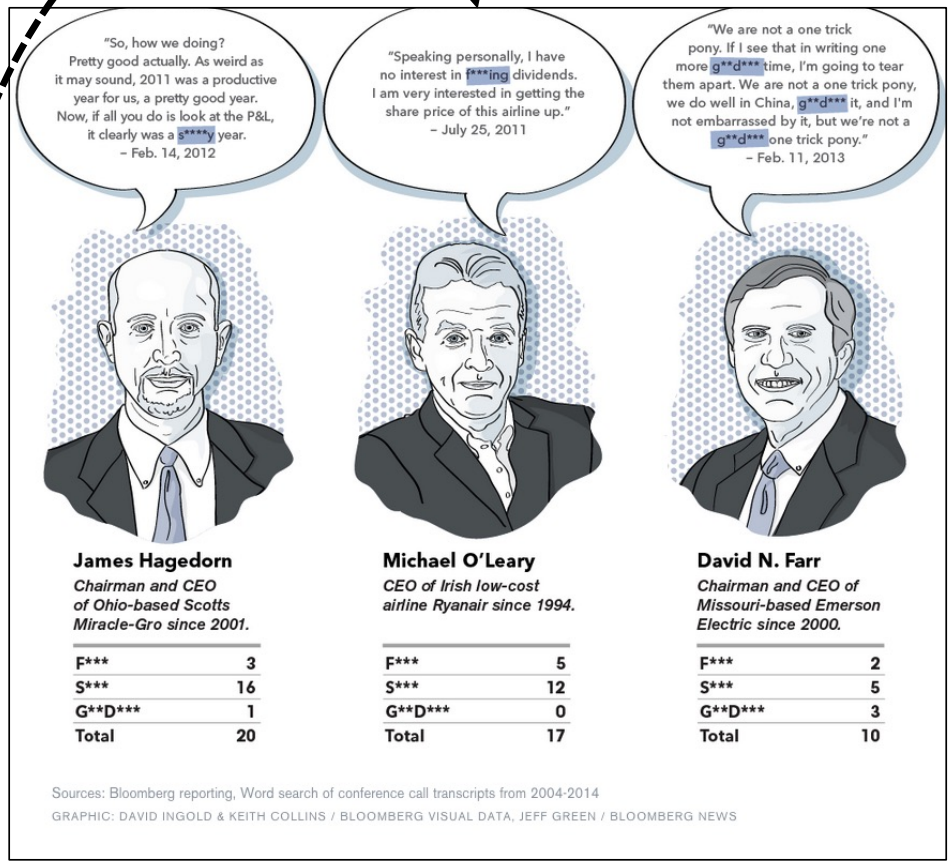


1 – 2014 figures year-to-date

Usage by word, 2004 to 2014¹



Top 3 salty CEOs: Quotes from conference calls More quotes



Source: https://www.bloomberg.com/graphics/infographics/graphic-language-the-curse-of-the-ceo.html

Consolidating: Expanding and Appending

- Expand: Add more attributes (“make it wieder”)
  - Often via. Feature Creation (see before)
  - Or acquire more attributed for the data (e.g. from a different source)

- Append: Add more tuples (“make it taler”)
  - Often relevant by time-related (streaming) data: add new tuples every period

Your Turn !

Exercise 2

Creating New Features

Data

1. Data Acquisition
2. Data Examination
3. Data Transformation
4.  Data Exploration

Widening the Viewpoint

- Examination ➔ get deeply acquainted with the data
- Exploration ➔ discover insights & qualities of understanding

➔ Move from ***looking*** at the data to ***seeing*** the data
  - What answers to your motivating curiosity can you find?
  - What other enlightening insights can you unearth?

➔ Find ”all” potentially interesting things you *could* show your audience (later: decide which one you *will* show!)

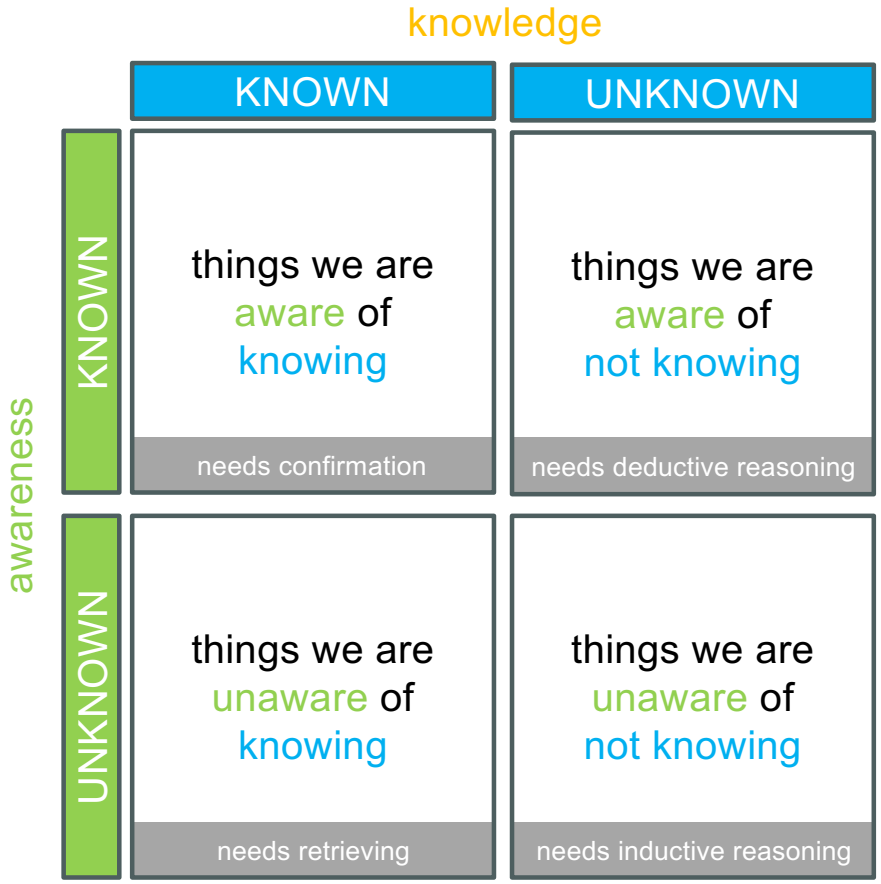
Widening the Viewpoint

Focus on

- Known Unknowns
- Unknown Unknowns

(i.e. the right side of the matrix)

Convert them into Knowns!

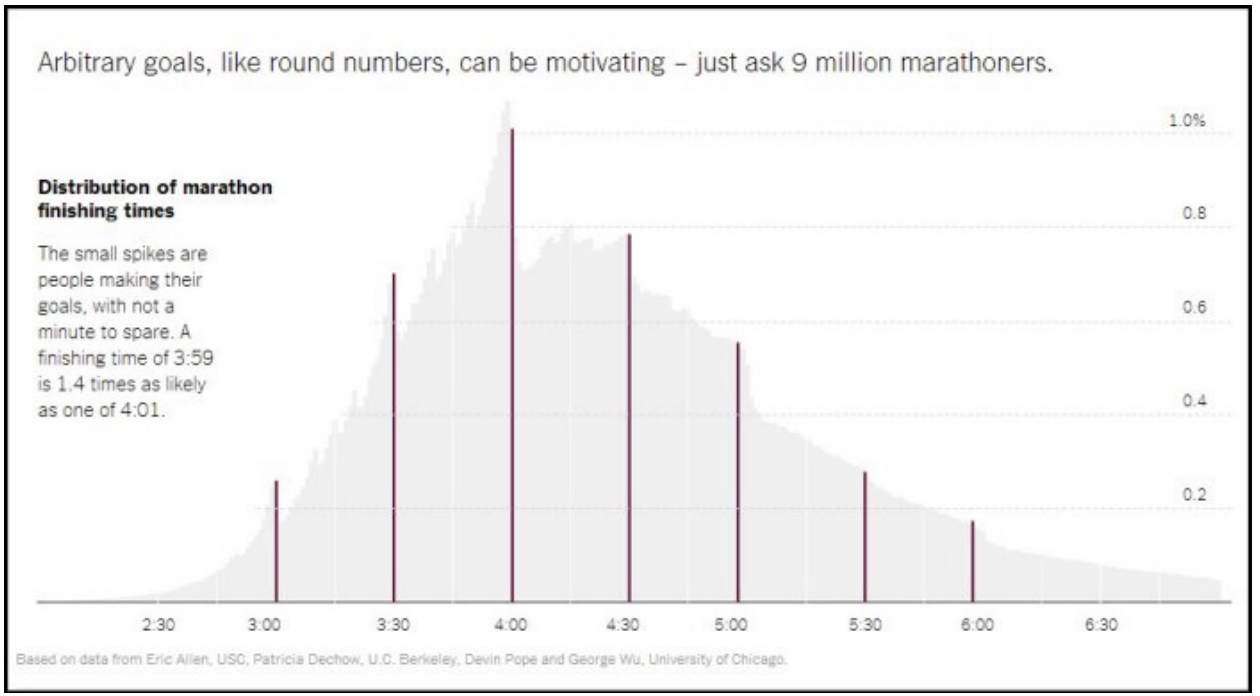


Rumsfield matrix in the version of Andy Kirk

EDA

- Exploratory Data Analysis (EDA) = Statistical + Visual Analysis of the Data
  - Show what is in the data and what is not!
- Example
  - Pattern in the finishing times of marathon runners
- How to find such unknown (un)knowns?

”EDA is an attitude, a flexibility, and a reliance on display, not a bundle of techniques.” (Edward Tufte)



Source: https://www.bloomberg.com/graphics/infographics/graphic-language-the-curse-of-the-ceo.html

EDA – finding the (un)known unknowns

So, what DO you need to conduct a successful EDA?

- Instinct of the analyst
  - Natural inquisitiveness and sense to “feel” what (statistical/visual) techniques to use and when
- Reasoning
  - Deductive Reasoning – confirm/reject the known unknowns
  - Inductive Reasoning – “play around” to find unknown unknowns
- (Domain) Research
  - Interpreting the results of statistical/visual tools only possible with domain knowledge
- Handle Nothings

Key Takeaways

- Data Acquisition
  - Identify Data Requirements, Find Data Sources, Document Metadata
- Data Examination
  - Inherent Data Types, Ranges, Quality, Representativeness
- Data Transformation
  - Cleaning, Creating, Consolidating
- Data Exploration
  - Broaden the viewpoint
  - Find (Un)Known Unknowns
  - EDA

Photo by Dragonfly Ave on Unsplash

Project Work

Team Setup

- Today, we will set up the teams for the PStA (Project)
- Teams will decide on the Visualization they want to create
- Teams will consist of usually 4 students (3 depending on number of participants)
- Signing up for a team means signing up for the class
  - If you sign up for a team today and do not finish the project, you will be graded “not passed” and you will have to repeat the class 1 year from now (it is only offered once per year), which may extend the number of semesters you study → right now is the last chance to simply “walk away” from this class!
- You will have to work together on the project outside of class

➔ Please group yourself into teams of 3-4 students now

➔ Agree on a ±3h slot in the week where all of you are available to work on the project (you will not need this slot every week, but you need to have one!)

Brainstorming of Vis Ideas

- Spend the reminder of todays class brainstorming first ideas for your Vis Project
  - Brainstorming Idea Pointers:

    Think about your hobbies!

    Think about your friends' hobbies!

    Think about interesting dataset you came across recently!

    Think about recent news and events!

    Did you wonder about anything strange recently?

    Anything you have always wanted to know more about?

- Formulate and Write Down potential Motivating Curiosities + Intended Effects