

ID430B: Data Analytics for Designers 디자인 특강V <디자이너를 위한 데이터 분석>

# Lecture 7

## Data Visualization 3/3

Tak Yeon Lee <takyeonlee@kaist.ac.kr> (takyeonlee.com)  
AI-Experience-Lab (reflect9.github.io/ael)

# Things to Learn

## 1. What shall we do after EDA

## 2. Data Reporting

How to convert informal observations to formal reporting

## 3. Data Storytelling

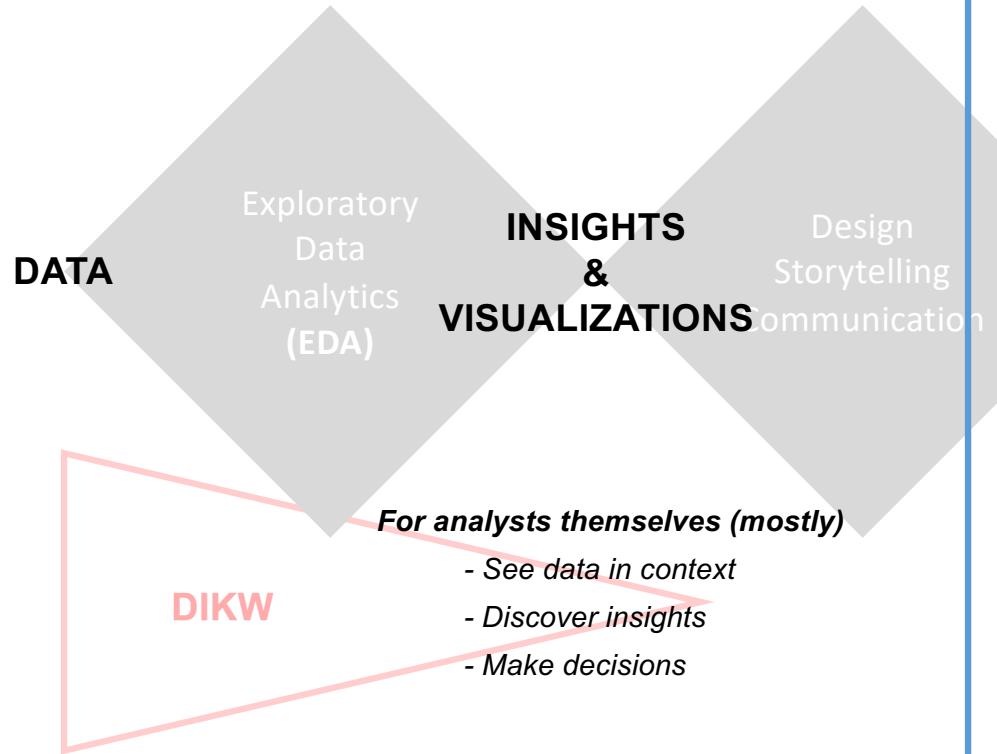
Goals, Techniques

Infographics, Data Videos, Journalism Articles

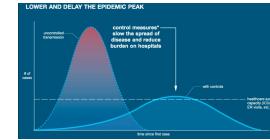
## 4. Guiding users to find their own insights

Product and Service features that employ data visualization

# What shall we do after EDA



## This Week



**Infographics**

**Dashboards**

**Products/Services**

**Data Videos**



**For other people**

- Guide the viewer's attention
- Communicate data and insights to the viewer
- Enable the viewer to get their own insights

# **Converting informal observations to formal reporting**

# 1. Introduce Dataset

Report how it was collected, processed, and analyzed

Report **data issues** and how you addressed

Report both **frequency and ratio**

Consistent number formatting for the same type of data

## Example: Reporting participant flow

Of the 298 participants who completed the initial screening survey, **78 (26.1%)** participants were **excluded for not meeting study criteria**, as they did not drink caffeine (**11.5%**) or they used prescription medication (**15.1%**).

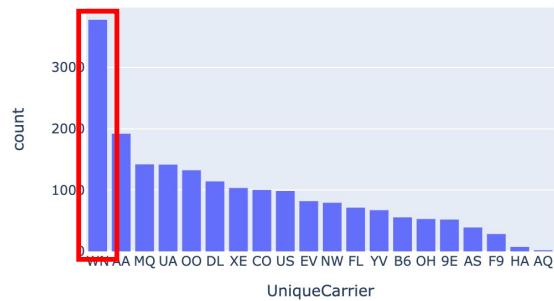
The remaining 220 participants were invited to complete the online study survey in exchange for study credit. However, an additional 12 participants failed to complete it, resulting in a final total of 208 participants.

<https://www.scribbr.com/apa-style/results-section/>

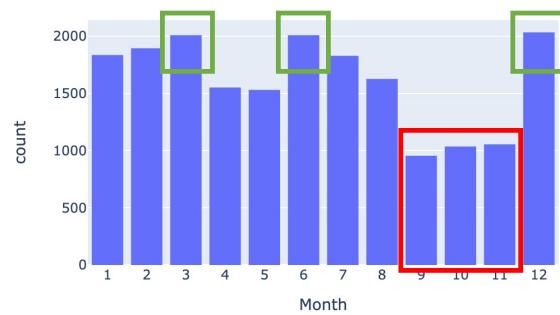
## 2. Summarize Data

Report couple important columns (univariate or combined)

Frequency Count for nominals



- Report the most frequent value(s) if there's significant gap between the rest
- Report both frequency and ratio
  - *E.g. Among the airline carriers, WN has the largest portion of delayed flights (3875; 28.5%), which is almost double (195.3%) of AA the second largest.*

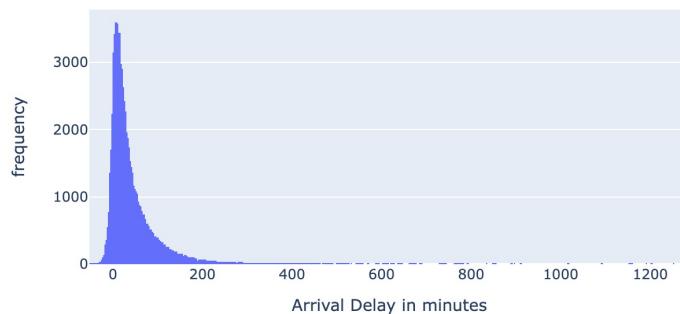


- There is no hard rule for choosing what to report. It depends on your goal of EDA.
- **Similar items** could be reported as a group
  - *E.g. The numbers of delayed flights from September to November (Average 997.5 flights per month) are less than half of the largest December*
- **Largest and smallest (groups of) items** are reported in usual.

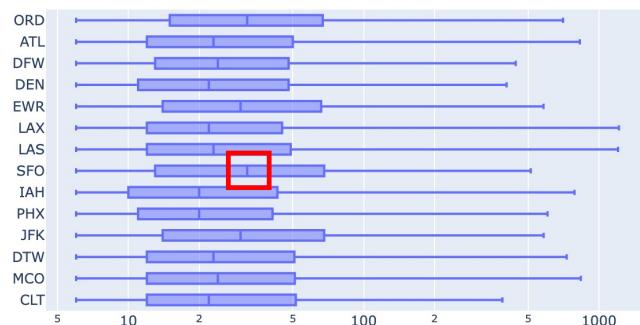
## 2. Summarize Data

Report couple important columns (univariate or combined)

Descriptive Statistics for quantitative columns



- Report the statistical distribution with parameters
  - E.g. Arrival Delay has a long-tail distribution on both sides ( $M = 42.1$ ,  $SD = 56.5$ )
  - Note that the parameter names are capitalized and italic. There is a space around '='. Number formatting is consistent.



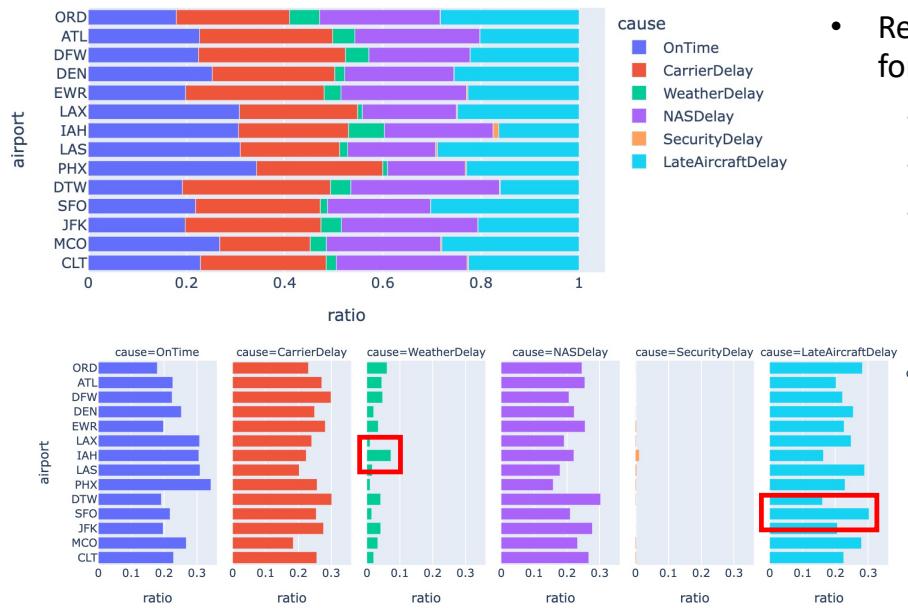
- Box plots are convenient for creating formal reports
  - E.g. SFO has the longest departure delay ( $Median = 32.0$ ,  $SD = 21.5$ )

## 2. Summarize Data

Report couple important columns (univariate or combined).

It's okay to start with a bivariate analysis. However, make sure the chart is not too complex for readers to understand. Also, you might need to break it down.

### Bivariate analysis



- Report the smallest and largest ratios with consistent number formatting

- E.g. ORD has the lowest ratio of on-time flights (17.9%)
- E.g. PHX has the highest ratio of on-time flights (34.3%)
- E.g. SFO (30.2%), LAS (28.9%), ORD (28.2%), and MCO (23.2%) are heavily impacted by belated aircrafts.

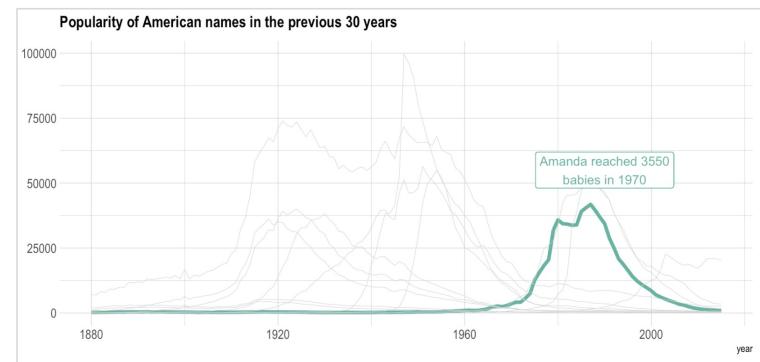
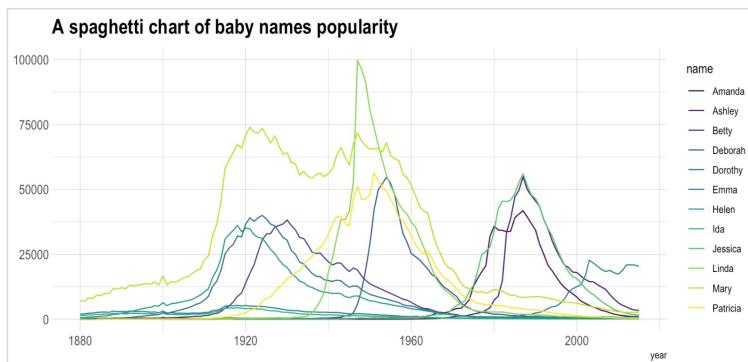
- Report with **hypothetical reasoning** for further analysis

- E.g. IAH suffers from humid tropical weather, as 7.3% of the entire flights were delayed by the weather condition
- E.g. Flights going to SFO are most likely to be delayed, since 30.2% of flights departing SFO got delayed for the cause

## 2. Summarize Data

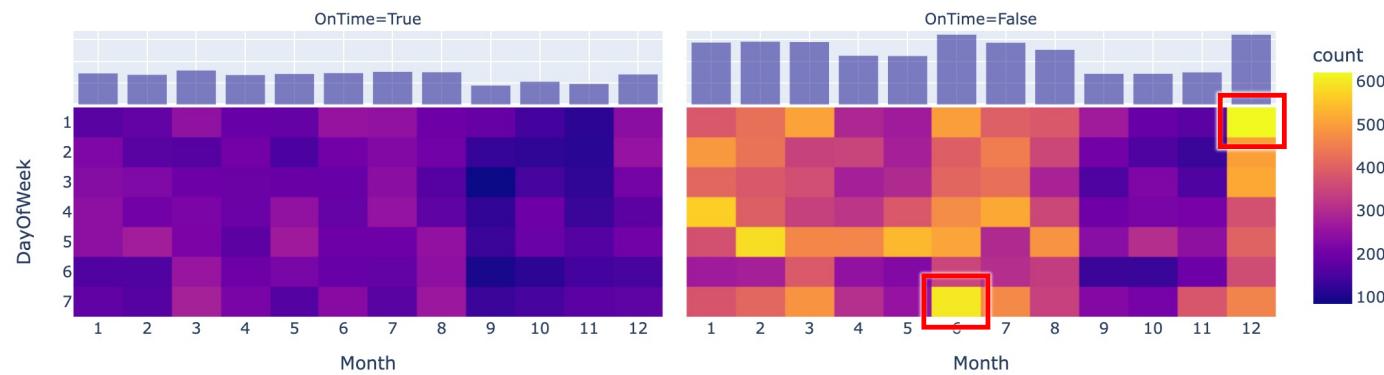
Use Annotations and Highlights for interesting findings in complex charts

- If the chart is too complex for readers to see the finding, repeat the same chart and
  - highlight the marker of interest and/or
  - add an annotation



### 3. Report Relationships

#### Multi-variate Analysis



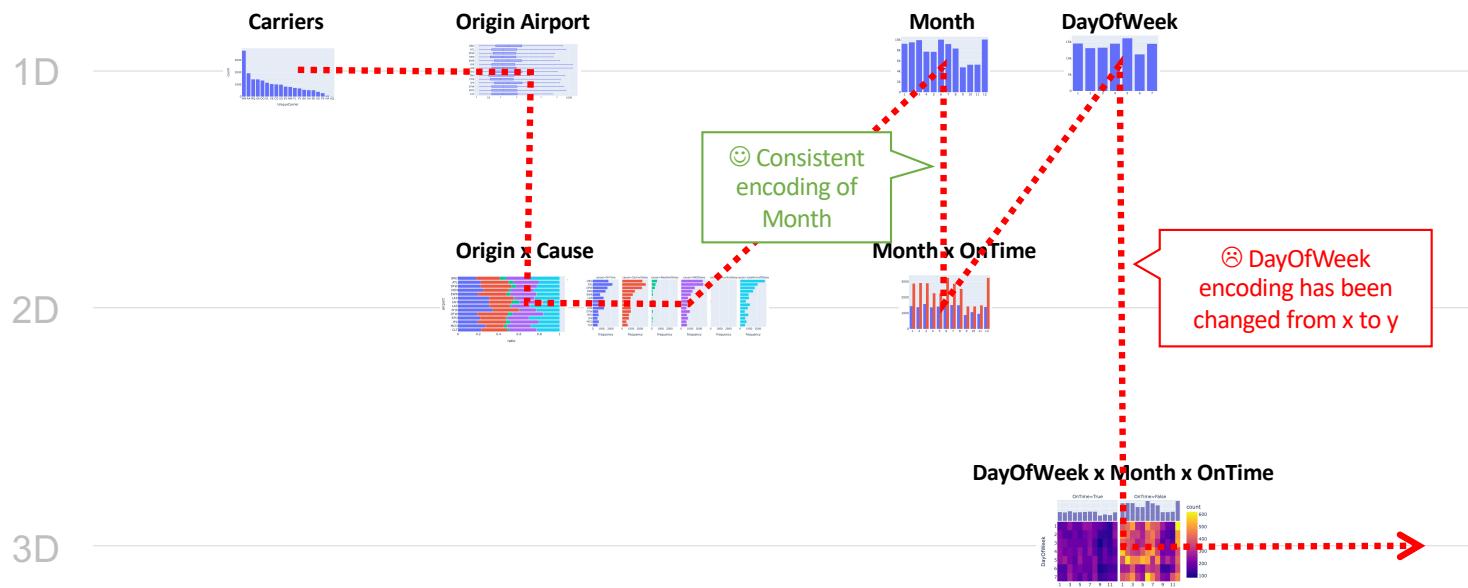
- With multivariate analysis you can get not only insights but also hypotheses (i.e. potential reasons) that lead to further analysis or ML models.

E.g. June is one of the two months with the highest volumes (3235; 20.3% of the whole delayed flights) of delayed flights. In June, delayed flights are mostly on Sunday (18.6% of all delayed flights in June). A potential reason is that lots of holiday travelers are returning home.

E.g. In December (Month:12), delayed flights are mostly on Monday (19.2% of all delayed flights in December). A follow-up analysis is needed to verify whether the delays were due to the cold weather in early morning and back-to-office commuters.

# Common Flow of Reporting

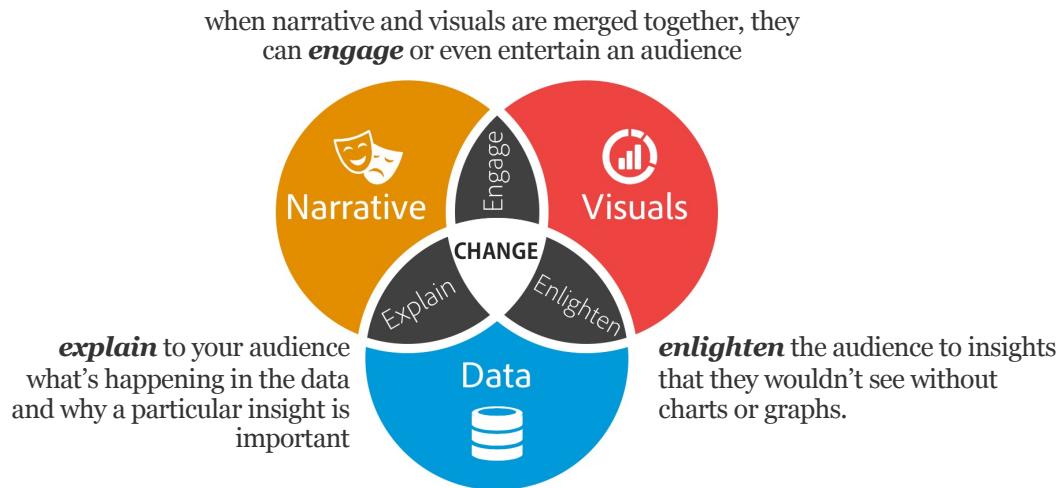
- Begin with univariate analysis (1D chart), and gradually build up multivariate analysis (2D, 3D, and higher dimension)
- Make sure that readers are aware of the previous steps that are relevant to the current step. For instance, before showing the (DayOfWeek x Month x OnTime) heatmap, readers must have seen (and still remember) the three columns.
- Avoid dramatic changes of column encoding



# Data Storytelling

# What is Data Storytelling

- Data storytelling is much more than just creating visually-appealing data charts. Data storytelling is a structured approach for **communicating data insights**, and it involves a combination of three key elements: **data, visuals, and narrative** [1]



[1] <https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/?sh=128a815952ad>

# What is Data Storytelling

“

Data storytelling couples data visualization with compelling narratives that help audiences better comprehend and take action based on data analysis. While effective data visualization helps people grasp and remember key takeaways, data storytelling is essential for helping them understand why those takeaways matter.

**Lydia Hooper**, Data Visualization Expert and Information Design Writer



From Venngage's Data Storytelling in Marketing: Benchmark Report 2021

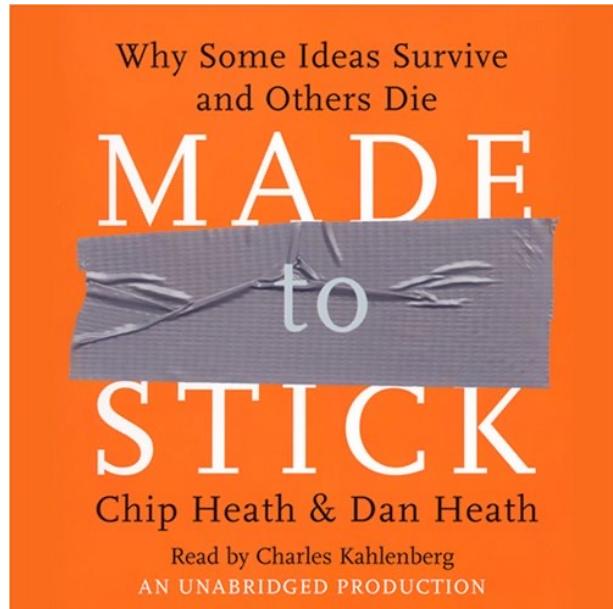
# **Three Goals of Data Storytelling**

Memorability

Persuasiveness

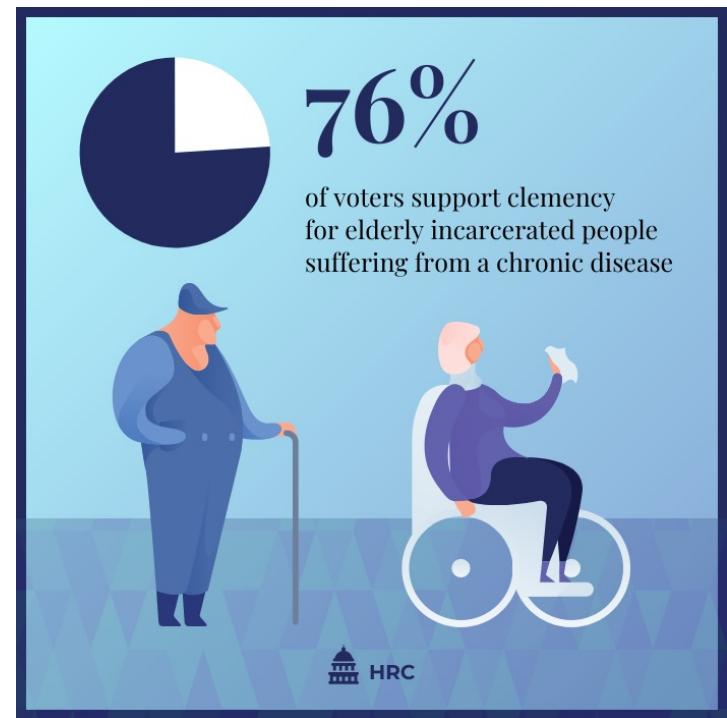
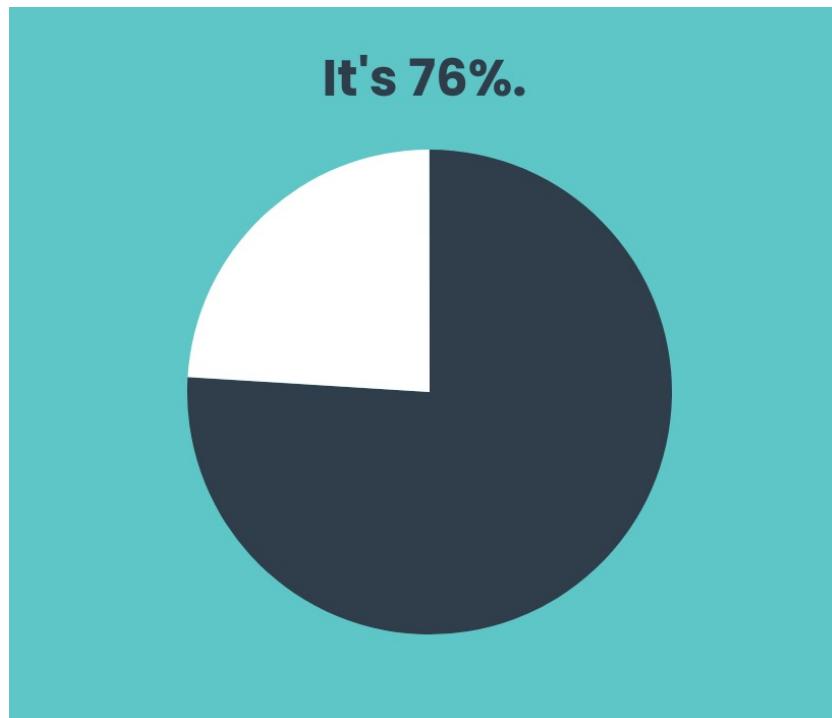
Engagement

# Memorability



"In the average one-minute speech, the typical student use 2.5 statistics. Only one student in ten tells a story. Those are the speaking statistics. The "remembering" statistics, on the other hand, are almost a mirror Image: When students are asked to recall the speeches, 63% remember the stories. Only 5% remember any individual statistic." – Made to Stick, Heath

# Persuasiveness



# Engagement

“In listening to stories we tend to suspend disbelief in order to be entertained, whereas in evaluating statistics we generally have an opposite inclination to suspend belief in order not to be beguiled.” - [stories vs. statistics](#) by John Allen Paulos



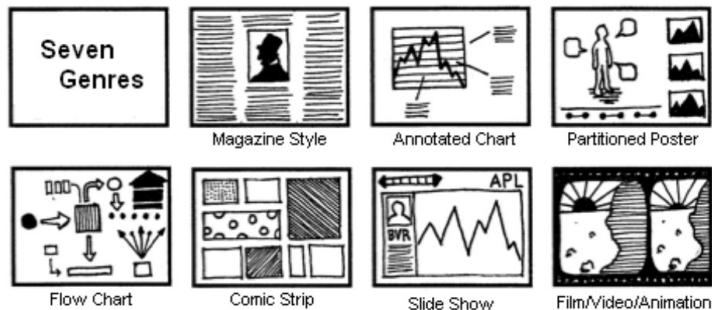
[Hans Rosling's 200 Countries, 200 Years, 4 Minutes](#)

– The Joy of Stats - BBC Four

Review Paper (Segel and Heer, 2010)

# Narrative Visualization: Telling Stories with Data

## Genres



## Author-driven / Reader-driven stories

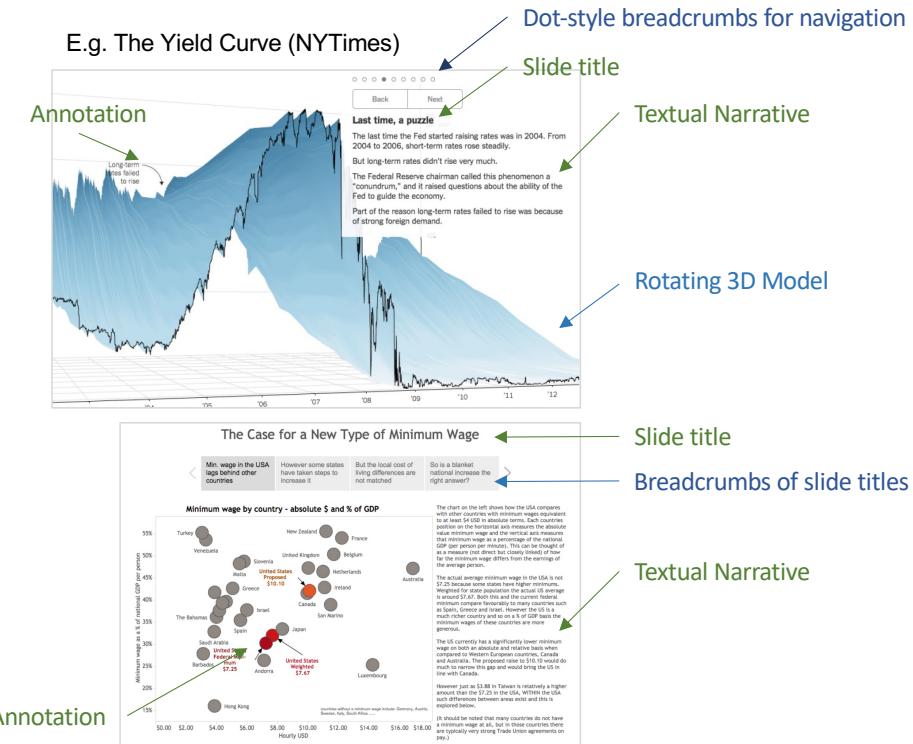
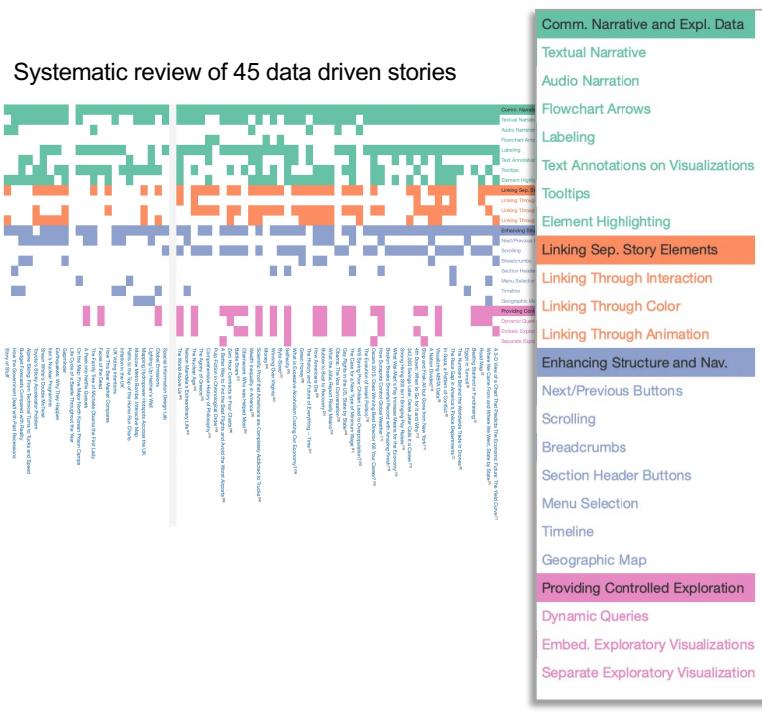
- 
- The image shows three icons representing narrative approaches:
- Martini Glass**: An illustration of a martini glass with a narrow base and a wide top, symbolizing a funnel-shaped narrative.
  - Interactive Slideshow**: A horizontal sequence of three rectangular slides, each with a waveform graphic.
  - Drill-Down Story**: An illustration of a hand interacting with a complex, branching interface, symbolizing a non-linear, exploratory narrative path.
- Martini Glass**
- Begins with author-driven introduction
  - Later let readers freely explore via interactivity
- Interactive Slideshow**
- Follows a typical slideshow format
  - Interactive mid-narrative within each slide
  - A more balanced mix of author-driven and reader-driven approaches
- Drill-Down Story**
- Presents the overview first
  - Readers choose what to explore
  - Put more emphasis on reader-driven approach

Full Paper (<http://vis.stanford.edu/files/2010-Narrative-InfoVis.pdf>)

Review Paper (Stolper et al., 2016)

# Emerging and Recurring Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Recent Stories

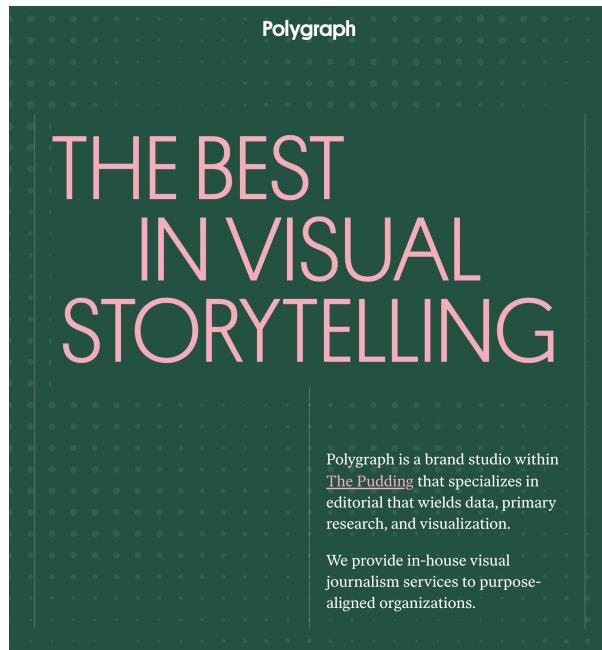
Systematic review of 45 data driven stories



Full Paper (<https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/MSR-TR-2016-14-Storytelling-Techniques.pdf>)

# Data-driven Journalistic Articles on the web

Two Prominent Resources of Journalistic Data Visualizations



<https://polygraph.cool/projects>

In a year with so many world-shaking moments, our strongest visual stories covered impeachment, outbreak, caucuses, primaries, donations, delegates, shutdown, jobs, lark deaths, coughing, hospitals, testing, George Floyd, protests, tear gas, dreams, nursing homes, policing, the election, stimulus, symptoms, Beirut, power, the Supreme Court, ballots, wildfires, Trump's taxes, turnout, swing states, masks, results, A.I., voter shifts, I.C.U.s, pollution and vaccines.

**2020: The Year in Visual Stories and Graphics.**

By The New York Times December 30, 2020

<https://www.nytimes.com/interactive/2020/12/30/us/2020-year-in-graphics.html>

# Vertical scroll through data storytelling

ThePudding

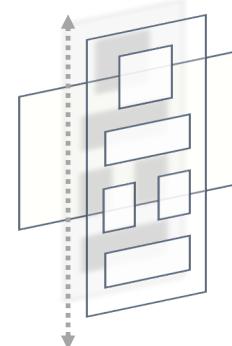
BY [RUSSELL GOLDENBERG](#)

## Craft beer – so hot right now. Could Sunnyvale, California, really be the microbrew capital of the US?

Text scrolling over visualizations creates the feeling of narrative space

Weighted Pivot Scatterplot visually connect two 1d metrics with a 2d scatterplot

*"Most audiences would simply scroll down and stop briefly, but rarely interact with charts"*  
- Zoya Bylinskii, Adobe Research

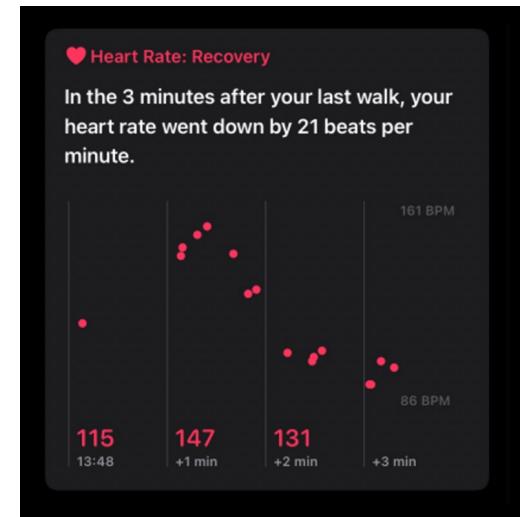
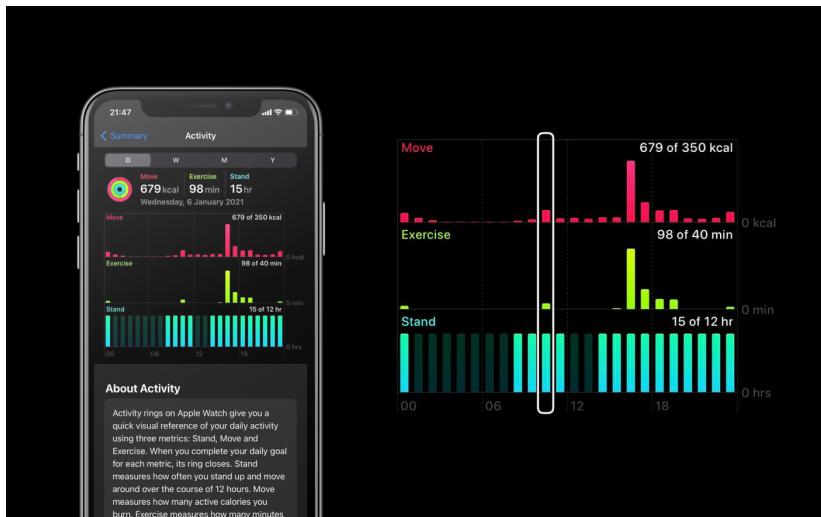


<https://pudding.cool/2017/04/beer/>

# **Guiding users to find their own insights**

Data Visualization as part of products and services

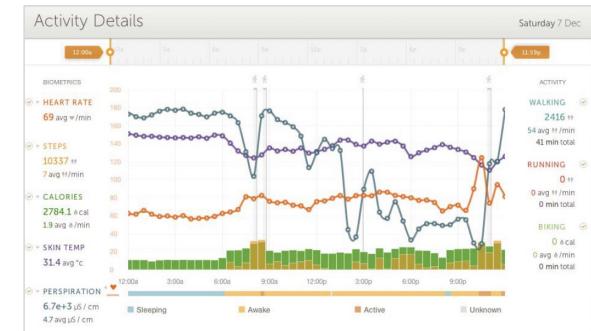
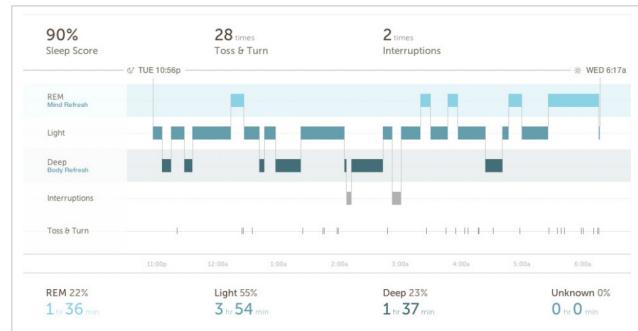
# Apple Health & Activity



**Personal informatics** is a class of tools that help people collect personally relevant information for the purpose of self-reflection and self-monitoring. These tools help people gain **self-knowledge** about one's behaviors, habits, and thoughts. It goes by other names such as living by numbers, personal analytics, quantified self, and self-tracking.

# Facilitating self reflection with data visualization [1]

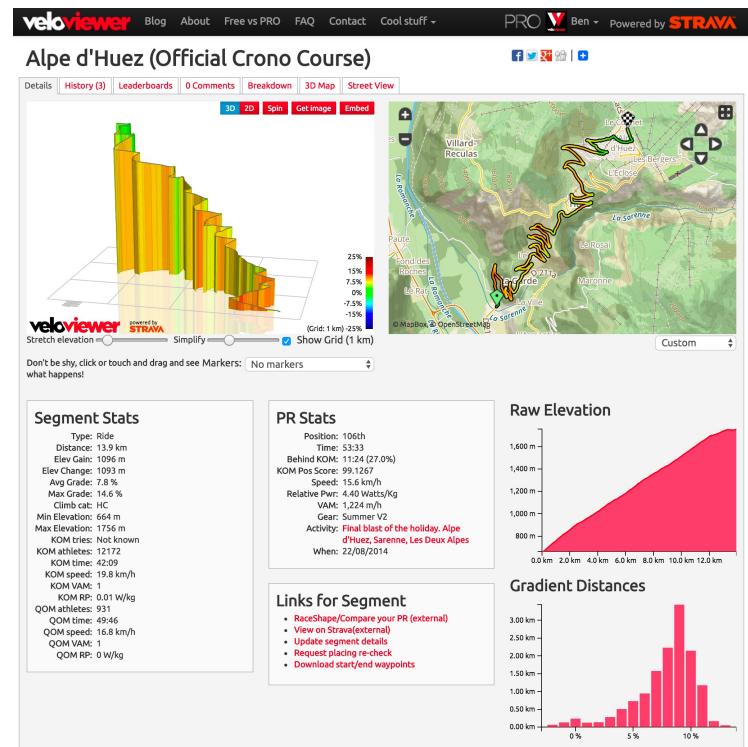
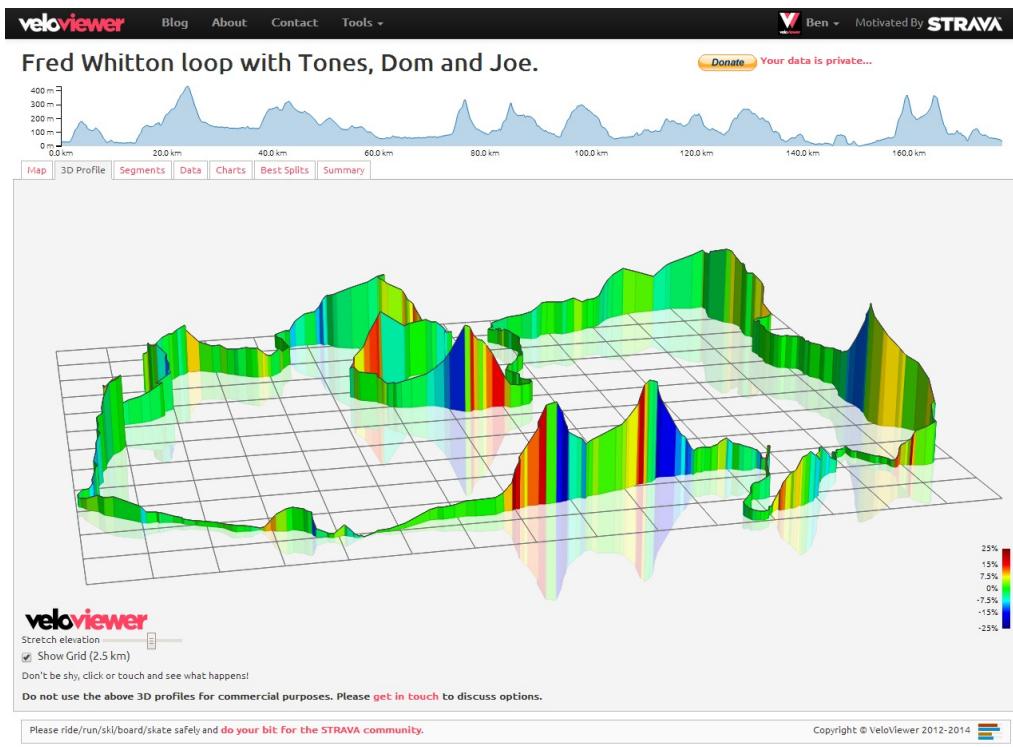
1. Make Data Interpretable at a Glance
2. Enable Exploration of Patterns in Time Series Data
3. Enable Discovery of Trends in Multiple Data Streams
4. Turn Key Metrics into Affordances for Action



[1] ["Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics"](#), Cuttone et al., 2014

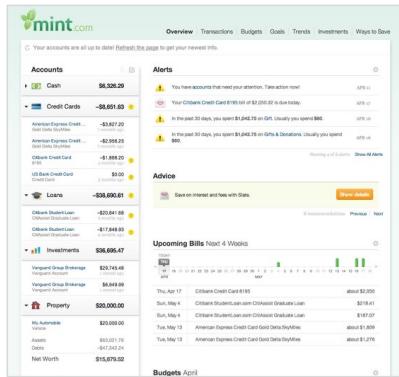
# Visualizing Bicycle Rides

Data visualization is often a core value of consumer products. How serious can it be? What do they need to support?



# Other Daily Products Employing Data Visualization

Personal Spendings



Sleep Tracker



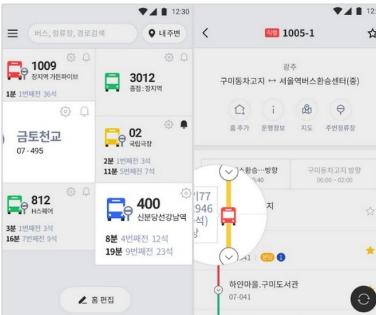
Electric Car – Energy Consumption



Weather Forecast



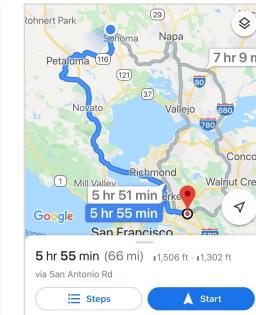
Utility Bill



Public Transportation



Stock Trading



Car Navigation

# Takeaways

- For many products and services, data visualization has become core features
- Most products do not support full EDA capabilities
- Many products focus on providing rich and expressive visual effects
- Goals may not be just “finding insights” but also
  - Reflection and motivation
  - Making quick decisions
  - Feeling of achievements
  - Simulating *What-If* scenarios
  - And more...
- Crucial to know both the capability of data and user’s needs

# Conclusion

- After fully understanding the data through EDA process, designers would know how to make use of it
- We covered a variety of design outcomes such as
  - Data Report
    - Guidelines of how to report findings from EDA
  - Data Storytelling (for memorability, persuasiveness, and engagement)
    - Taxonomy of Storytelling Techniques
    - Resources of journalistic articles on the web
  - Data-driven Features in products and services
    - Might not need full EDA capabilities, but designers must fully understand data through EDA
    - Goals might vary according to product context and user needs
    - Intersection of data capability and user needs