COSC 3337 : Data Science I

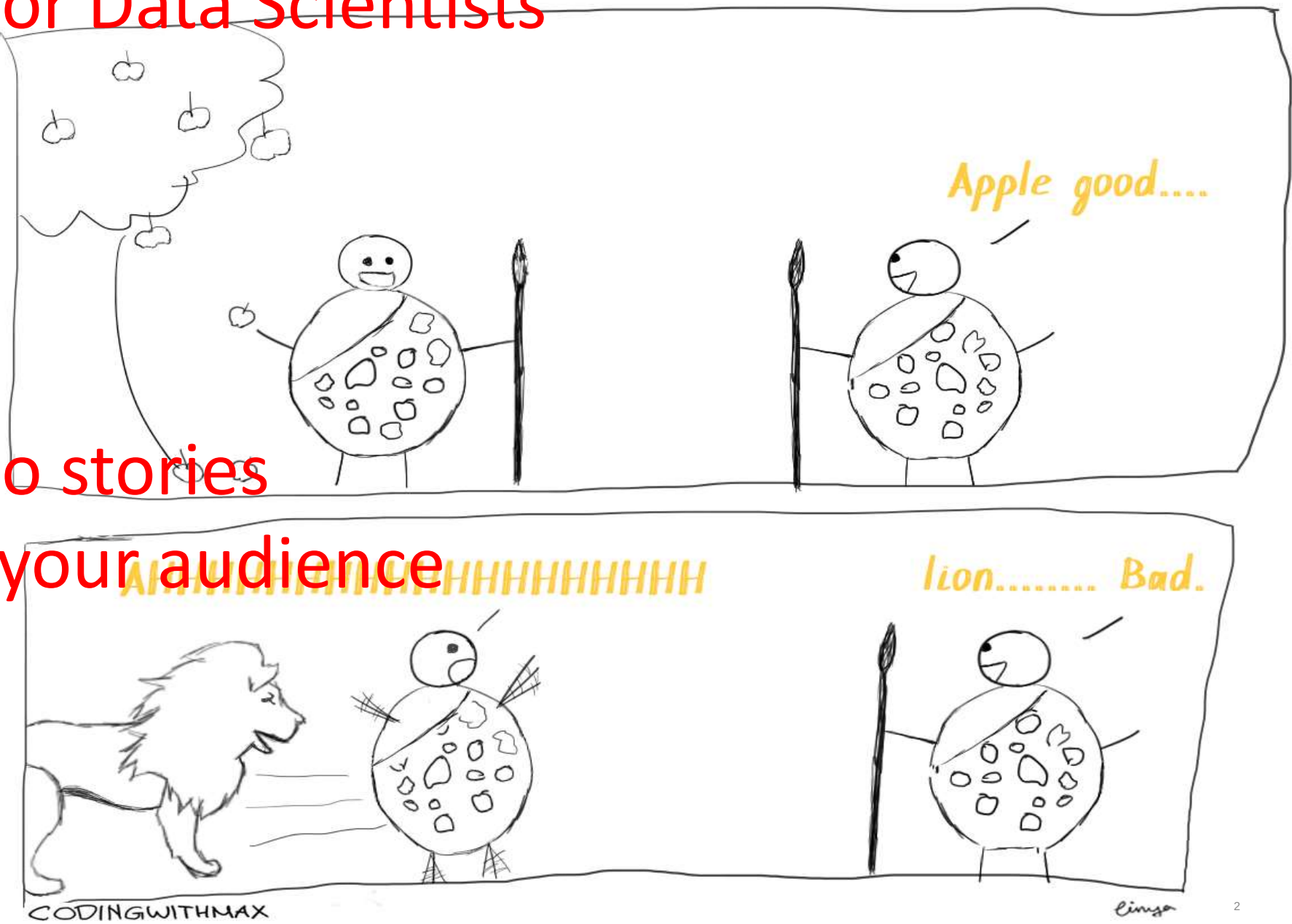
N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

Storytelling for Data Scientists



Turn data into stories
to persuade your audience

Storytelling

Data analysis is only half of the story

Data analysis is a very important skillset for scientists, because models are built on the results that we see in experiments, and if we are able to properly analyze our experimental data, we are able to formulate models that better represent reality.

The other half is that you also need to be able to:

1. Communicate your findings to others
2. Convince others that what you've found is indeed correct

Dr. Hal R.Varian

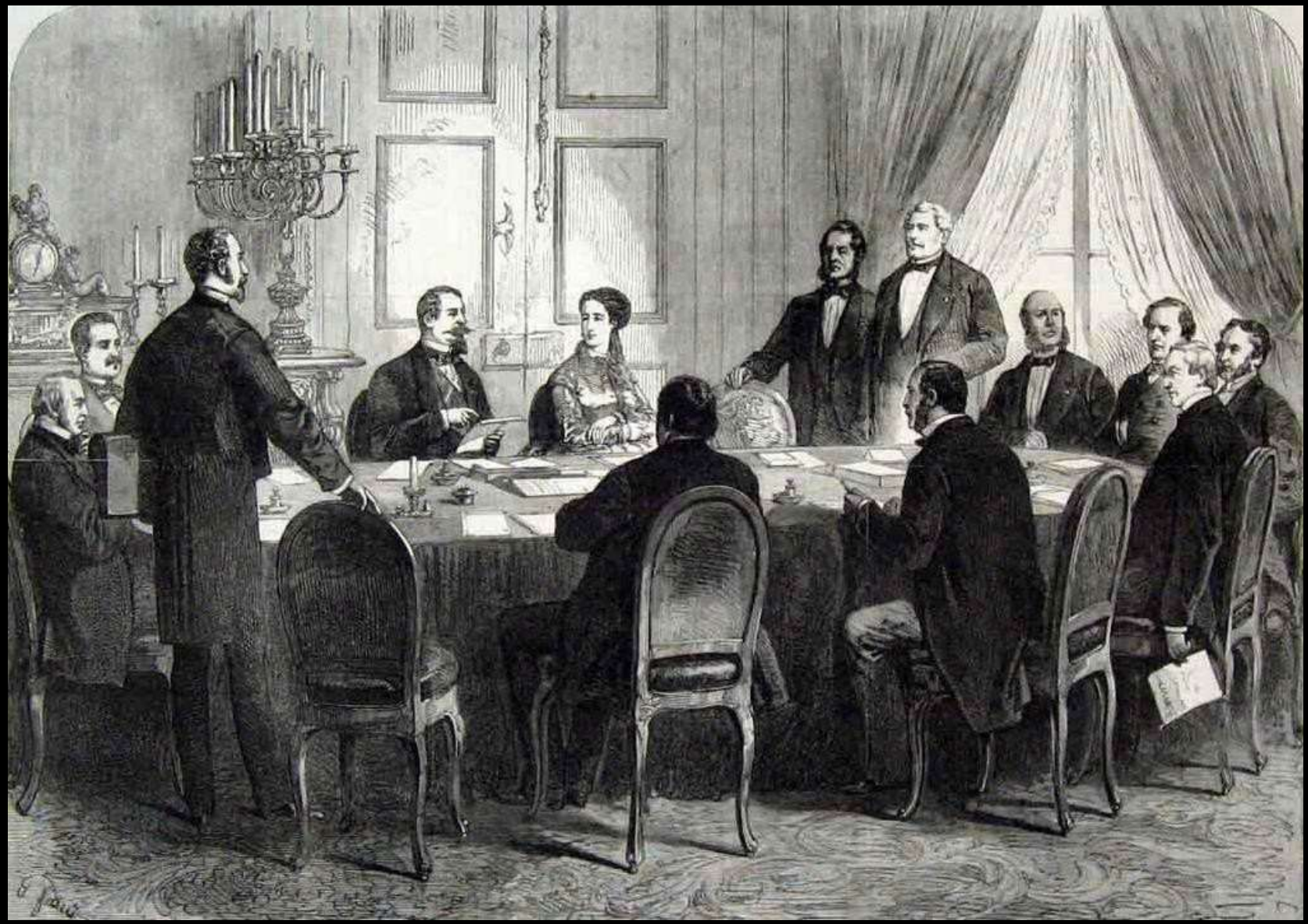
During a 2009 interview, Google's Chief Economist Dr. Hal R.Varian stated, "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades." Fast forward to 2016 and many businesses would agree with Varian's astute assessment.

Let me tell you a (data) story.

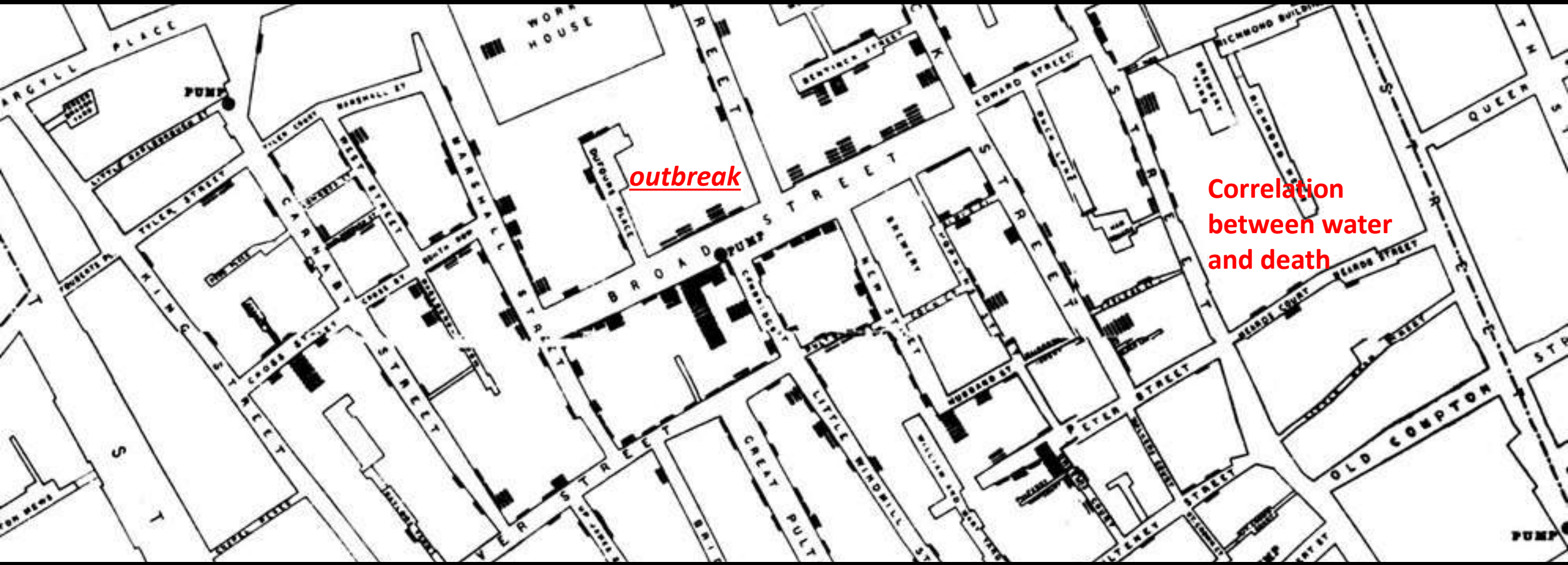

The Red Granite kerbstone
marks the site of the historic
BROAD STREET PUMP
associated with Dr. John Snow's
discovery in 1854
that Cholera is conveyed by water

London 1854

Within 3 weeks 127 died …why this disease is being spread



Stories are powerful. Cholera
Dr. John Snow told a *story* to prompt an action.



It's a myth that he used this map to convince people to remove the pump handle at Broad Street.

Story ? time, data, valgus (how many death),trend

Collect information
Who lives there?
When they died ?
Why they did not die?
Why they died?

Evidence : Brewery and workhouse survived

The ways in which organizations deliver business intelligence and analytics insights are evolving, notably in the rising use of what is called data storytelling.

This trend is an extension of the now dominant self-service model of Business Intelligence, combining explorative data visualization with narrative techniques to deliver insights in a way that engages with decision makers in a compelling and easily assimilated form.

What data storytelling is?
How it is evolving and how to best use it to go beyond reporting and dashboarding.

Key Issues

1. Beginning: What data storytelling is and why it matters
2. Middle: The how of data stories
3. End: Moderating the dark side of data narratives
   ➔Presenting complex data using a multidimensional aspects of visualization ➔discovery of correlation between water pump and the death of many people ➔prompt for an action

Story ➔Deliver information to get an action

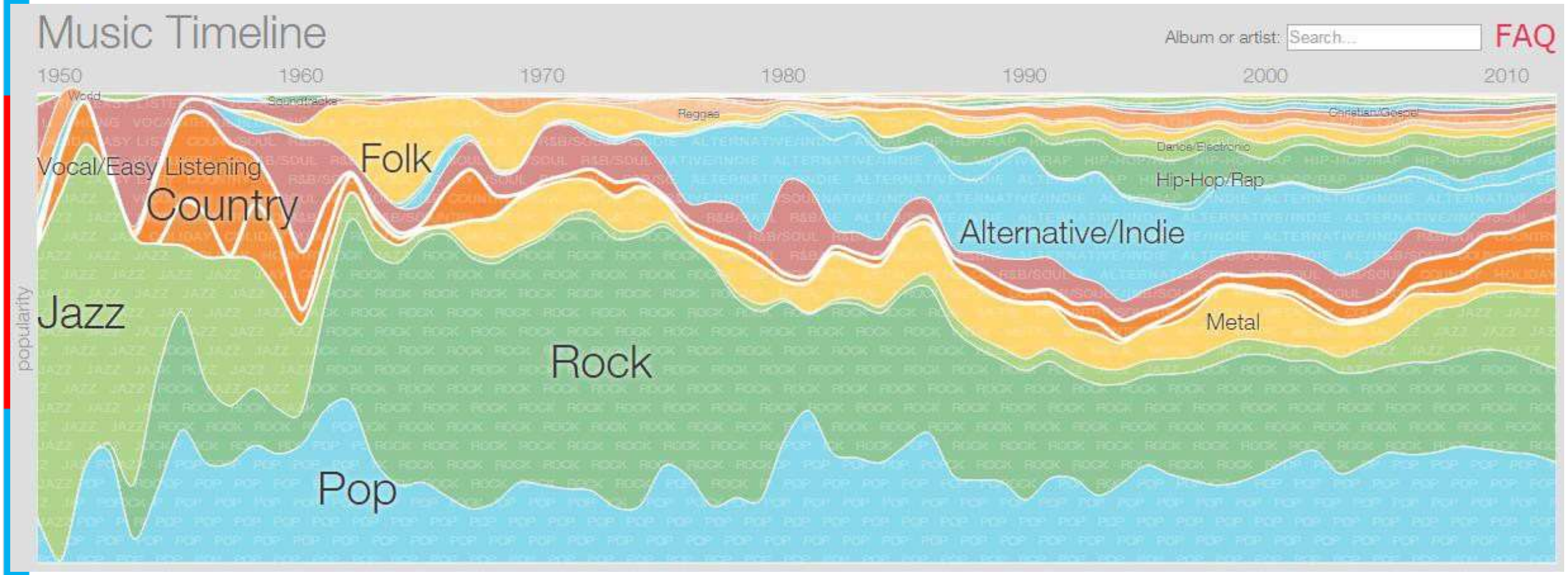
Key Issues

1. Beginning: What data storytelling is and why it matters
2. Middle: The how of data stories
3. End: Moderating the dark side of data narratives

Data Visualization Definition

- Interactive visualization enables the exploration of data via the manipulation of a battery of chart images, with the color, brightness, size, shape and motion of visual objects representing aspects of the dataset being analyzed.
- This capability enables people to analyze data by interacting directly with its visual representation.

Visuals Reveal Patterns, Trends, Changes and Clusters



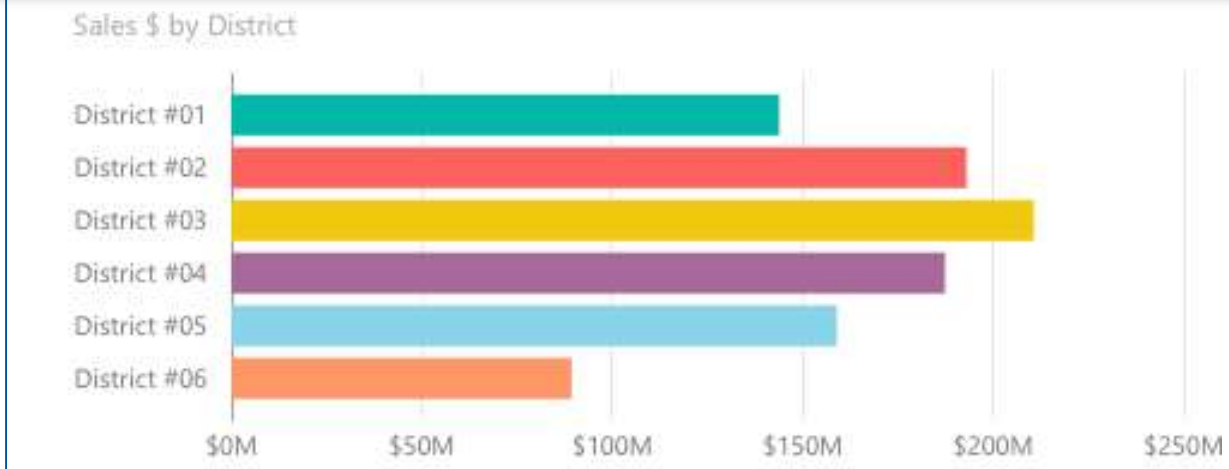
Source: http://research.google.com/bigpicture/music

Always the first question:

"What are we looking at?"

No visualization goes undescribed.

Natural Language Generation As Descriptor



Sales $ by District

This analysis measures Sales $ by District.

- Total Sales $ is $983.2 million across all six districts.
- The distribution ranges from $89.3 million (District #06) to $210.7 million (District #03), a difference of $121.3 million.
- The average Sales $ per district is $163.9 million and the median is $173.2 million.
- Sales $ is relatively evenly distributed across all the districts.
- The top two districts represent over a quarter (41%) of overall Sales $, and the top three districts account for over a half (60%).

Automated machine learning Descriptor analyzing the chart =>the data story telling is not necessarily the data itself but things around the data

Image source: http://resources.narrativescience.com/h/i/244761001-announcing-a-new-collaboration-between-microsoft-and-narrative-science

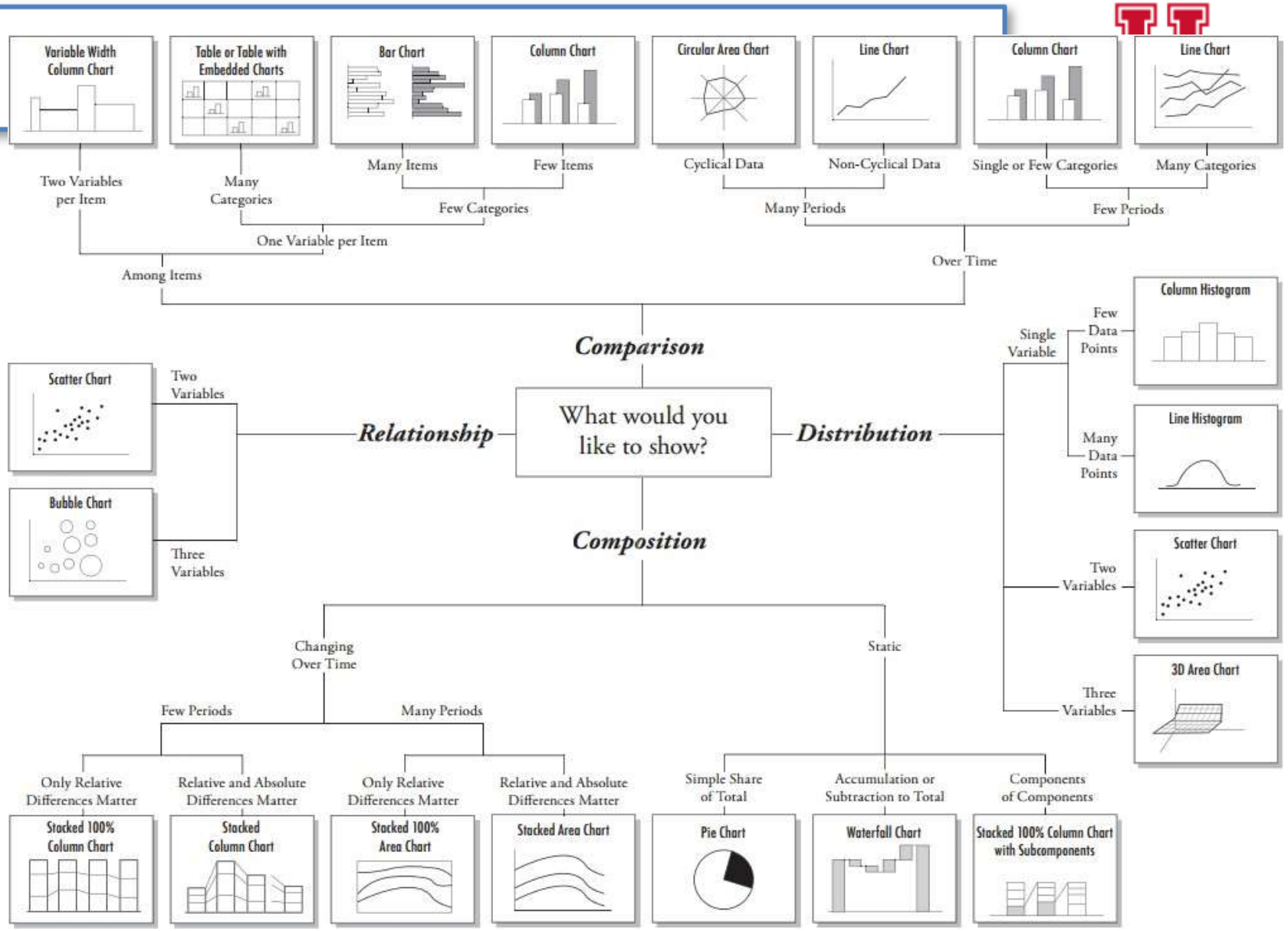
Narration Makes Sense of Visualization via Context

Data Storytelling <> Data Visualization

Visualization is one aspect of data stories.

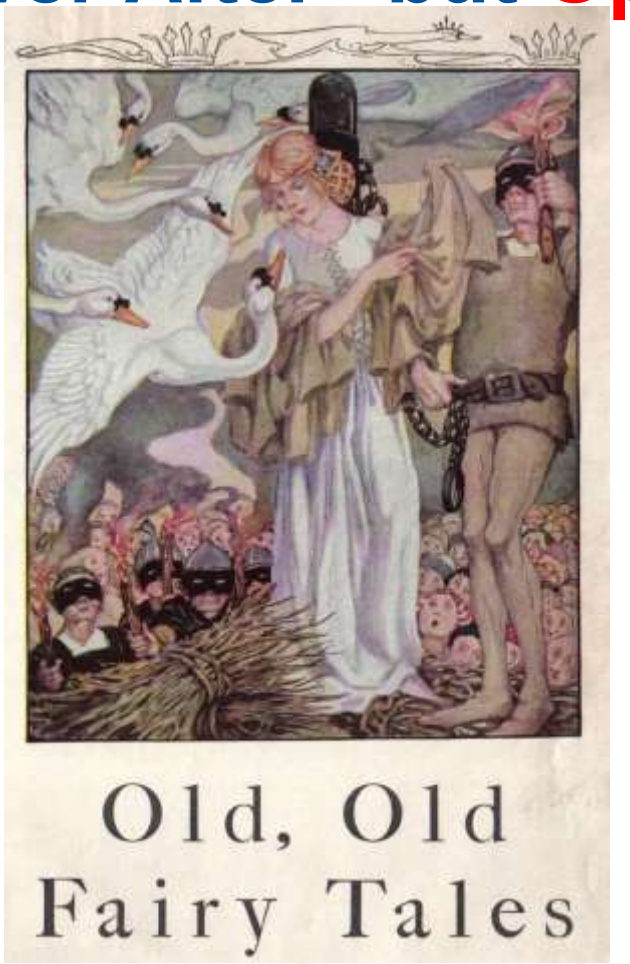
They work well together

Why? Only Some Visualizations Include Time



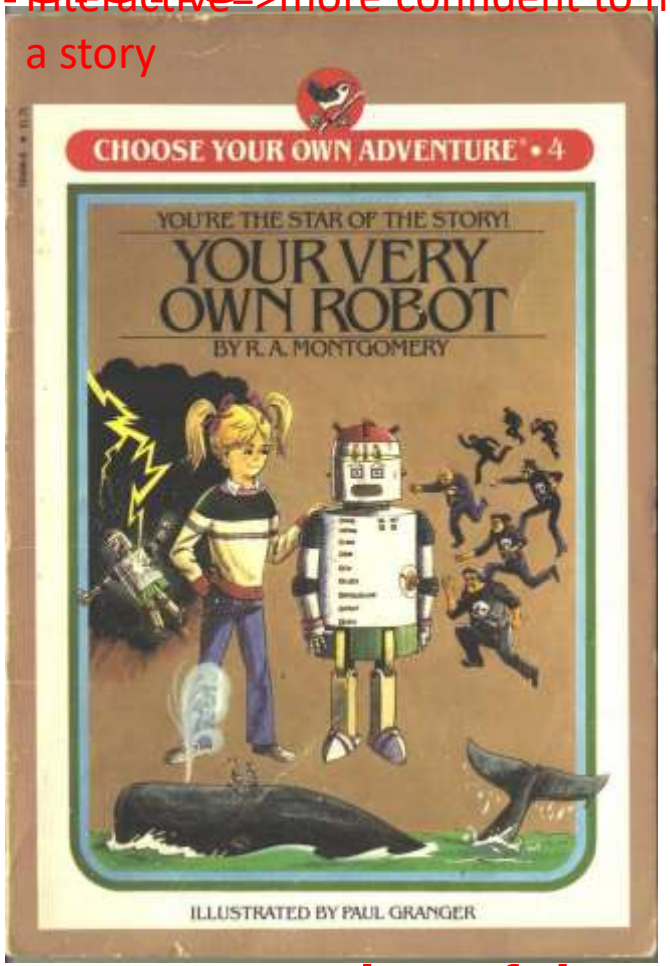
Source: http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

No Fairy Tales: Data Stories Aren't About "Happy Ever After" but Options and decisions



VS.

Interactive=>more confident to make a story



data story is not about the passive compensation of data (representation) =>it is interaction between audience and the data to get them closest to make decisions

Why we do Business Intelligence, Business Analytics and data science ?
➔is all about how to make decision from packaged data ?

A journalistic data story is a complete linear narrative to be passively consumed for information or entertainment.

An analytic data story is an unfinished, working narrative to be actively explored and questioned collectively to aid in decision making.

➔analytical data story(hypothesis put together for more exploration and decision)

Data Storytelling =

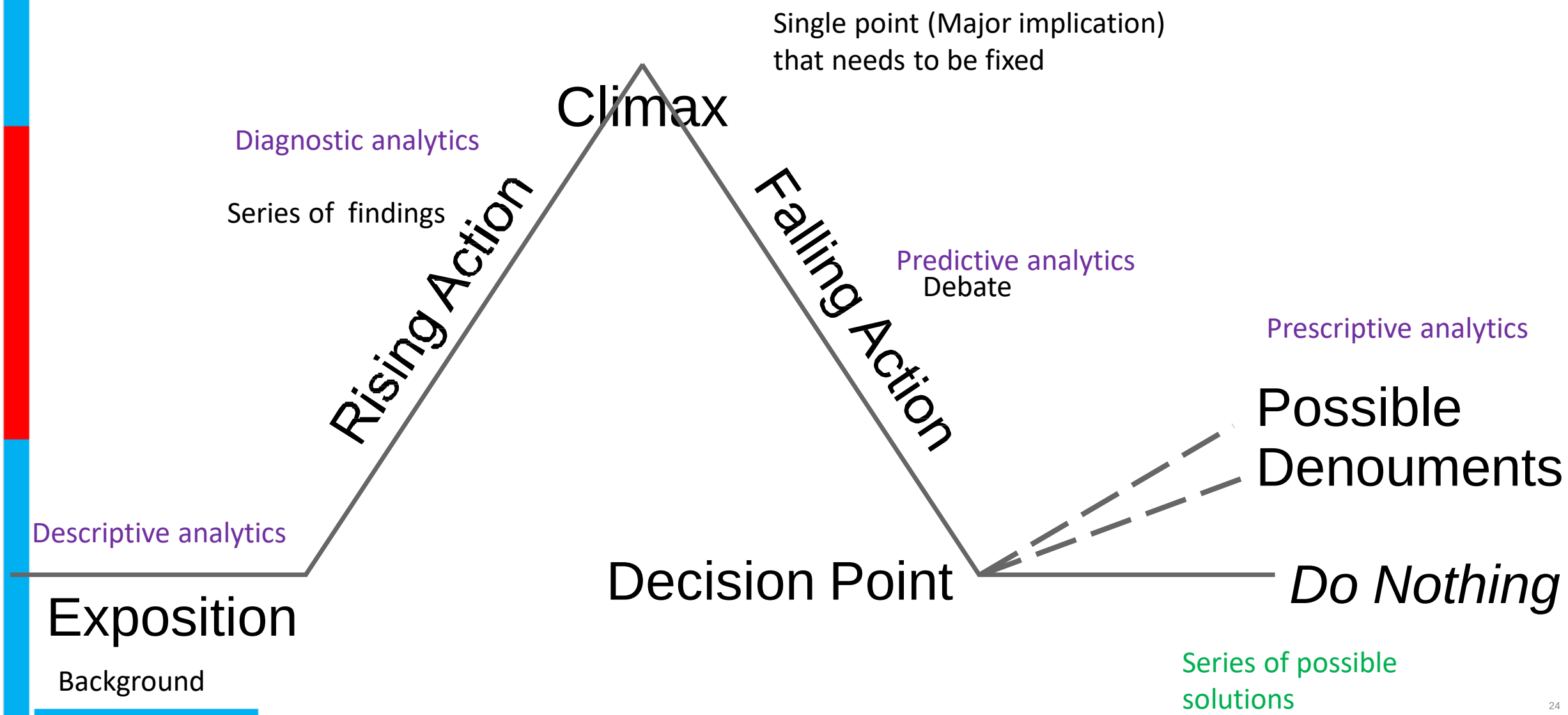
Visualization + Narrative + Context

Narrative describing the visualization (what is the audience is looking at) and the context what other things around the data that drives the trend based on human experience

Key Issues

1. Beginning: What data storytelling is and why it matters
2. Middle: The how of data stories
3. End: Moderating the dark side of data narratives

Stick to the Plot — Apply Freytag's Pyramid to Data Stories

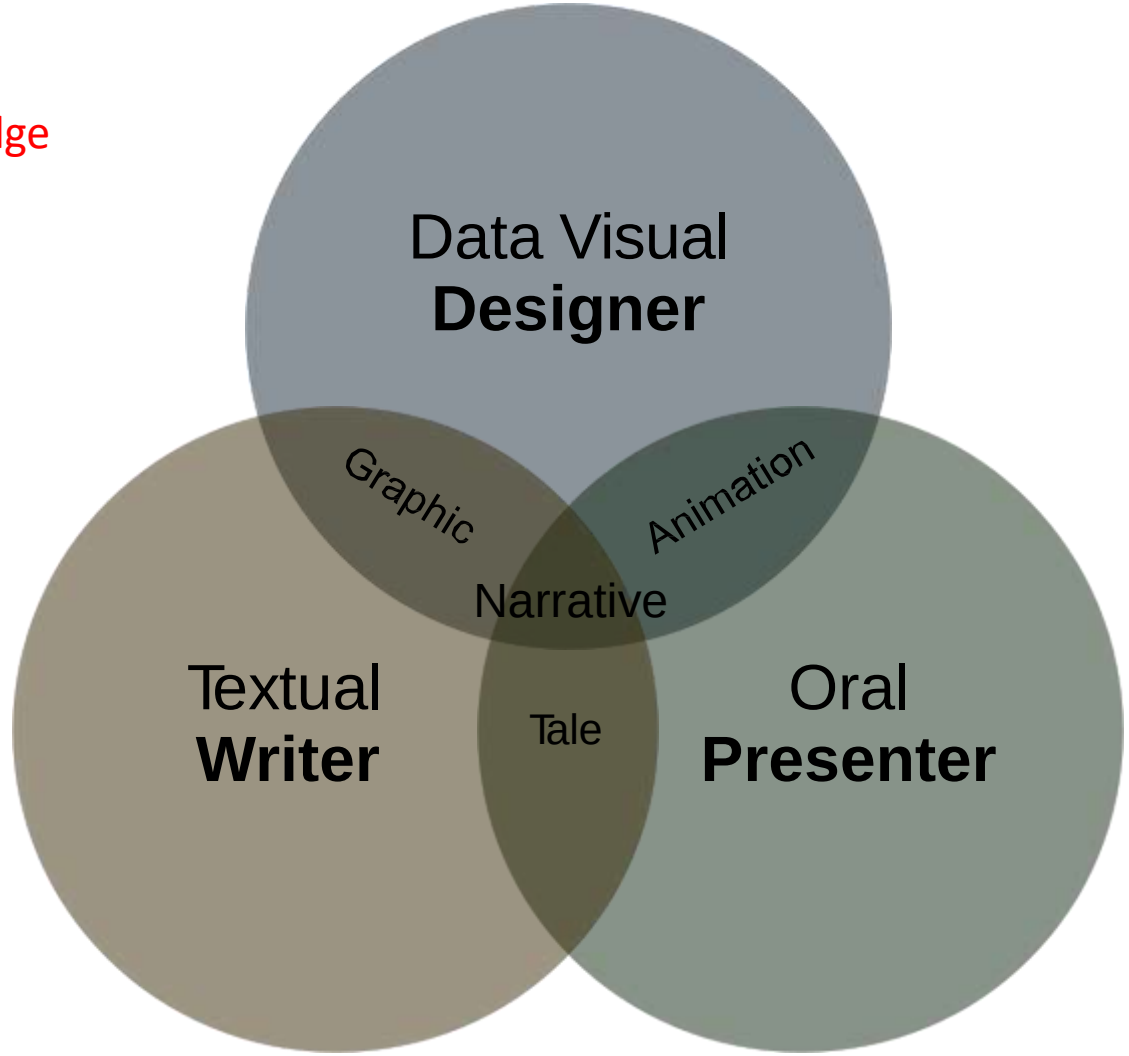


Tell Stories About Data People Care About

Hire or Teach Data Storytelling's Combined Skills

Designer (no knowledge of major impact)

Textual (data story is edited (analyst who like to talk and focus more on the core message to deliver)

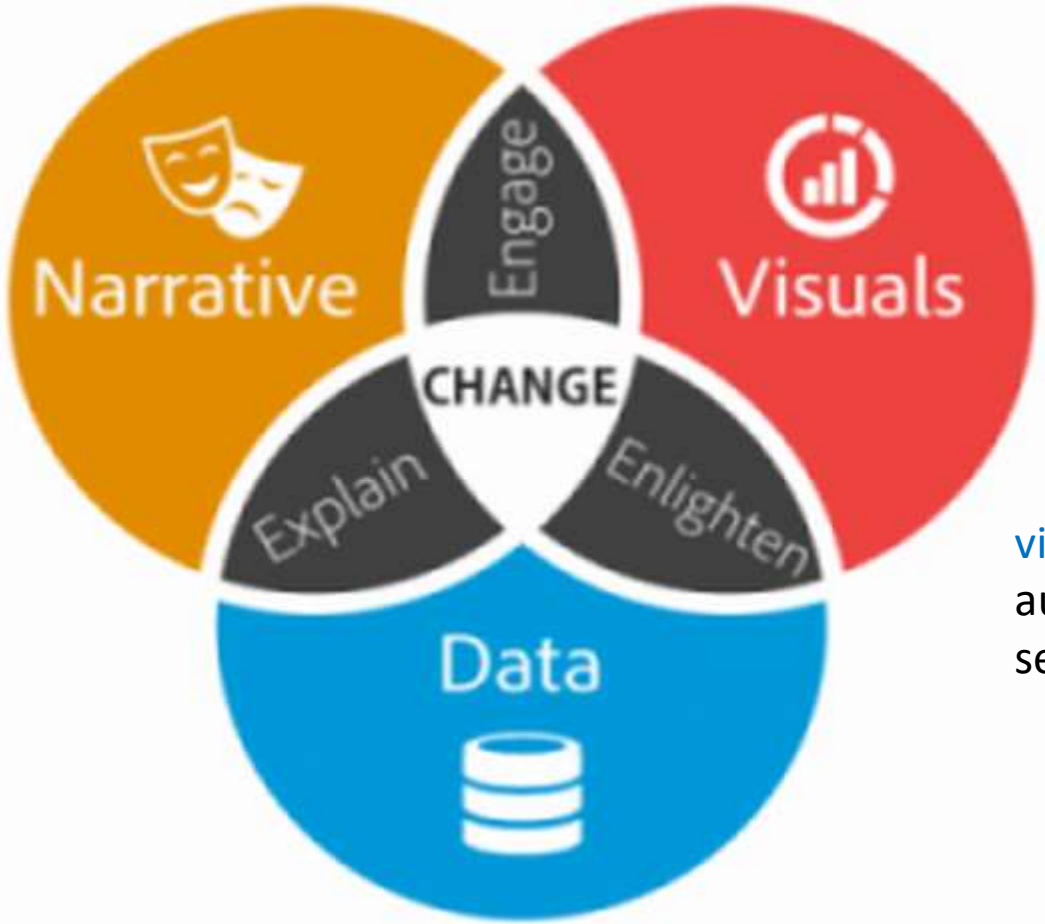


Right presenter with the right profile

Data storytelling is a structured approach for communicating data insights, and it involves a combination of three key elements: *data*, *visuals*, and *narrative*.

narrative and visuals are merged together, they can **engage** or even entertain an audience.

narrative +data, it helps to **explain** to your audience what's happening in the data and why a particular insight is important. Ample context and commentary is often needed to fully appreciate an insight. visualizations.



visuals +data, they can ***enlighten*** the audience to insights that they wouldn't see without charts or graphs.

Combining the right visuals and narrative with the right data, a data story that can ***influence and drive change***.

Why data storytelling is essential?

When you package up your insights as a data story, you build a bridge for your data to the influential, emotional side of the brain. When neuroscientists observed the effects detailed information had on an audience, brain scans revealed it only activated two areas of the brain associated with language processing: Broca's area and Wernicke's area. However, when someone is absorbed in a story, they discovered it stimulated more areas of the brain

Memorability: A study by Stanford professor Chip Heath (Made to Stick author) found 63% could remember stories, but only 5% could remember a single statistic.

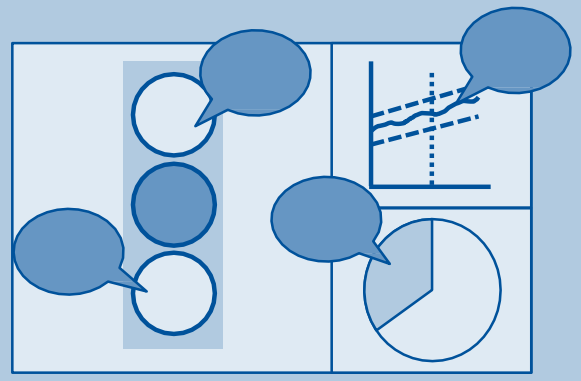
Persuasiveness: In another study, researchers tested two variations of a brochure for the Save the Children charity organization. The story-based version outperformed the infographic version by $2.38 to $1.14 in terms of per participant donations.

Engagement: Researchers discovered people enter into a trance-like state, where they drop their intellectual guard and are less critical and skeptical. Rather than nitpicking over the details, the audience wants to see where the story leads them.


THE 8 COMMANDMENTS OF STORYTELLING WITH DATA
1
BEGIN WITH A QUESTION
Set up your story. What is your audience going to learn?
2
END WITH AN INSIGHT
If we can't learn something useful from the data, the story isn't worth telling.
3
TELL A COMPELLING STORY
People remember stories, not data. Take them on your journey.
4
EXPLAIN WITH VISUALS, NARRATE WITH WORDS
People understand metrics, trends and patterns better with visuals. Use words to add your voice to the data.
5
BE HONEST AND CREDIBLE
The clients we want value honesty. Don't sugarcoat the negatives.
6
BE CLEAR AND CONCISE
Remove everything that is not part of your story. Save the good bits for another time.
7
KNOW AND CATER TO YOUR AUDIENCE
What are their interests and goals? Do they want the details, or just the high-level summary?
8
PROVIDE CONTEXT
Compare metrics over time or to industry benchmarks. Numbers are meaningless without context.
twoOctobers

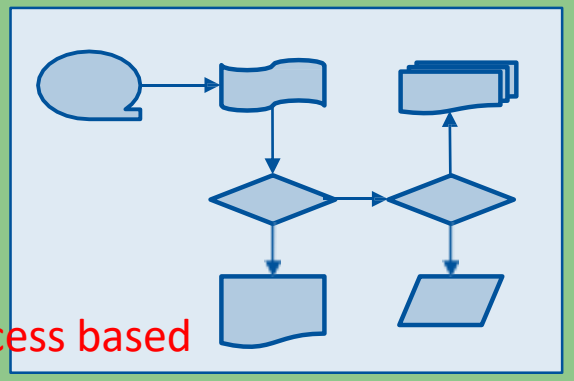
Select the Data Storytelling Form to Fit the Narrative

Annotated Dashboard



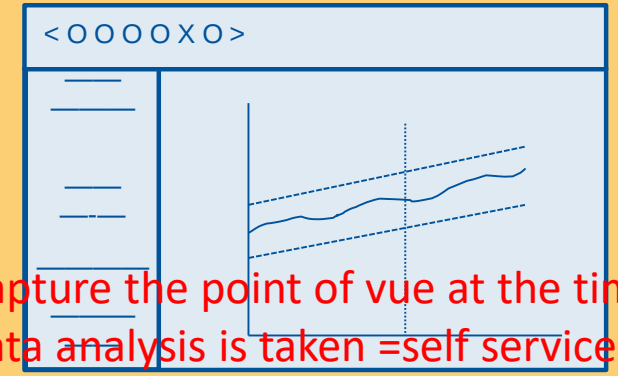
Group of people to capture the need of change (audience is the manager)

Decision Tree/Flowchart



Process based

Connected Slide Show

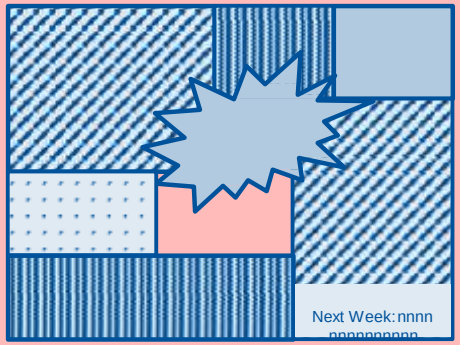


Capture the point of vue at the time data analysis is taken =self service BI

Infographic

Informative icons

Storyboard



Panel +flow of data

Example: Annotated Dashboard Data Story

Source: Decisive Data

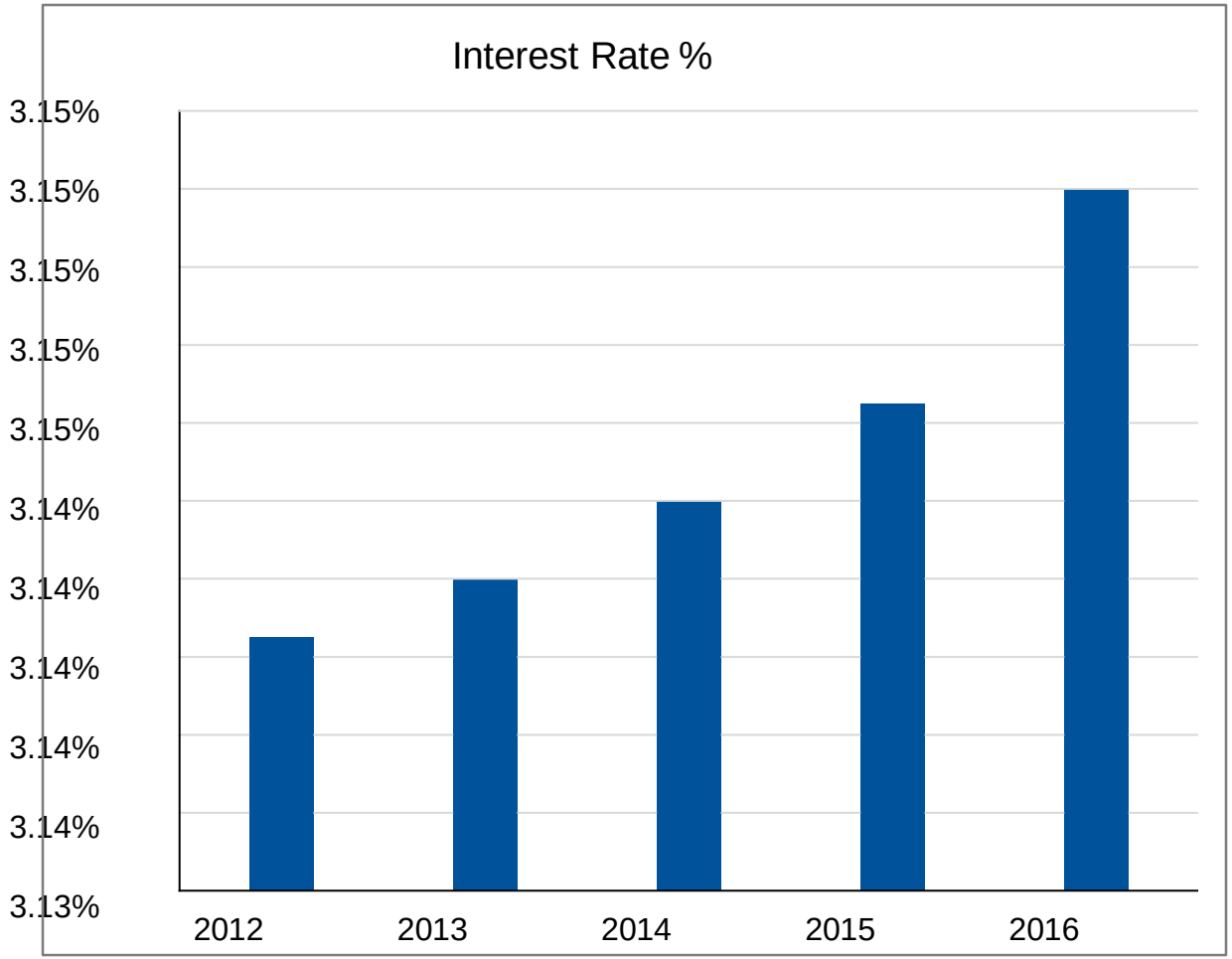
Key Issues

1. Beginning: What data storytelling is and why it matters
2. Middle: The how of data stories
3. End: Moderating the dark side of data narratives

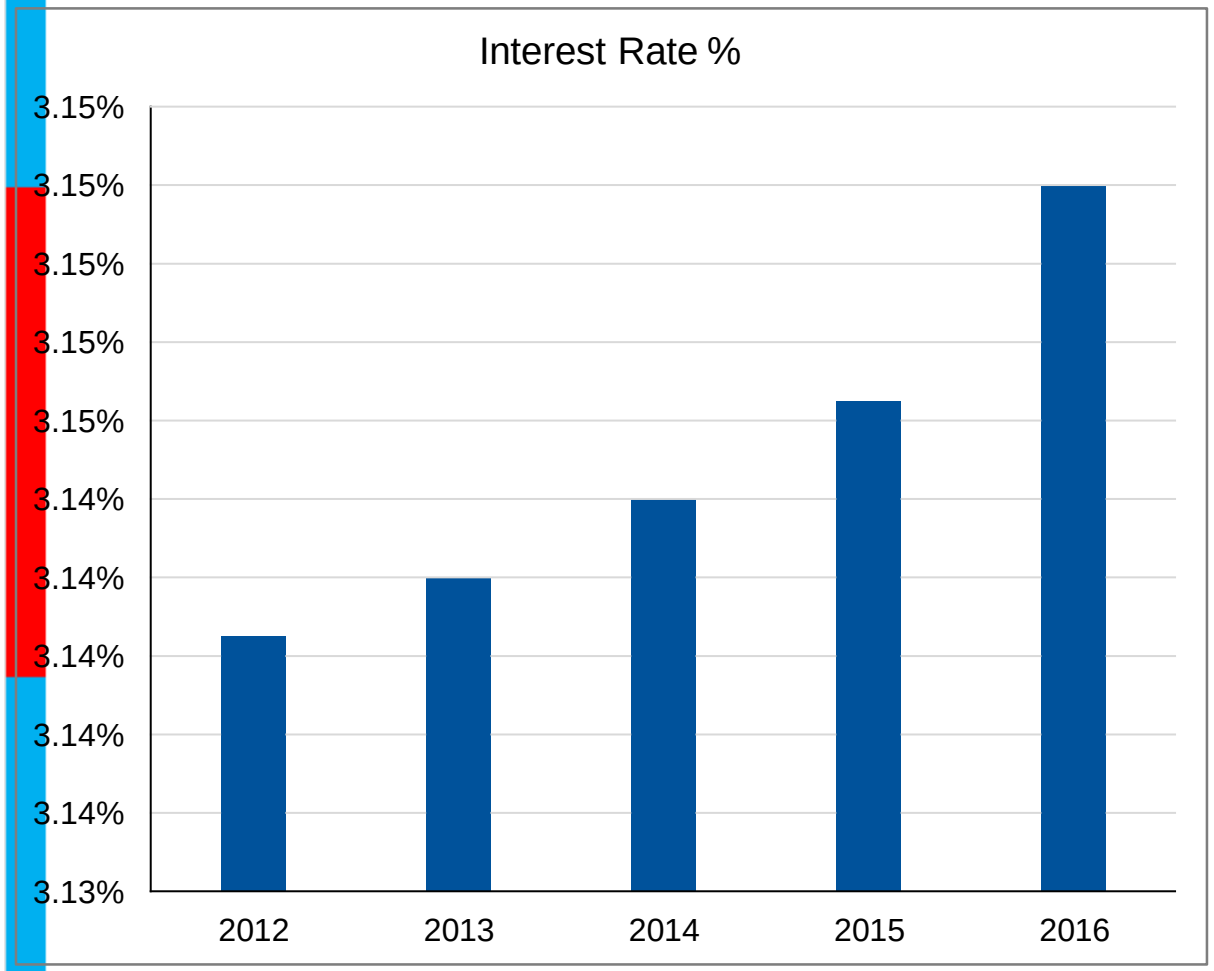
Some data narratives are unreliable.

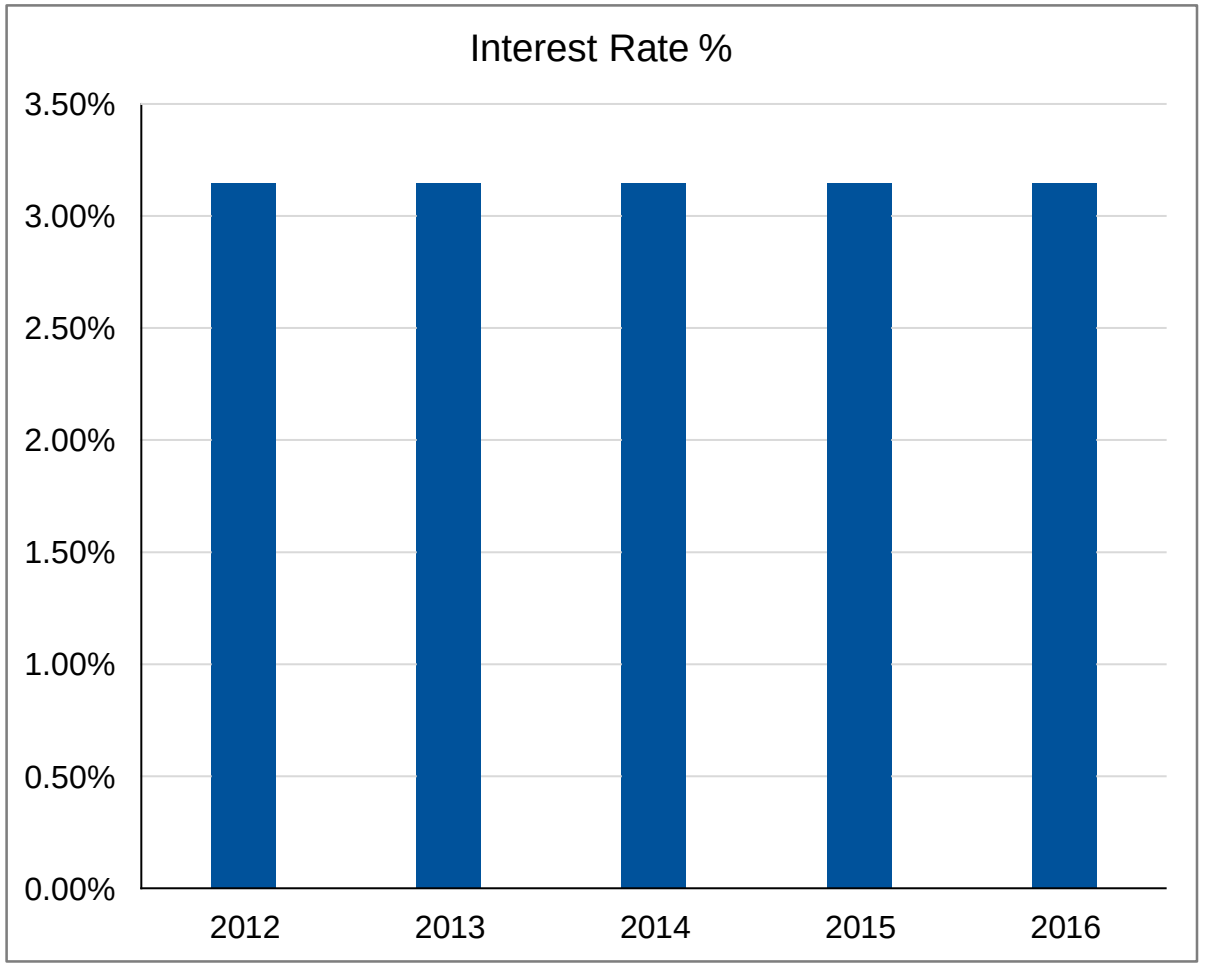
Humans can't help it. We're made that way.

Unreliable Visual Narrative

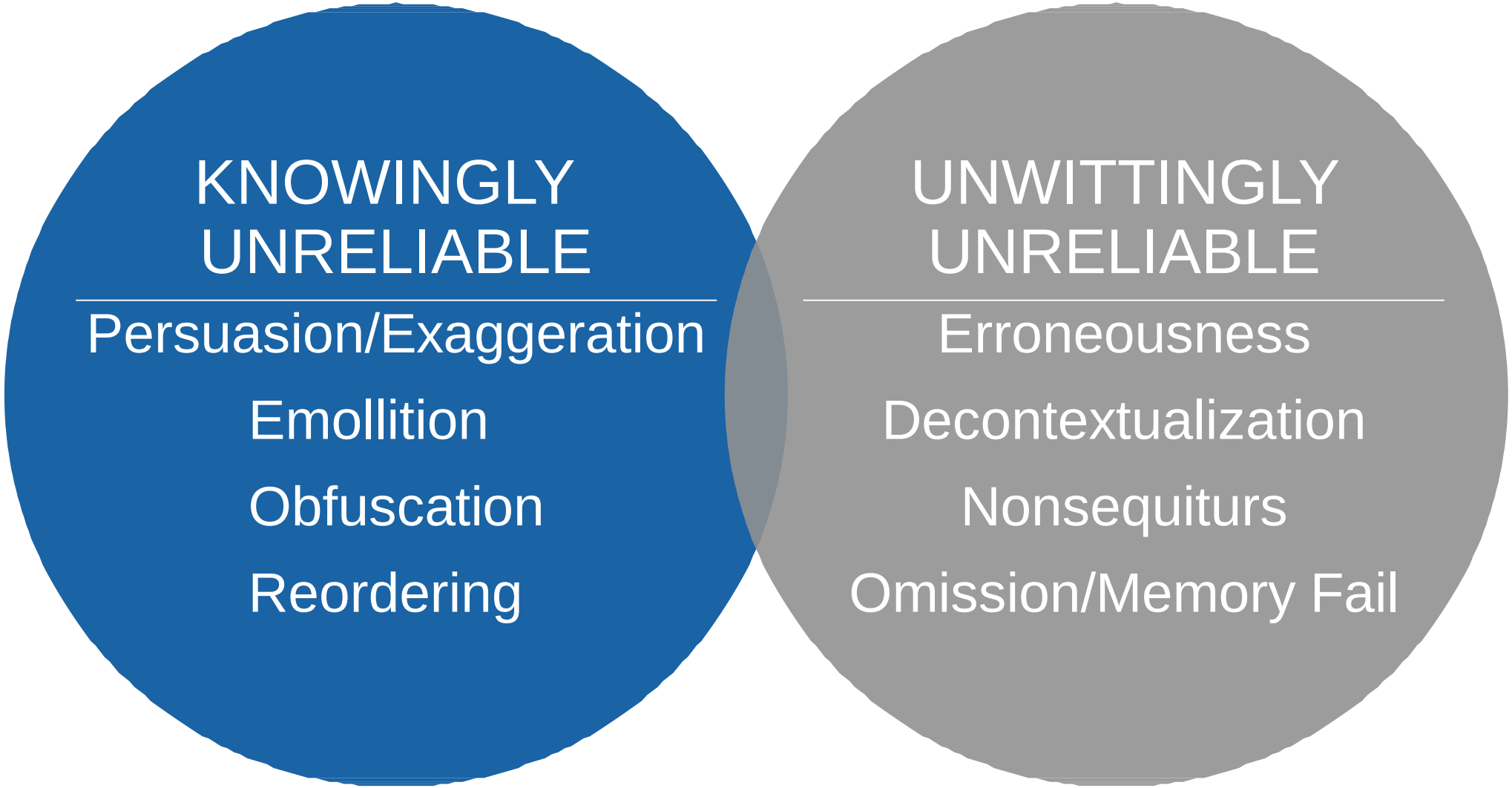


Unreliable Visual Narrative



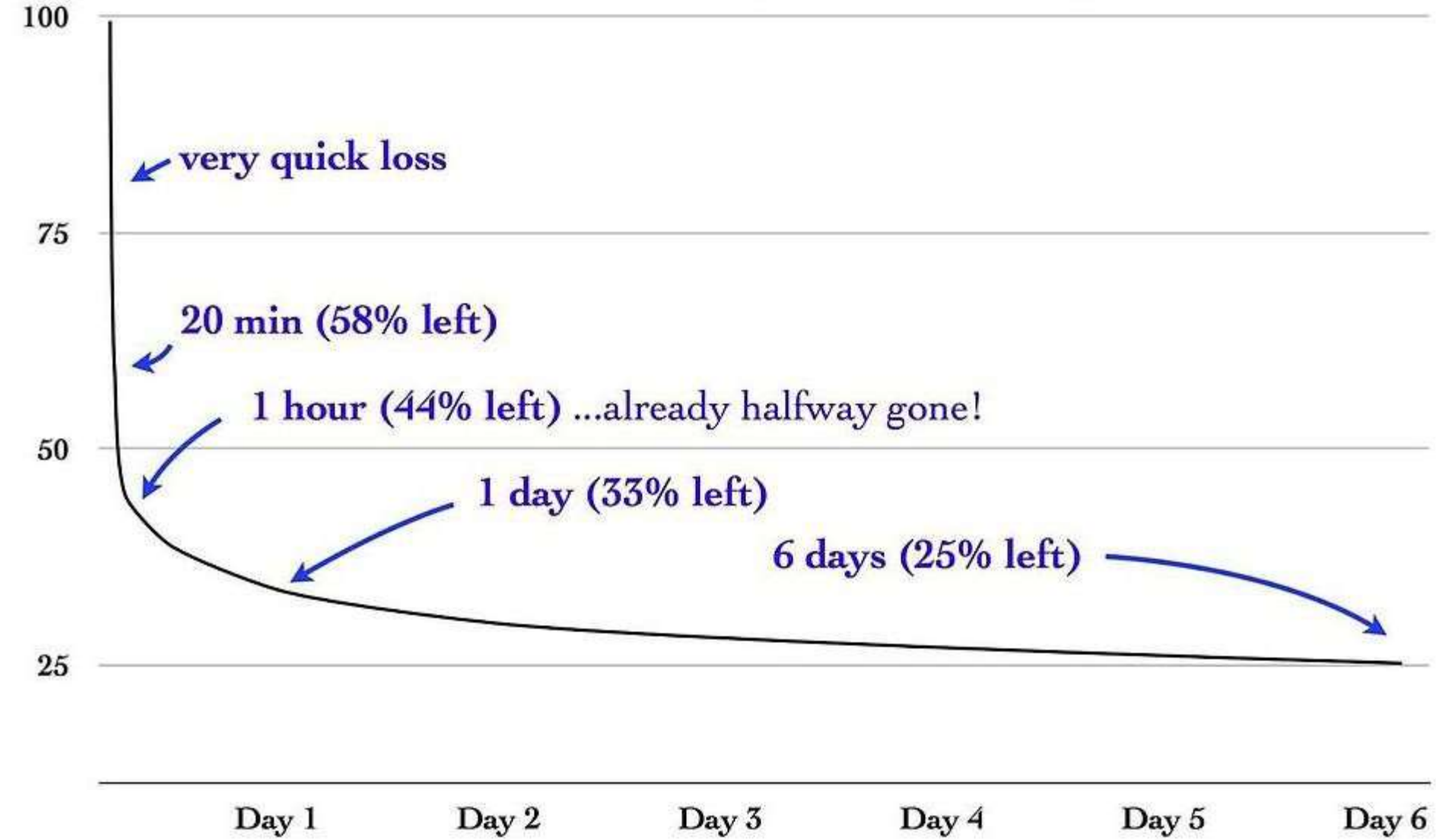


Unreliable Narration Factors Impacting Data Stories



Ebbinghaus' Forgetting Curve

(How much of something do we forget each day?)



Unreliable Narration Impacts Decision Making

A "narrator who may be in error in his understanding or report of things and who thus *leaves readers without the guides needed for making judgments.*"

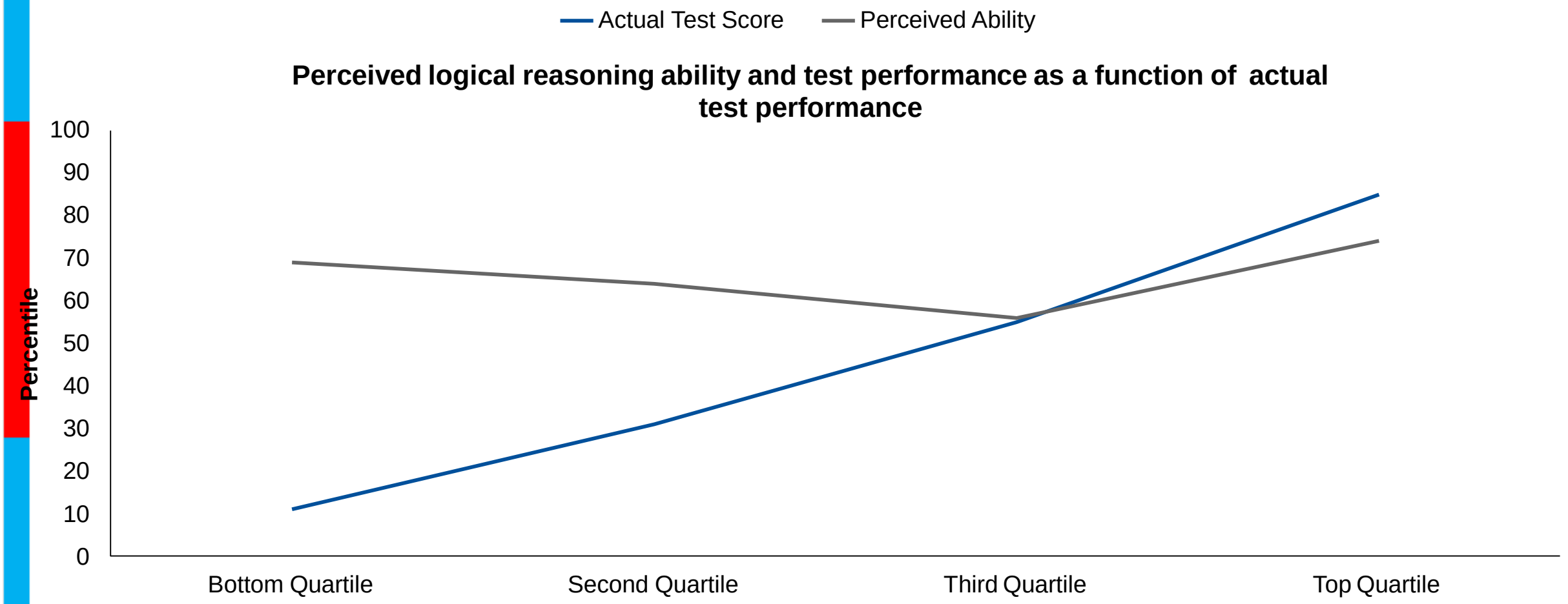
W. Harmon, "A Handbook to Literature (Tenth Edition)," 2006

Is the Data Story Reliable?
Use Carl Sagan's "Baloney Detection Kit"

1. "Wherever possible there must be independent confirmation of the 'facts.'"
2. Encourage substantive debate on the evidence by knowledgeable proponents of all points of view.
3. Arguments from authority carry little weight — "authorities" have made mistakes in the past.
4. Spin more than one hypothesis.
5. Try not to get overly attached to a hypothesis just because it's yours.
6. Quantify.
7. If there's a chain of argument, every link in the chain must work …
8. Occam's razor … when faced with two hypotheses that explain the data equally well, to choose the simpler.
9. Always ask whether the hypothesis can be, at least in principle, falsified."

Quoted from "The Demon-Haunted World: Science as a Candle in the Dark"

People Overate Their Abilities: Debating Data Stories Helps



Note: Chart redrawn from the original in "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments" (1999) by Justin Kruger and David Dunning.

Future of Data Stories?
Machine(s) Learning to Tell Disinterested Stories Already

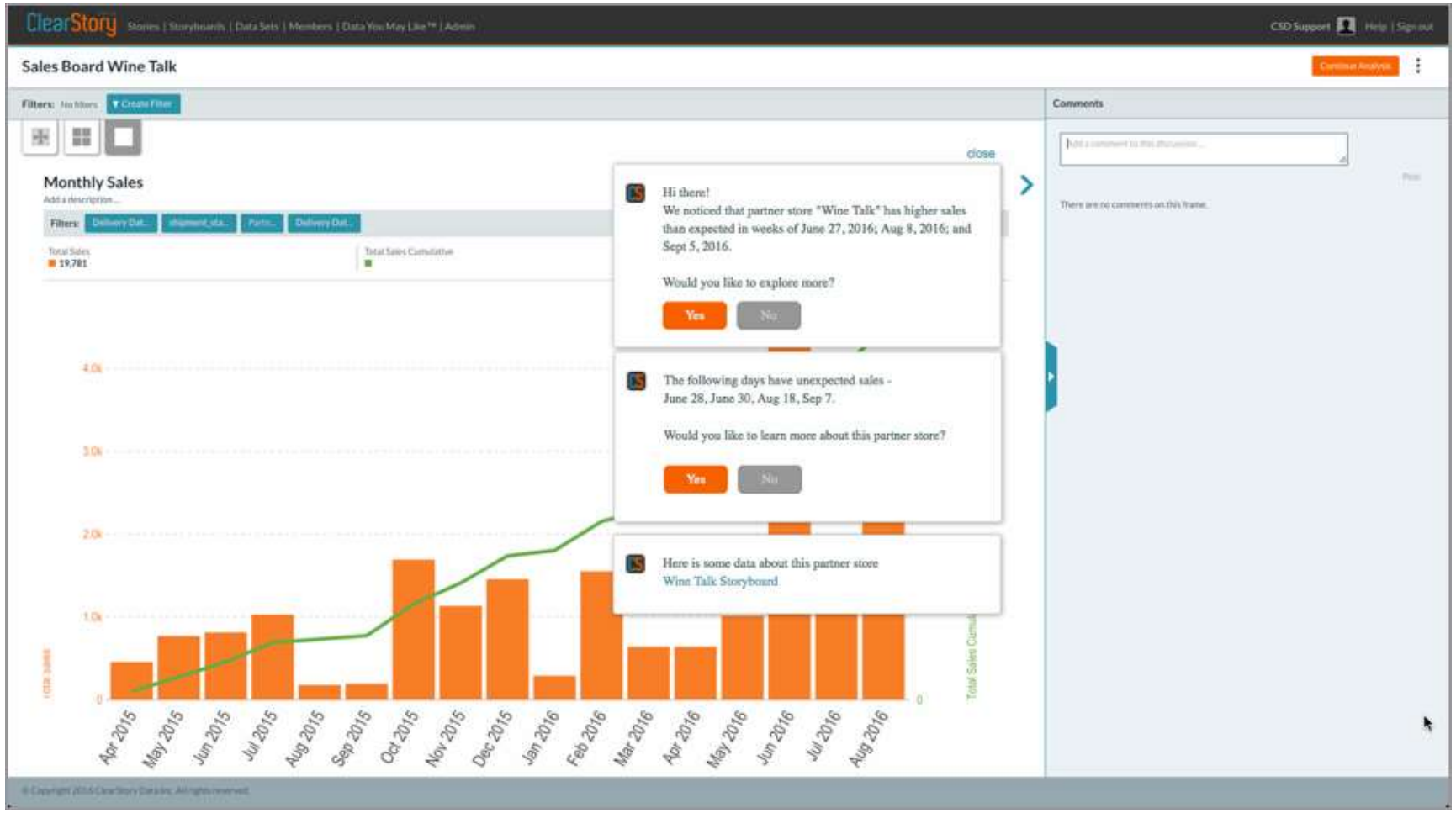


Image courtesy of ClearStory Data

Denouement

"No one ever made a decision because of a number. They need a story."

Daniel Kahneman, Quoted in Vanity Fair article "How Two Trailblazing Psychologists Turned the World of Decision Science Upside Down," November 2016

Recommendations

- ✓ Investigate data storytelling to advise business leaders and users on applying narrative techniques to data to help communicate its importance to decision makers.
- ✓ Evaluate and experiment with the data storytelling capabilities of modern business intelligence platforms.
- ✓ Prepare programs to develop and instill narrative skills for business analysts and data-savvy end users, including a virtual team of certified data storytellers.
- ✓ Appraise your organization's managerial and decision-making culture as to its fitness for accepting data as stories.

Example

Attendance across marking periods

Explore how different student factors relate to attendance.
Specific questions:

1. Do absences and/or tardies vary across marking periods?

2. Are there particular "at risk" student populations?

3. Are the same students at risk for the different types of attendance issues (unexcused absences, excused absences, and tardies)?

```
#import the relevant libraries

import pandas as pd
import numpy as np
import hypertools as hyp
import seaborn as sns
%matplotlib inline
```

```
#load in the data

fname1 = 'attendance_data_16-17.csv'
fname2 = 'attendance_data_17-18.csv'

columns = ('id', 'grade', 'age', 'school',
'sex', 'homeless', 'disadvantaged',
'specialneeds',
            'excused1', 'unexcused1', 'tardy1',
            'excused2', 'unexcused2', 'tardy2',
            'excused3', 'unexcused3', 'tardy3',
            'excused4', 'unexcused4', 'tardy4')

y1_data = pd.read_csv(fname1, skiprows=[0],
names=columns)
y2_data = pd.read_csv(fname2, skiprows=[0],
names=columns)

#use student IDs as the index
y1_data.set_index('id', inplace=True)
y2_data.set_index('id', inplace=True)
```

```
#do some data cleaning

#in "disadvantaged" column, replace "YES" with 1 and NaN with 0
y1_data['disadvantaged'] = y1_data['disadvantaged'].map({np.nan: 0, 'YES': 1})
y2_data['disadvantaged'] = y2_data['disadvantaged'].map({np.nan: 0, 'YES': 1})

#in "specialneeds" column,
y1_data['specialneeds'] = y1_data['specialneeds'].map({np.nan: 0, 504: '504', 'IEP':
'IEP'})
y2_data['specialneeds'] = y2_data['specialneeds'].map({np.nan: 0, 504: '504', 'IEP':
'IEP'})

#replace '---' with 0 (Fourth marking period columns)
y1_data.replace('---', 0, inplace=True)
y2_data.replace('---', 0, inplace=True)

#replace 'K' with 0 and all other non-number grades with -1
y1_data['grade']= y1_data['grade'].map({'K': 0, 'PD': -1, 'PA': -1, 'PP': -1, 'AW':
-1, '1':1, '2':2, '3':3, '4':4, '5':5, '6':6, '7':7, '8':8, '9':9, '10':10, '11':11,
'12':12})
y2_data['grade']= y2_data['grade'].map({'K': 0, 'PD': -1, 'PA': -1, 'PP': -1, 'AW':
-1, '1':1, '2':2, '3':3, '4':4, '5':5, '6':6, '7':7, '8':8, '9':9, '10':10, '11':11,
'12':12})
```

HyperTools plot of the dataset

We'll define an "attendance score" at the total number of absences and tardies across all marking periods. Then we'll color students according to their attendance score.

```
y1_data.head()
```

```
features = ['grade', 'age', 'sex', 'homeless', 'disadvantaged',
'specialneeds']
attendance_factors = ['excused1', 'excused2', 'excused3', 'excused4',
'unexcused1', 'unexcused2', 'unexcused3', 'unexcused4', 'tardy1',
'tardy2', 'tardy3', 'tardy4']
attendance_score = y1_data[attendance_factors].sum(axis=1)

hyp.plot(y1_data[features], '.', group=attendance_score, model='TSNE');
```

```
#create a dataframe just for the excused absences
y1_excused = y1_data[['excused1', 'excused2', 'excused3', 'excused4']]
y1_excused.head()
```

Does attendance vary across marking periods?

```
h = sns.factorplot(data=y1_excused,
kind='bar', color='k')
h.set_xlabels('Marking period')
h.set_xticklabels([1, 2, 3, 4])
h.set_ylabels('Average number of excused
absences')
```

```
statistics = ['unexcused', 'tardy']
names = ['unexcused absences', 'tardies']
marking_periods = [1, 2, 3, 4]
for x in np.arange(len(statistics)):
    columns = list(map(lambda j: statistics[x] + str(j),
    marking_periods))
    df = y1_data[columns]
    h = sns.factorplot(data=df, kind='bar', color='k')
    h.set_xlabels('Marking period')
    h.set_xticklabels([1, 2, 3, 4])
    h.set_ylabels('Average number of ' + names[x])
    sns.plt.show()
```

Summary : There are clear differences in attendance across marking periods.

Key (tentative findings):

1."Soft" attendance issues (excused absences and tardies) peak in the 3rd marking period (March -- June).

2."Hard" attendance issues (unexcused absences) rise throughout the year (dropout?)

Data story ……

1. How to gather data
2. How to identify interesting stories in data
3. How to curate assets (icons, photos) to go with the data and message
4. How to use data visualization tools (charts and graphs) for your chosen story
5. How to use design elements (such as colors or layout)
6. How to provide references that bring about credibility

1. How to gather data

a. Public Free resources

US Census Bureau: A large resource on US citizens which covers population data, geographic data, and education.

European Union Open Data Portal: Another place to explore government data, but it is based on data gleaned from institutions in the EU of course.

Datacatalogs.org: Yet another open government data resource, except this one allows you to peer into data from the US, EU, Canada, and more.

NHS Health and Social Care Info Centre: You can find health data sets from the UK National Health Service in this database.

Amazon Web Services Public Datasets: This is a huge resource of public data, which includes the 1000 Genome Project, as well as NASA's database of earth satellite imagery.

Google Finance: This is a database made up of 40 years' worth of stock market data, which Google updates in real time.

b. Internal data resources

Facebook Insights : A tool for marketers to track and analyze user interaction on their Facebook fan pages. It helps you determine the best time of day and week to post, and to also figure out which types of content your audience would like to read. Also useful when it comes time to gathering data for your social media report on Facebook page performance.

Mailchimp Analytics – A tool for marketers to dive deeper into the performance of their email campaigns, and to also learn more about their readers. It comes in handy when it comes time to crunch some numbers for your email reports, and pairs nicely with the likes of Google Analytics if you want to do more digging.

Google Analytics – A digital analytics tool that you can use to analyze data from various touchpoints, whether it's from the traffic that you're getting to the site in terms of numbers and geography, to zeroing in to the way that people are interacting with it. It is a great help in offering a deeper understanding of the customer experience and related behaviors, and it's easy to share insights with your team or client.

2. How to identify interesting stories in data

Develop a story structure and central theme

a. **Identify Trends:** For example, are people using their PCs less in favor of tablets and smartphones? Is there a growth in online shopping behavior in a certain country or region?

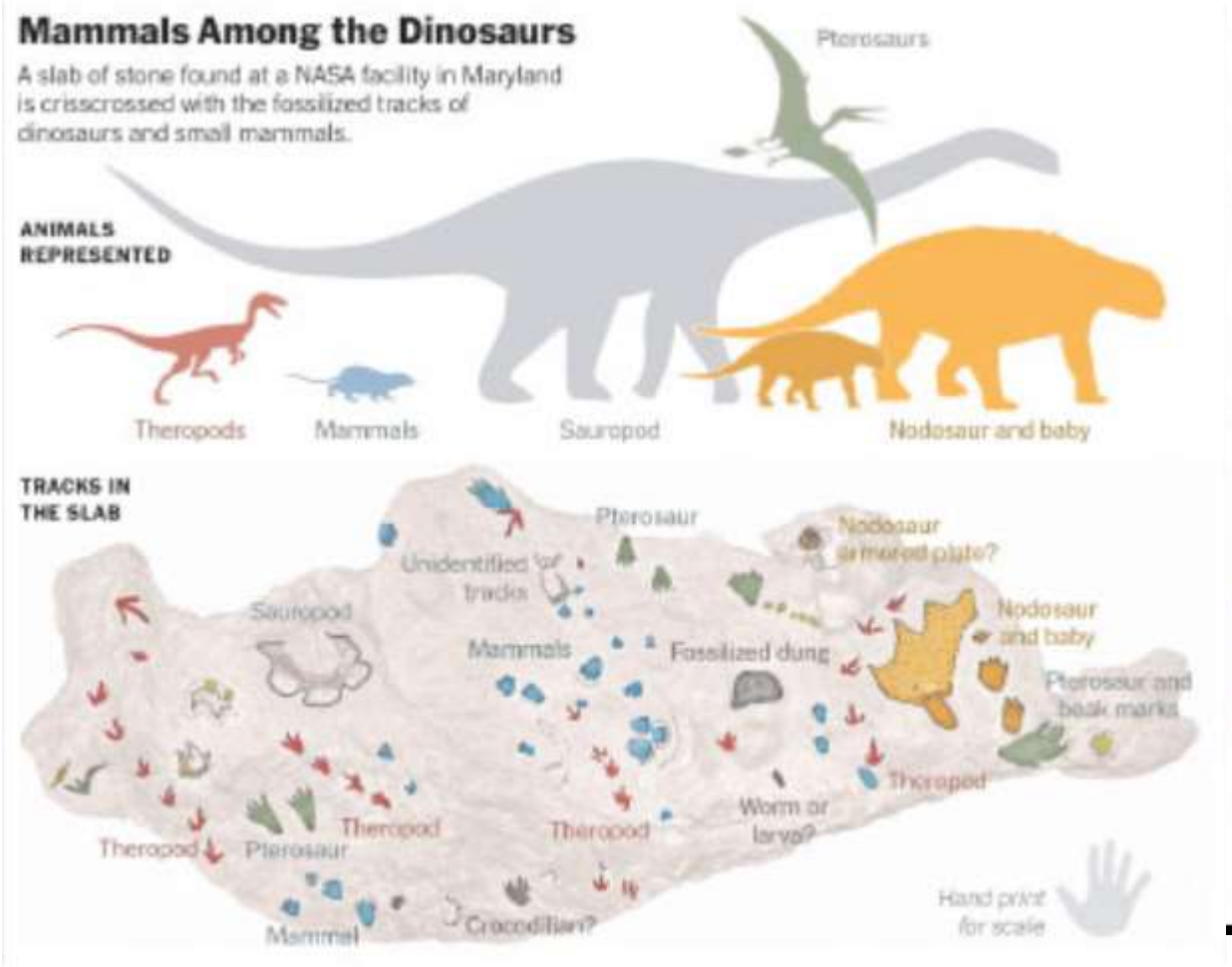
b. **Using rankings:** For example, is Vienna consistently being ranked number one in the list of most livable cities, while Vancouver is in second place? Are there certain areas in the United States that have higher crime rates than others? Rankings tell a story using data about the relationship between items on a list.

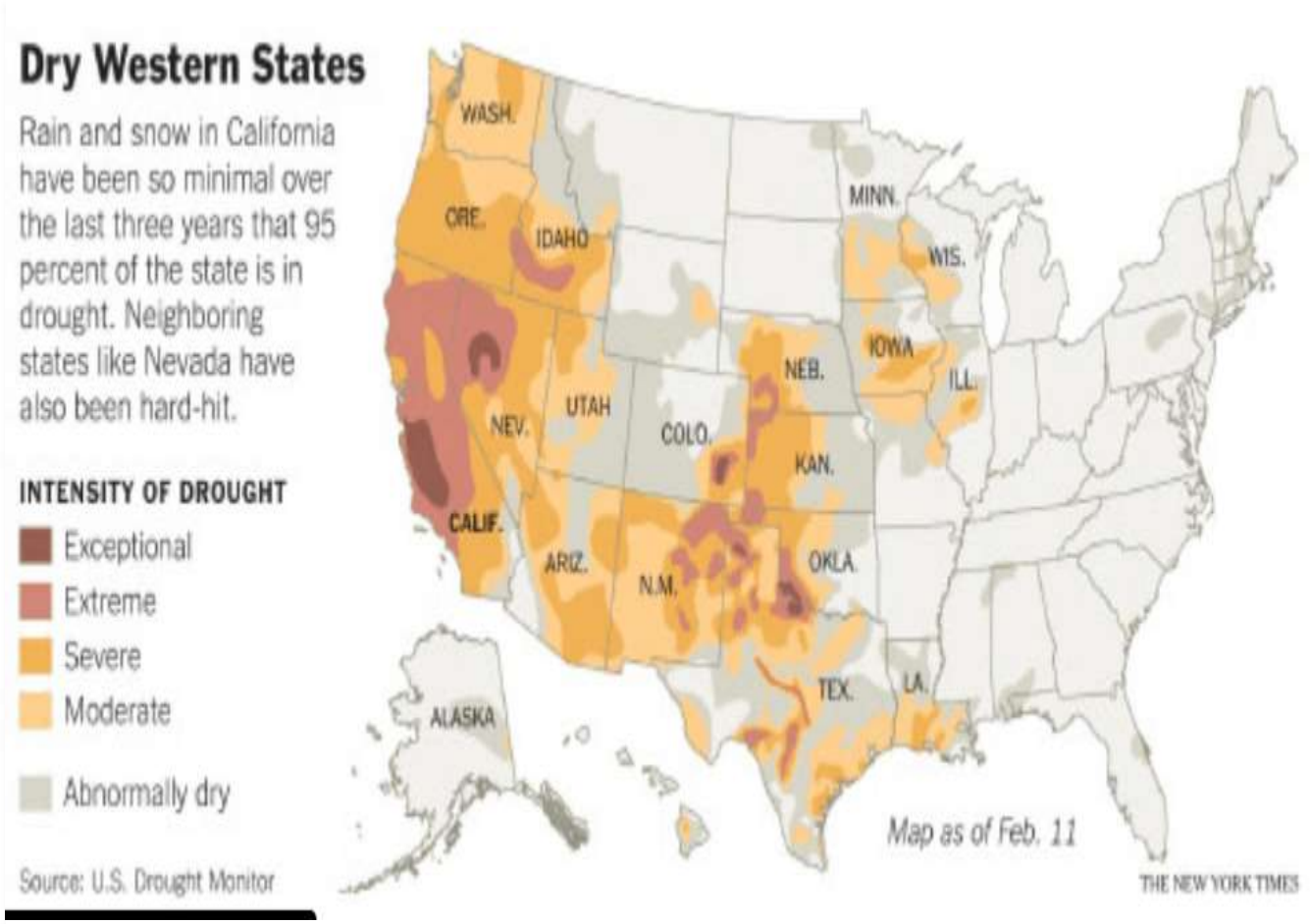
c. **Draw comparisons:** For example, how is Apple performing on the stock market in comparison to Microsoft? How much more dedicated to work-life balance is France, compared with Japan? Comparisons tell a side-by-side story between either polar opposites, or very similar things.

d. **Look for surprising or counterintuitive data:** For example, the sale of pop tarts apparently increased sevenfold before a hurricane. Seems surprising, but if you dig deeper – it seems that in anticipation for a natural disaster, people seek out comfort foods. Data that challenges previously confirmed knowledge of what people know to be true, tells a great story.

e. **Point out the relationship between data points:** For example, the influx of bitcoin mining companies moving to Canada is leading to rising costs in energy prices for local residents. Relationships between data points tell a story by showing a connection or correlation between a

3. How to curate assets (icons, photos) to go with the data and message





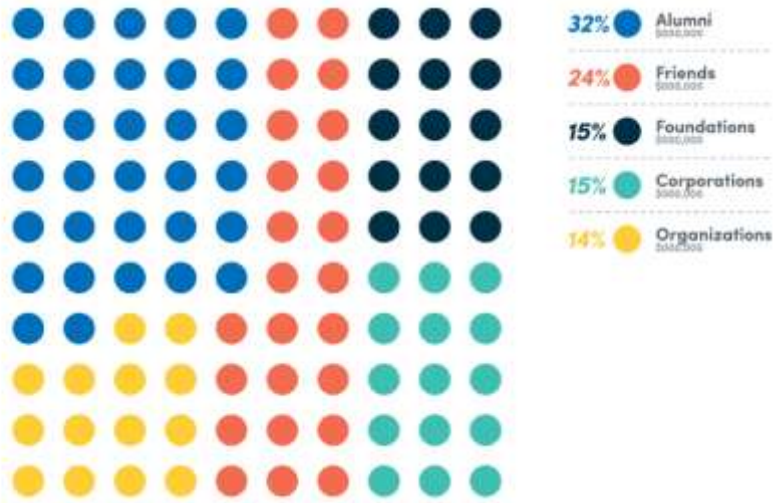

Land bank
70000
hectares
120000 m²
agricultural premises
1300
agricultural technical units
1200
sheep
2500
forage cows
7260 ha
crops & legumes
4840 ha
feeds
2530 ha
vegetables & perennial herbs


VISUAL
Storytelling
101
How to Create Beautiful Stories That Matter to Your Customers

4. How to use data visualization tools (charts and graphs) for your chosen story

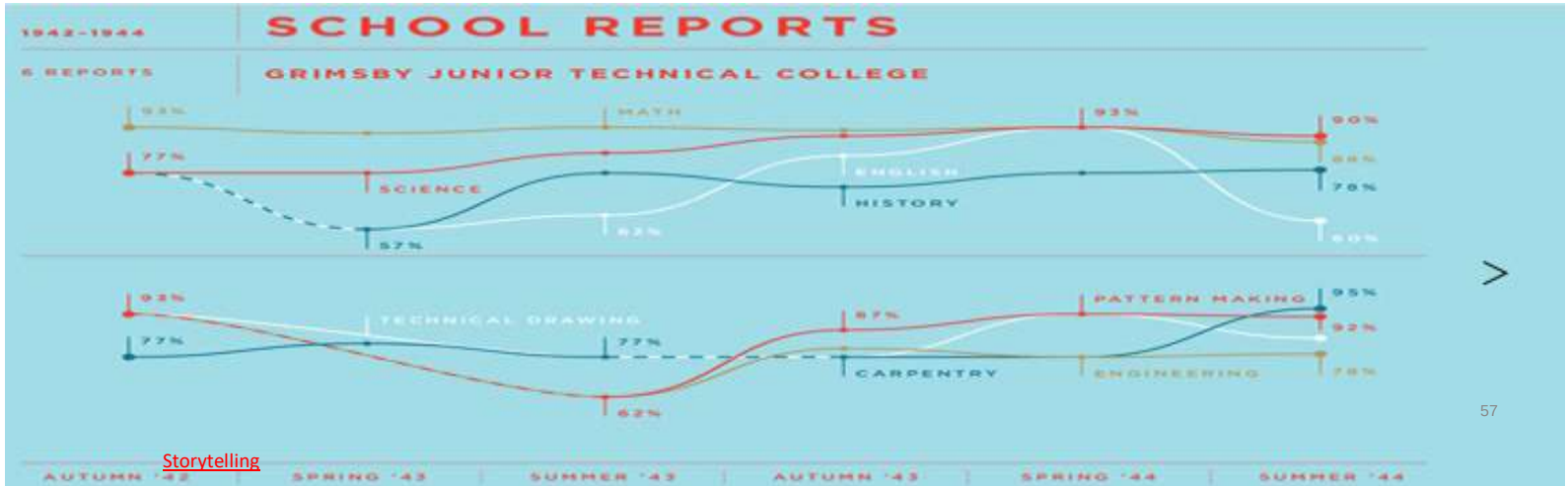
Dot Matrix Chart: Displaying data in units of dots, these charts can use a myriad of colors to represent a particular category which are grouped together in a matrix.

How can they be used? They can be used to give a quick overview of the distribution and proportions in data categories, and to also draw comparisons across other datasets – if you are seeking patterns.



Line graphs: Similar to bar graphs, line graphs can be used to track changes over short and long periods of time. They are typically more nimble than bar graphs and are helpful in representing when smaller changes exist.

How can they be used? Line graphs can be used to compare changes over the same period of time for more than one group. So for example, the performance of students at Grimsby Junior Technical College in various disciplines over a period of two years.



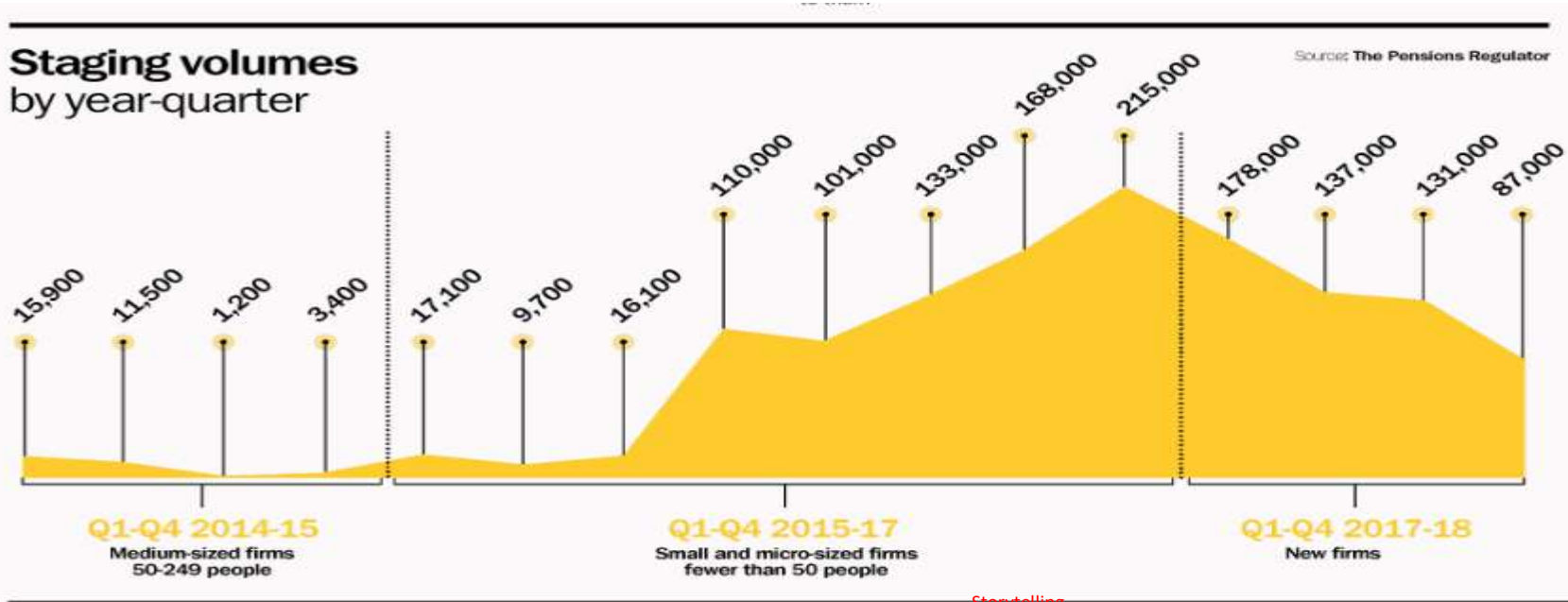
4. How to use data visualization tools (charts and graphs) for your chosen story

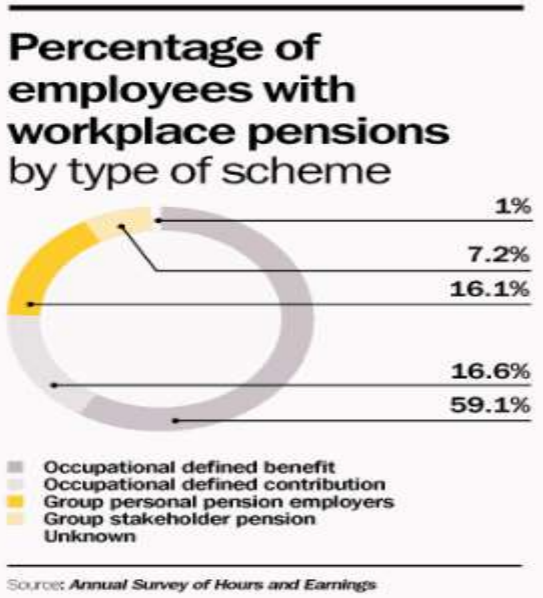
Bar graphs: Consisting of two axes, bar graphs are especially useful in comparing data among a handful of categories at a glance, and can run horizontally or vertically.

How can they be used? Bar graphs are typically used to show big changes in data over time, for example comparing electricity costs in cities across North America.

Area graphs: Area graphs is like the sister graph to line graphs. They can also be used to track changes over time for one or more groups. It would make more sense to use an area graph to track changes in two or more related groups that make up one whole category.

How can they be used? For example, you can use an area graph to compare the staging volumes across three firms of varying sizes.





5. How to use design elements (such as colors or layout ..Visualization)

1. Layout: Pay attention to proportions. Your readers should be able to take a ruler, measure the length or area of your graph, and find that it matches the relationship in your underlying data.

Your graphs should be free of any decorative elements, although you can use graphics, such as icons, that serve to support interpretation.

2. Colors: When it comes to your color scheme, make sure that it is intentional. It should either your organization's or client's brand colors.

Be aware of how your use of color can direct the eye. Pieces of data that are considered supporting or less important, should be in muted or grey colors.

3. Typography: Consider hierarchy when it comes to your text. Your titles should be the largest in size, followed by subtitles, then your labels, and finally – source information. The larger the text, the higher in importance. Your text should be horizontal only, and in 9 points (if in print) and 20 points (if on screen).

Data Design Do's And Don'ts

Do use icons to improve comprehension and eliminate the need for too much labelling

Do visualize your data in a way that makes it easy to compare values

- Don't use more than six colors in a single layout as it can scatter attention
- Don't use 3D charts as it can skew the perception of your data visualizations

6. How to provide references that bring credibility

Things to consider listing out for your data visualizations would be:

- **The author**
- **Dates of data generation**
- **Date of access**
- **Where the data is housed**
- **The publisher of the data**
- **The URL**


Great stories are:
RELATABLE
BELIEVABLE
INCLUSIVE
UNIFYING
USEFUL
IMPORTANT
INSPIRING
EVOCATIVE
EMOTIVE
CONTAGIOUS
Tell yours.