HANDS-ON WITH DATA VISUALIZATION

Darren Chiu

FOR PYDATA HONG KONG
OCT 2018 MEETUP

*This deck does not represent Google nor reflect any Google standpoints

DARREN CHIU

Technologist
+
Educator

Academic Background: #Computer Engineering #Business #Psychology
Work: #Software Engineer #Entrepreneur #Tech Consultant
Industry: #Investment Banking #Education #Media #Digital Marketing

Data Science

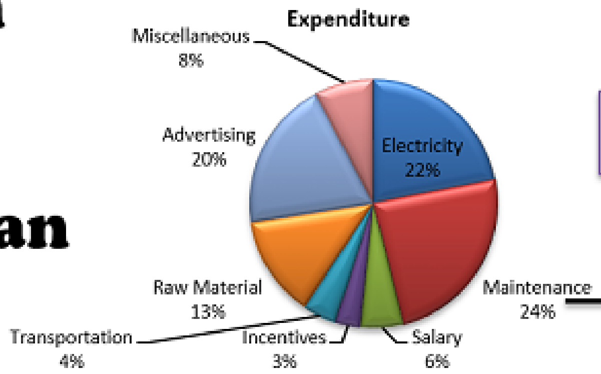
- Data Science = Analysis + Prediction
- EDA (Exploratory Data Analysis): Find useful insights thru analysing past data
- Prediction: Usually done by Machine Learning technology

Exploratory Data Analysis... How?

- "Dimension" as a framework to guide your EDA journey
- 1D : descriptive statistics
- 2D : correlational study
- 3D+ : multi dimensional correlational study

1D: Descriptive Statistics

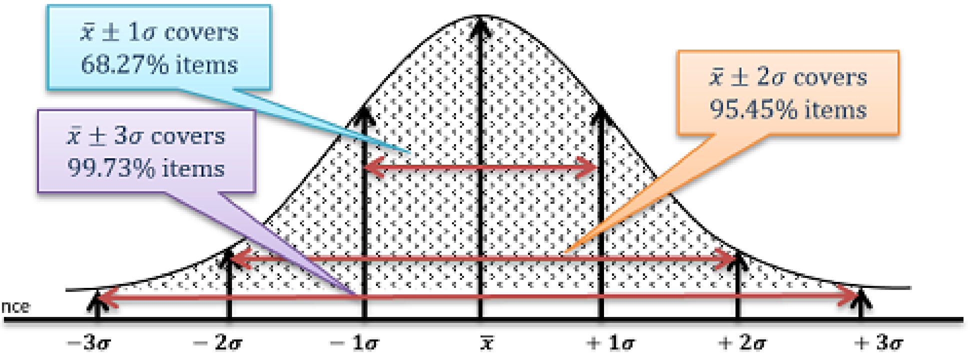
Mean

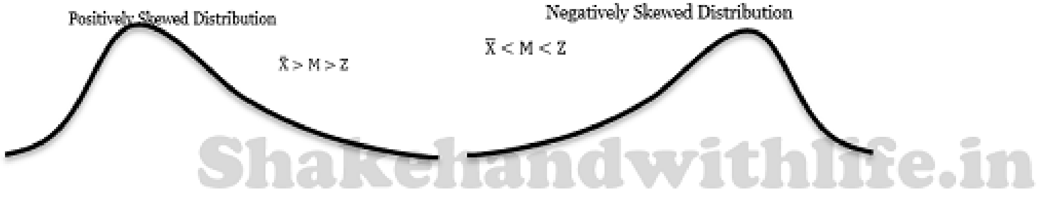


Median

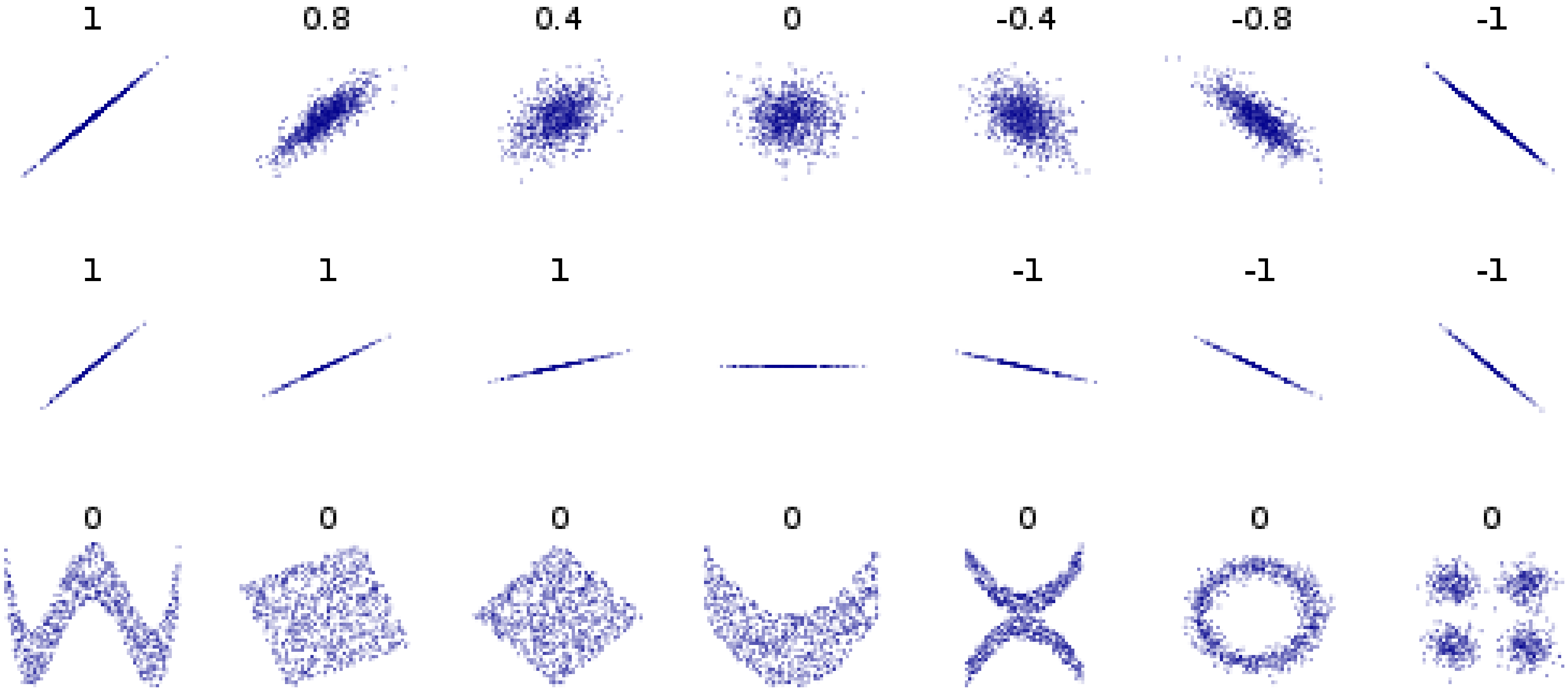
Mode

$$Std.Dev.\sigma = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n}}$$



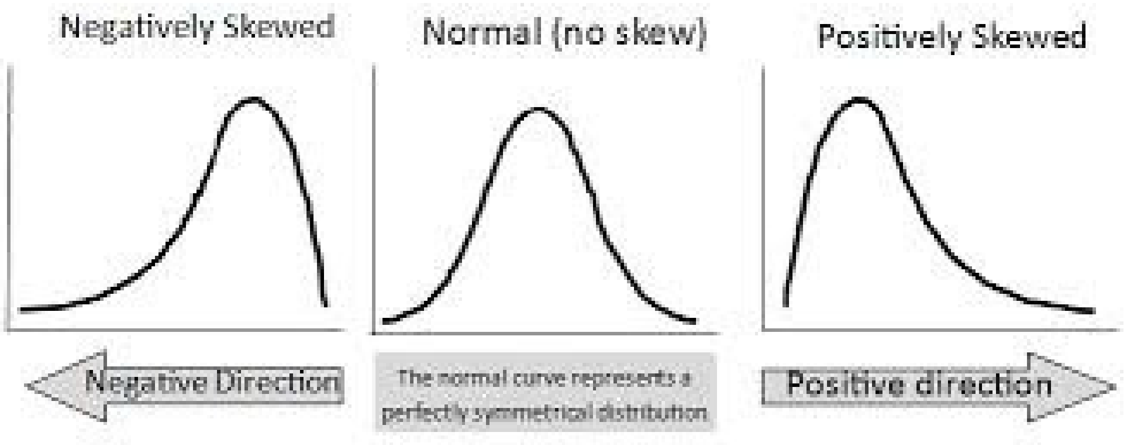

Positively Skewed Distribution
$\bar{X} > M > Z$
Negatively Skewed Distribution
$\bar{X} < M < Z$

2D: Correlational Analysis



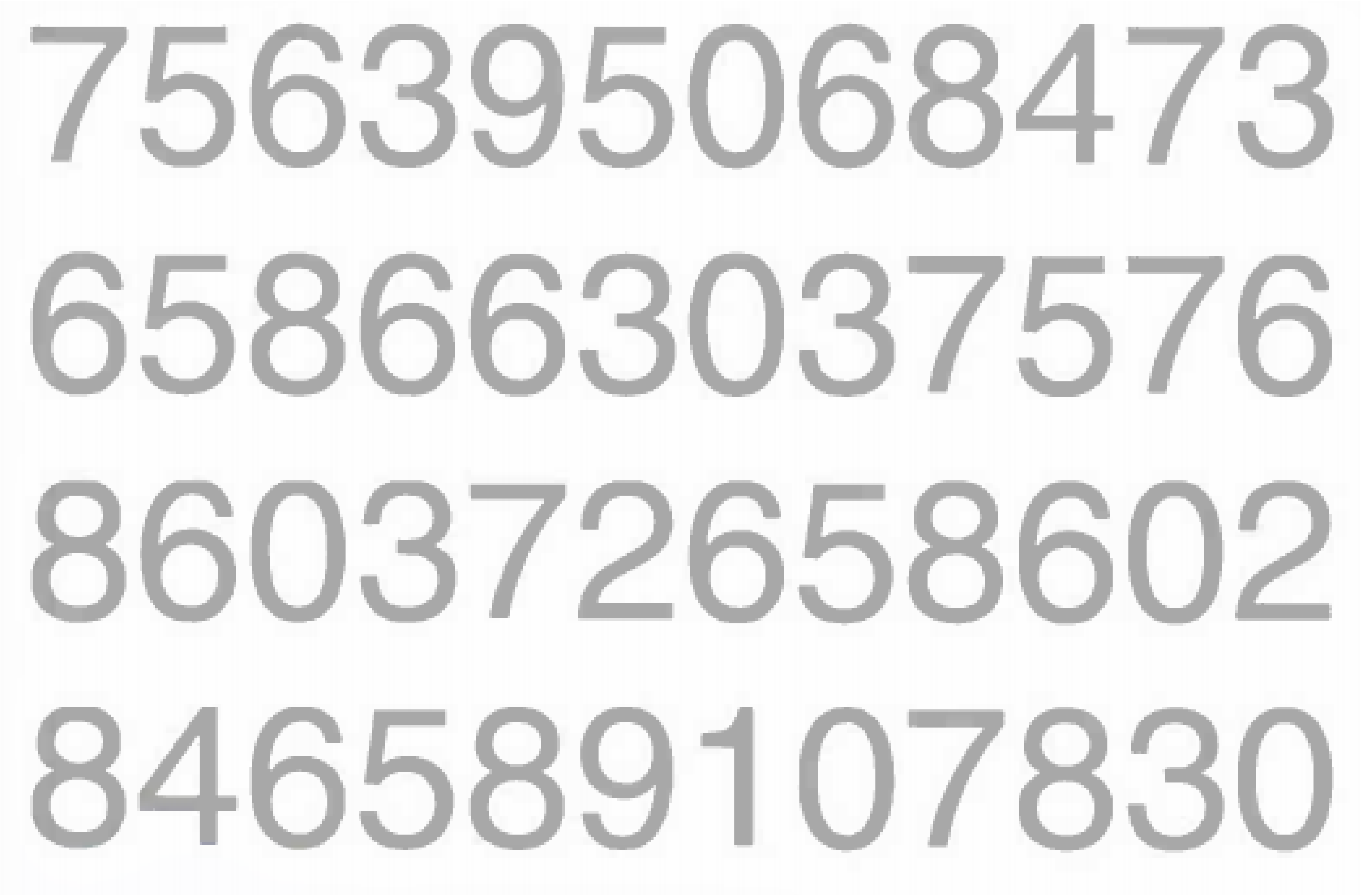
Why we need Visualization

- Visualization is the most powerful tool for EDA
- human brain process image >>> numbers/text



$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

I still don't believe!



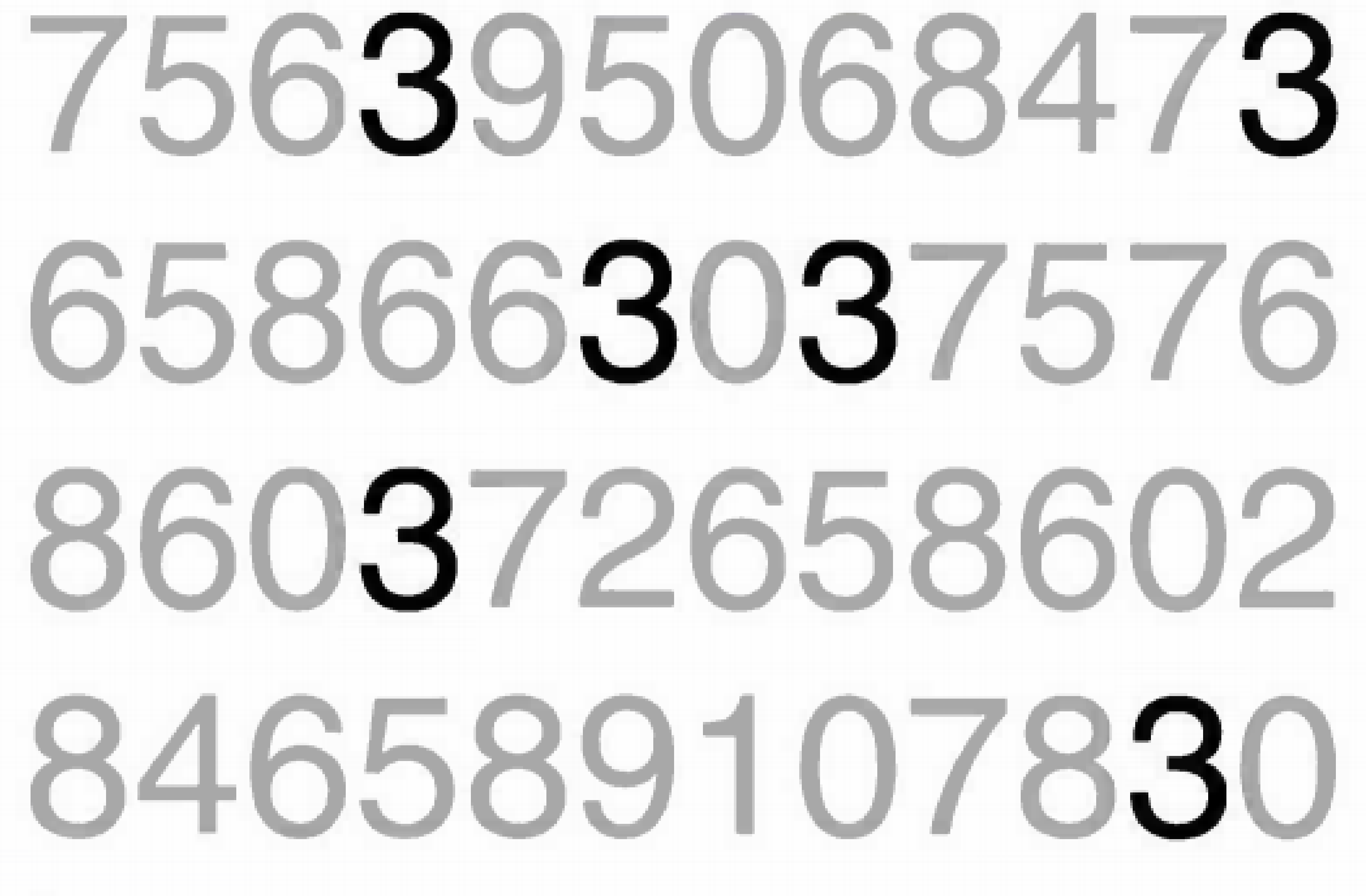
756395068473

658663037576

860372658602

846589107830

I still don't believe!



756**3**950684 7**3**

65866**3**0**3**7576

860**3**72658602

846589107 8**3**0

So what I can leverage?

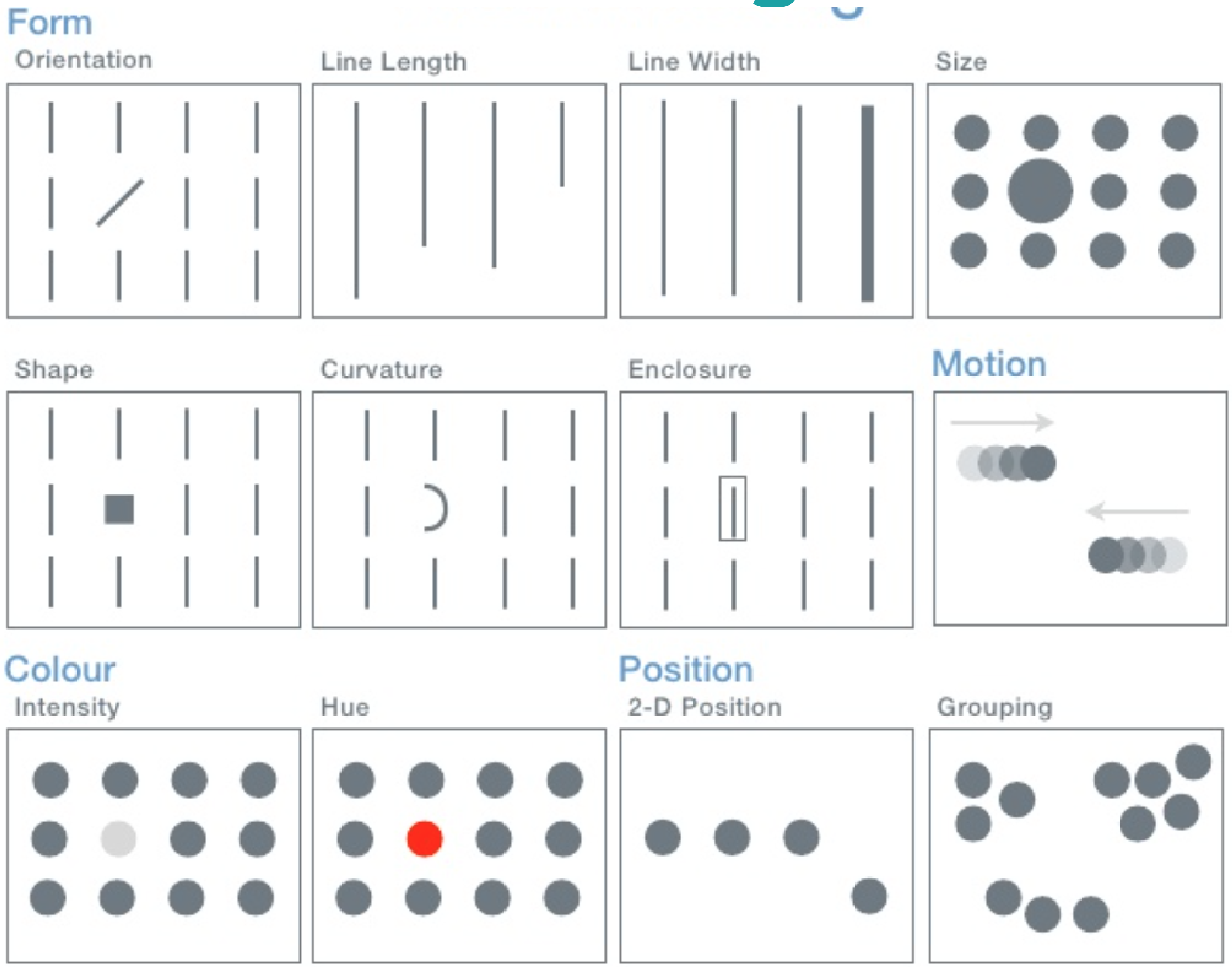
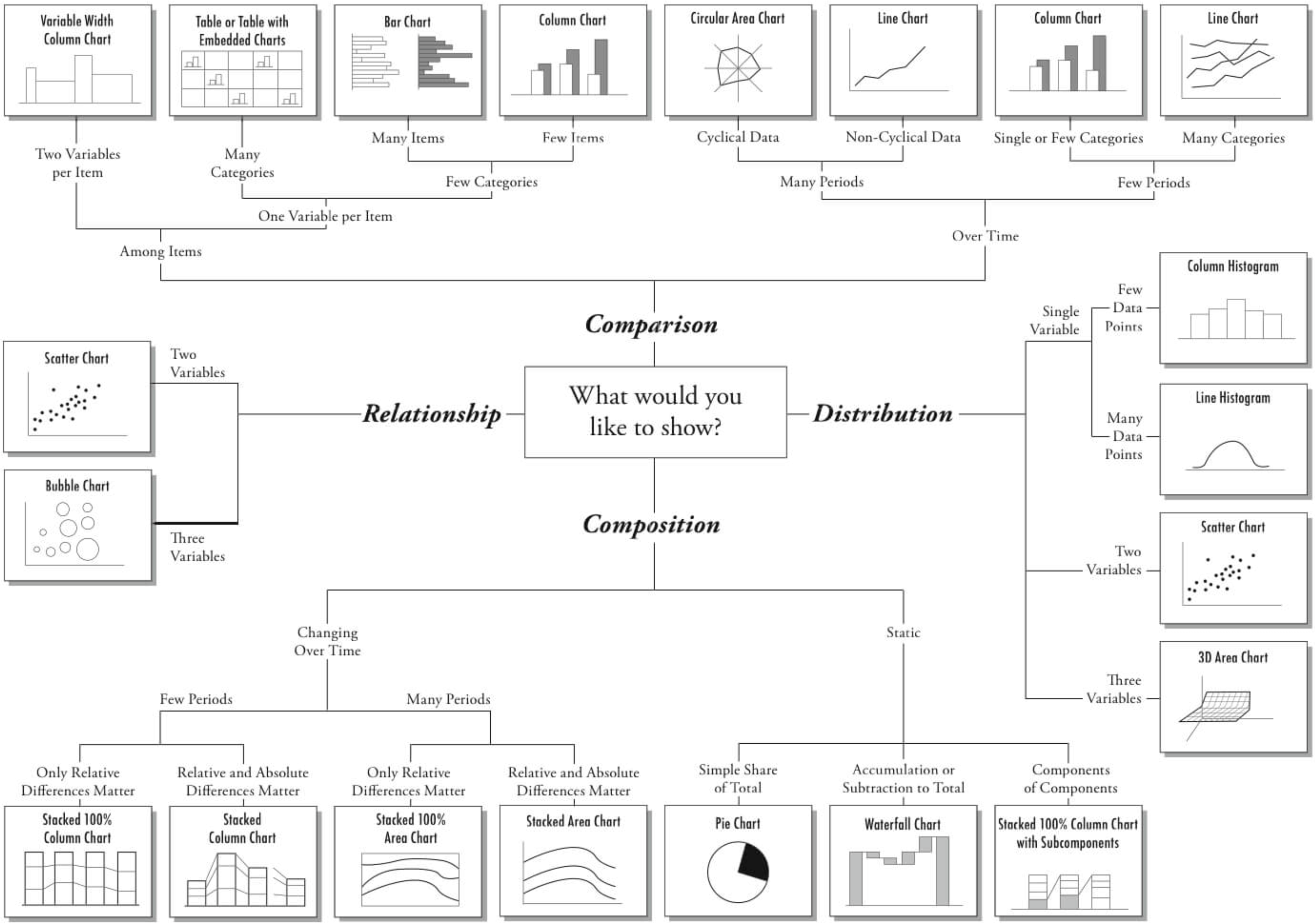


Chart Suggestions—A Thought-Starter



What would you like to show?

Comparison

Among Items

Two Variables per Item — Variable Width Column Chart

One Variable per Item

Many Categories — Table or Table with Embedded Charts

Few Categories

Many Items — Bar Chart

Few Items — Column Chart

Over Time

Many Periods

Cyclical Data — Circular Area Chart

Non-Cyclical Data — Line Chart

Few Periods

Single or Few Categories — Column Chart

Many Categories — Line Chart

Relationship

Two Variables — Scatter Chart

Three Variables — Bubble Chart

Distribution

Single Variable

Few Data Points — Column Histogram

Many Data Points — Line Histogram

Two Variables — Scatter Chart

Three Variables — 3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter — Stacked 100% Column Chart

Relative and Absolute Differences Matter — Stacked Column Chart

Many Periods

Only Relative Differences Matter — Stacked 100% Area Chart

Relative and Absolute Differences Matter — Stacked Area Chart

Static

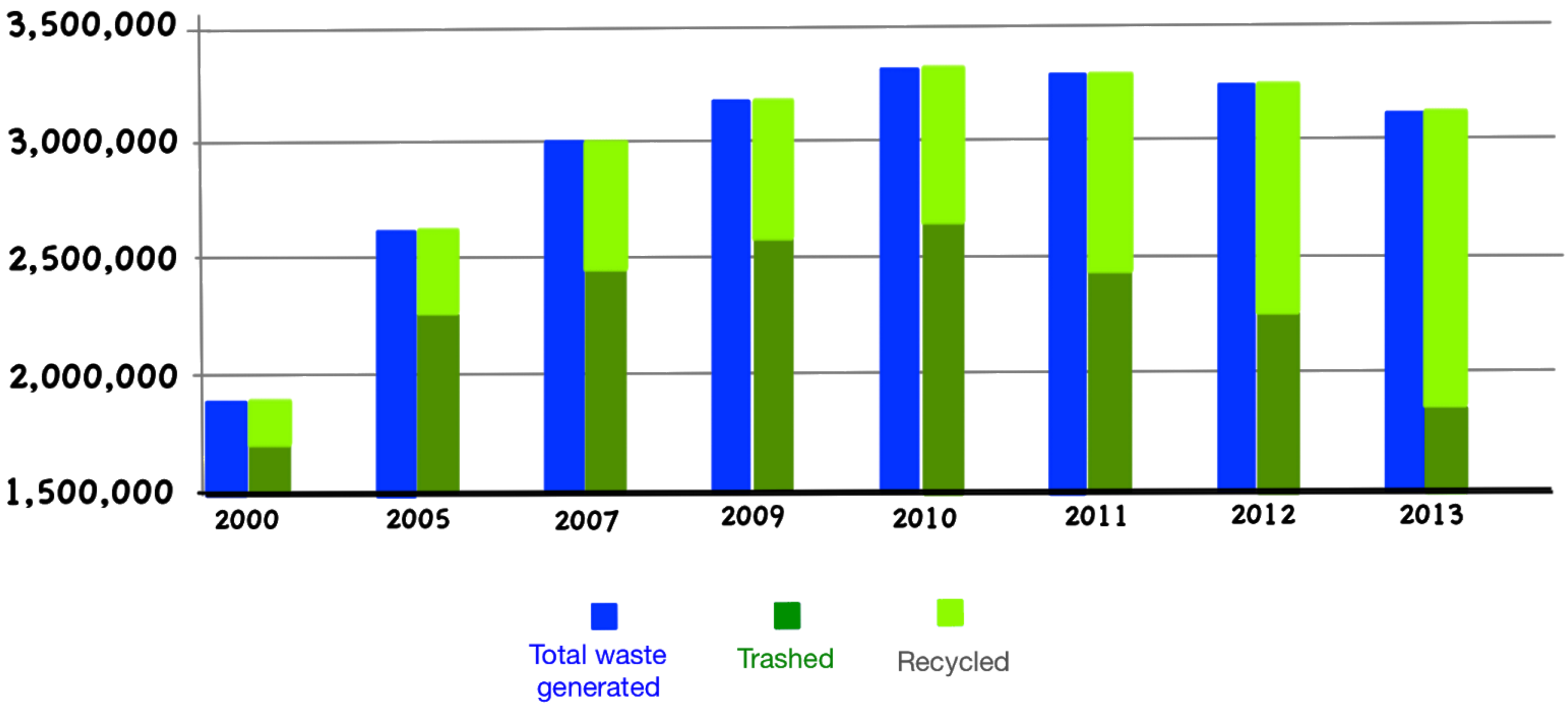
Simple Share of Total — Pie Chart

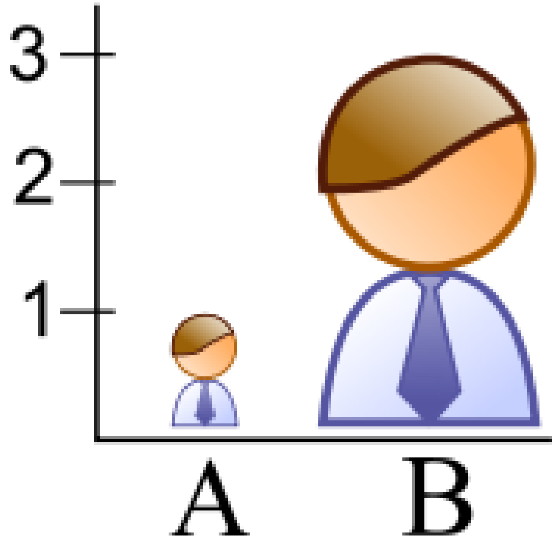
Accumulation or Subtraction to Total — Waterfall Chart

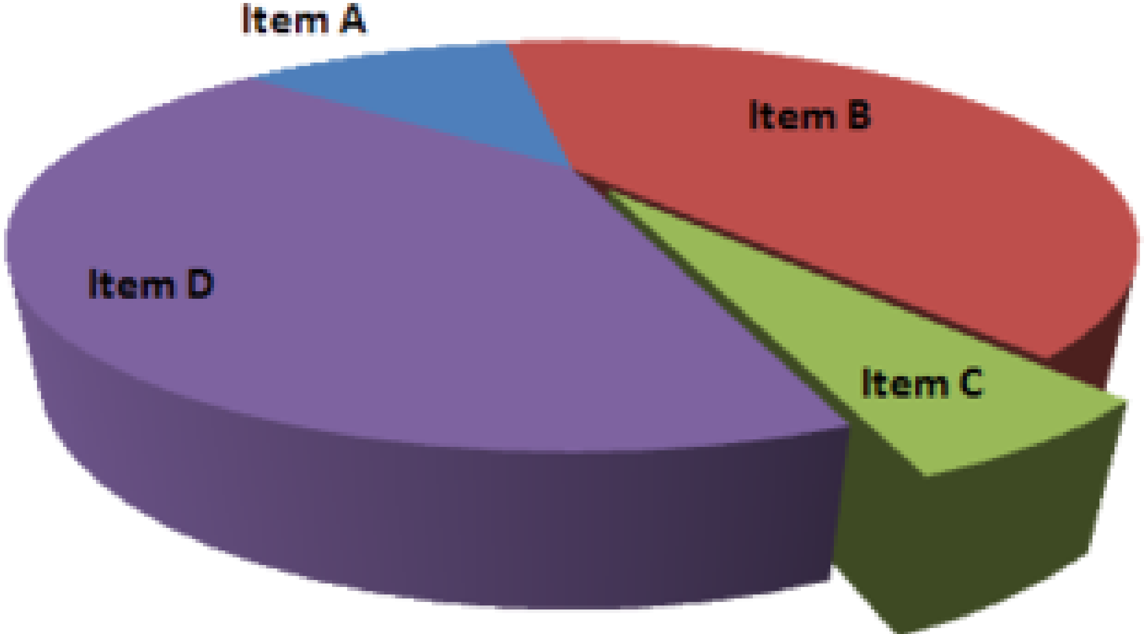
Components of Components — Stacked 100% Column Chart with Subcomponents

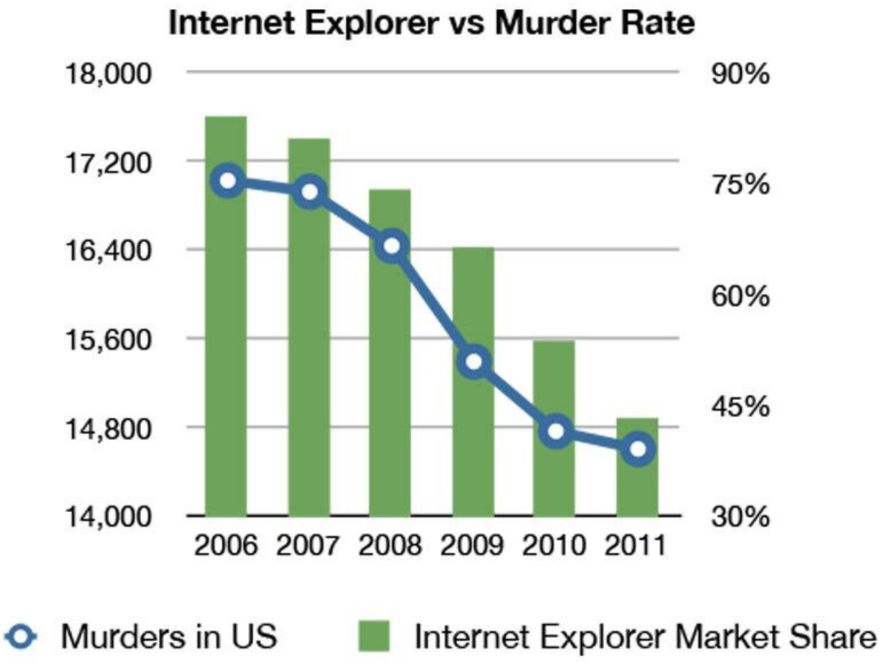
Let's look at
some misleading visualizations...

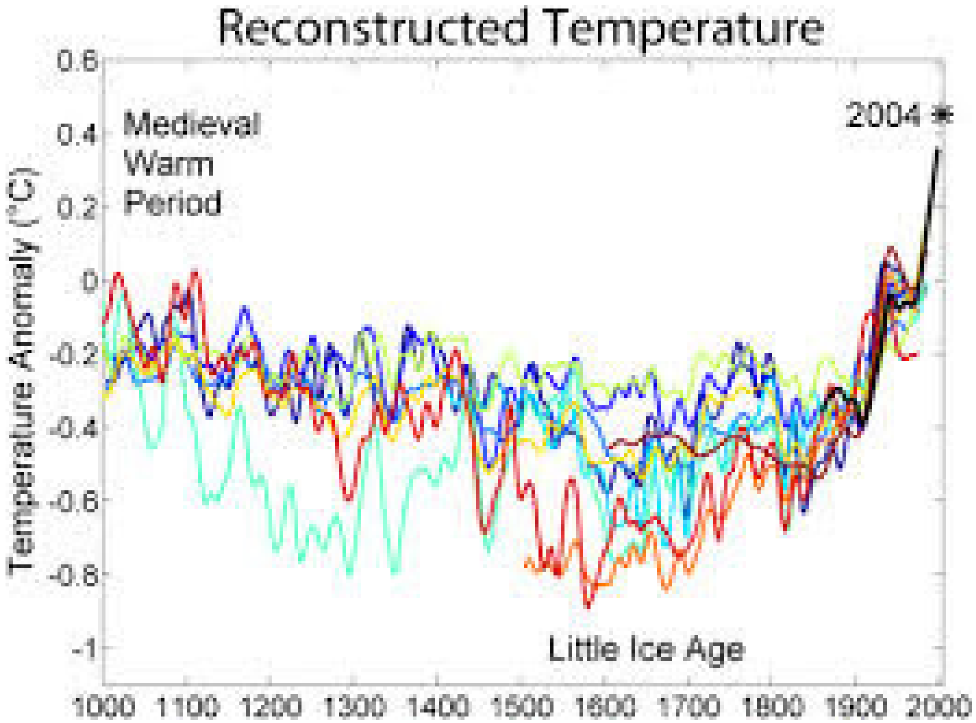
Amount of Environmental Waste Created and Recycled

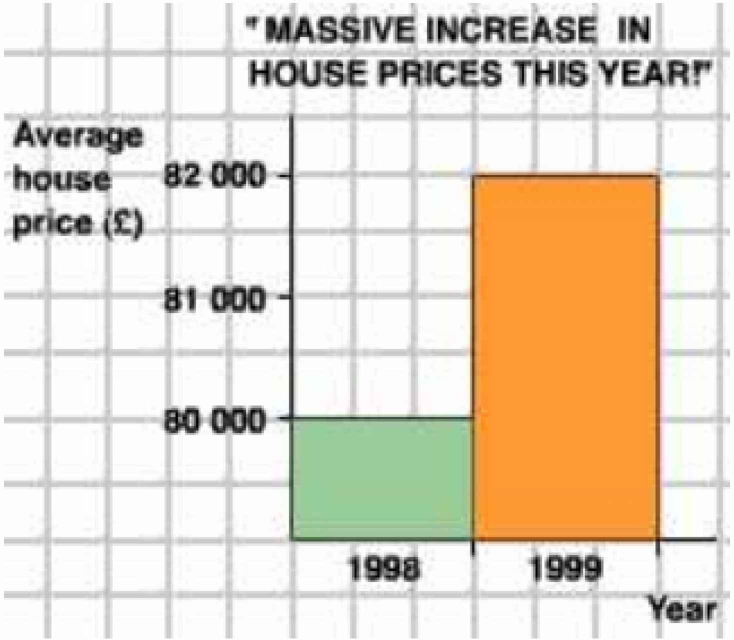







Internet Explorer vs Murder Rate
18,000
17,200
16,400
15,600
14,800
14,000
90%
75%
60%
45%
30%
2006
2007
2008
2009
2010
2011
Murders in US
Internet Explorer Market Share


Reconstructed Temperature
Temperature Anomaly (°C)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Medieval
Warm
Period
2004
Little Ice Age
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000


"MASSIVE INCREASE IN
HOUSE PRICES THIS YEAR!"
Average
house
price (£)
82 000
81 000
80 000
1998
1999
Year

Visualization Tools

Google Data Studio

- Beta since 2016
- went out of beta on 25th Sept
- Hundreds of data connectors
- Run on cloud
- Even high school students could create dynamic interactive data visualization

*This slide does not represent Google nor reflect any Google standpoints

Let's do some exercise...

Google Data Studio:
https://datastudio.google.com

Data Source:
https://www.kaggle.com/vikalpdongre/us-flights-data-2008

Importing Data to Google Cloud BigQuery:
https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv

Exercise 1:

Find out the relationship of month and flight delay

Exercise 2:

So is it because there are more flights in Dec?

Exercise 3:
Then what is the source of flight delay?

Exercise 4:
We now know that weather is the reason of delay...Then is it related to where you are flying?

Exercise 5:
But what if I really want a Christmas holiday… When should I fly?

Exercise 6:

If you still interested - which airline we should go for …

Conclusions - some tips...

- Always start with message you want to deliver
- use only one color tone in one graph
- highlight the thing you want to highlight
- only show data that are relevant
- only show data that can be process by human brain
- Tell the story you want vs tell the right story?

-Don't risk deceiving yourself-