Exploratory Data Analysis

Visualization Process & Role

The Visualization Process

From **data** to **ideas**: efficiently and accurately generating understanding to support decision making.

The roles of visualization

Visualizations can be designed to fulfil one or more roles:

Communication
   *describing and explaining ideas,* e.g. entertainment or learning

Confirmation
   monitoring information streams, e.g. IoT sensor measurements

Exploration
   exploring and analysing information, e.g. public service data

Visualizations are created and exist within a subjective context as images that are interpreted by people. This is influenced by aesthetic, social, cultural, emotional and epistemological context.

From data to wisdom
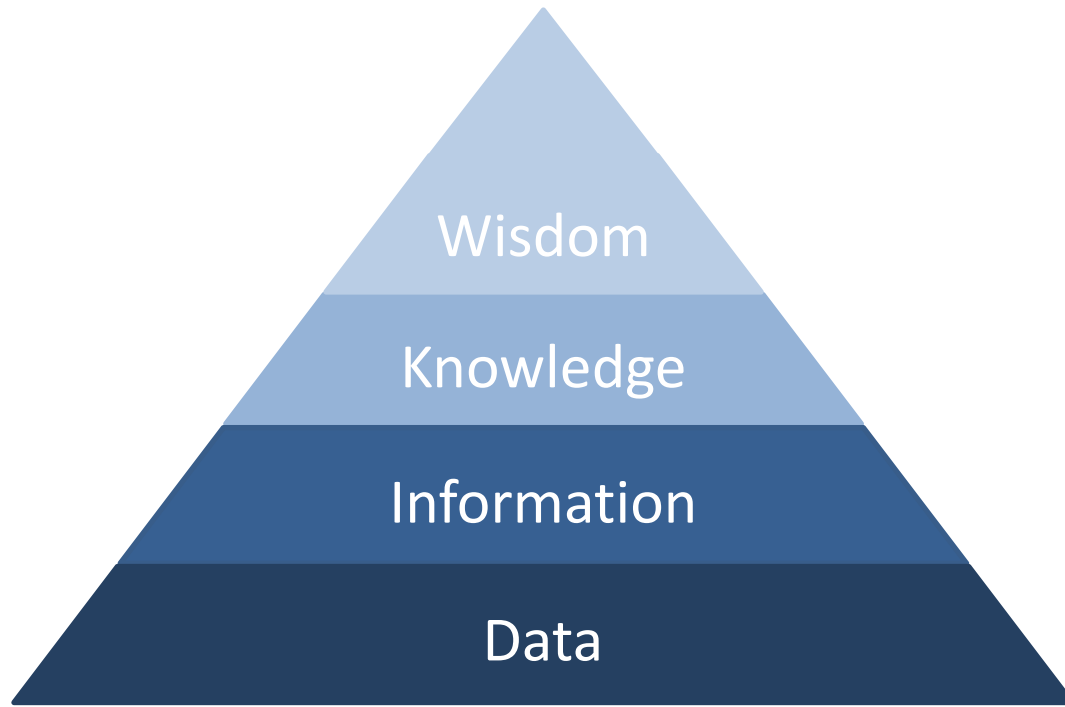
Wisdom: principles and values that inform decisions.

Knowledge: information that is understood by a human being.

Information: structured data with clear meaning or purpose.

Data: symbols and signs representing something in the world.

The goal of many people with data is to highlight information from it that generates knowledge to support better decisions.
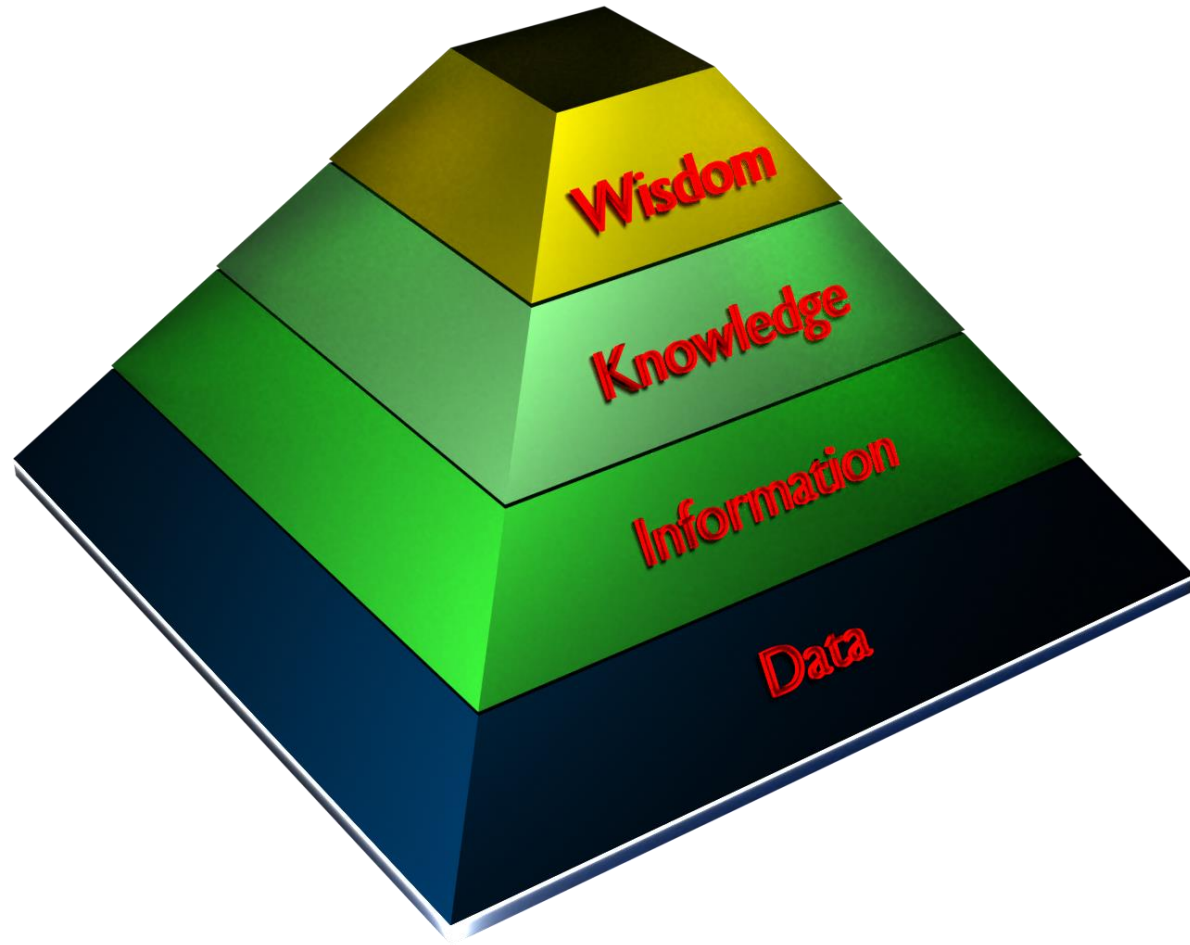
Why use (data) visualization?

"Visual understanding is the essential and only true means of teaching how to judge things correctly, the learning of numbers and language must be definitely subordinated."

Johan Heinrich Pestalozzi
1746-1827

"The greatest value of a picture is when it forces us to notice what we never expected to see."
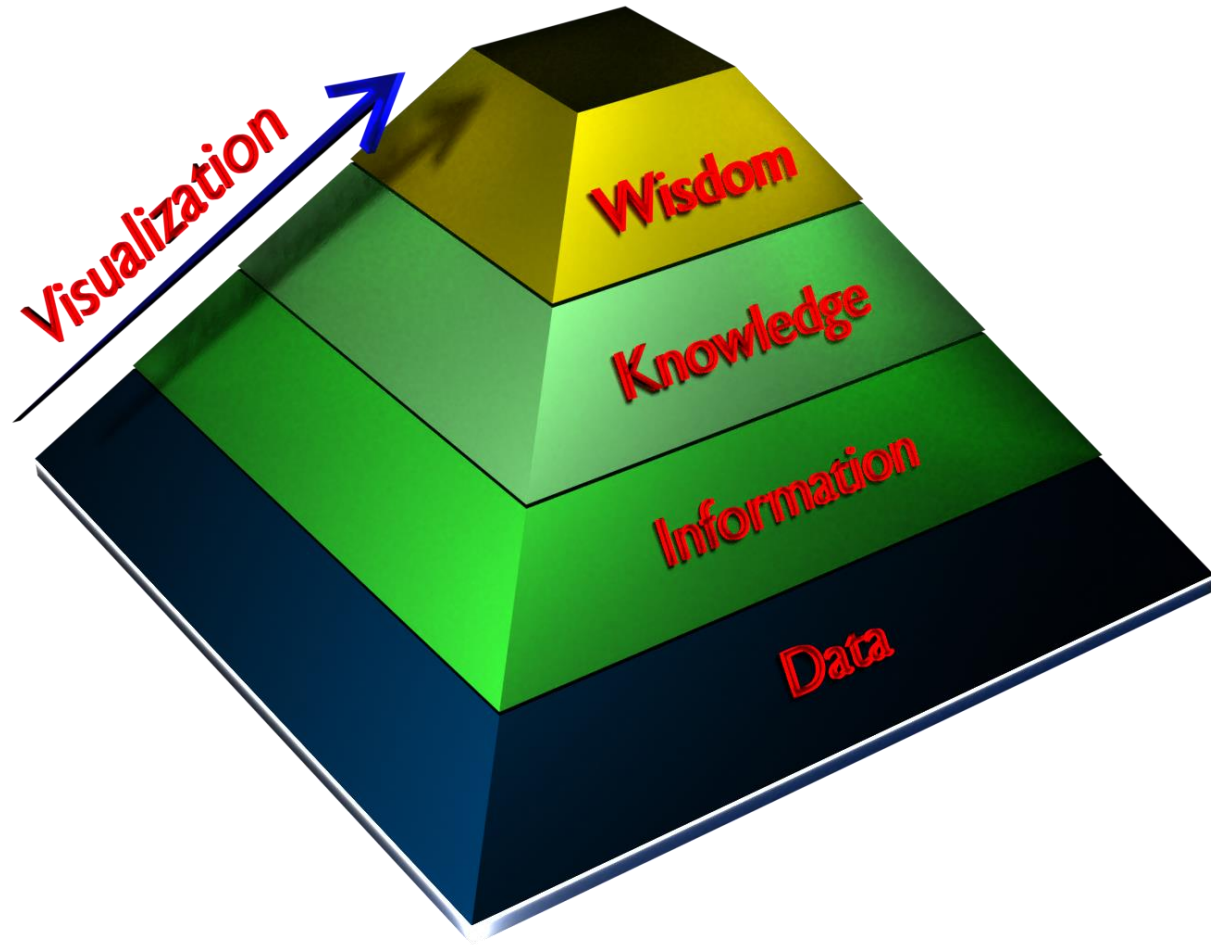
John Tukey, 1977

Why use (data) visualization?

"Visual understanding is the essential and only true means of teaching how to judge things correctly, the learning of numbers and language must be definitely subordinated."

Johan Heinrich Pestalozzi
1746-1827

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey, 1977

Visualization is a tool that supports faster more accurate human cognition.

Importance of visualization

"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, … because now we really do have **essentially free and ubiquitous data**."

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

Unconvinced?

The table below shows tri-glyceride levels by gender, age and income.

Which gender or income group shows a noticeably different effect with age?

Income group	Males		Females	
	Under 65	65 or over	Under 65	65 or over
0-$24,000	250	200	375	550
$25,000 +	430	300	700	500

A visualization

Triglyceride Level

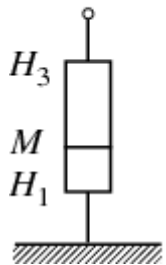Slide after Stephen Kosslyn, Clear and to the Point
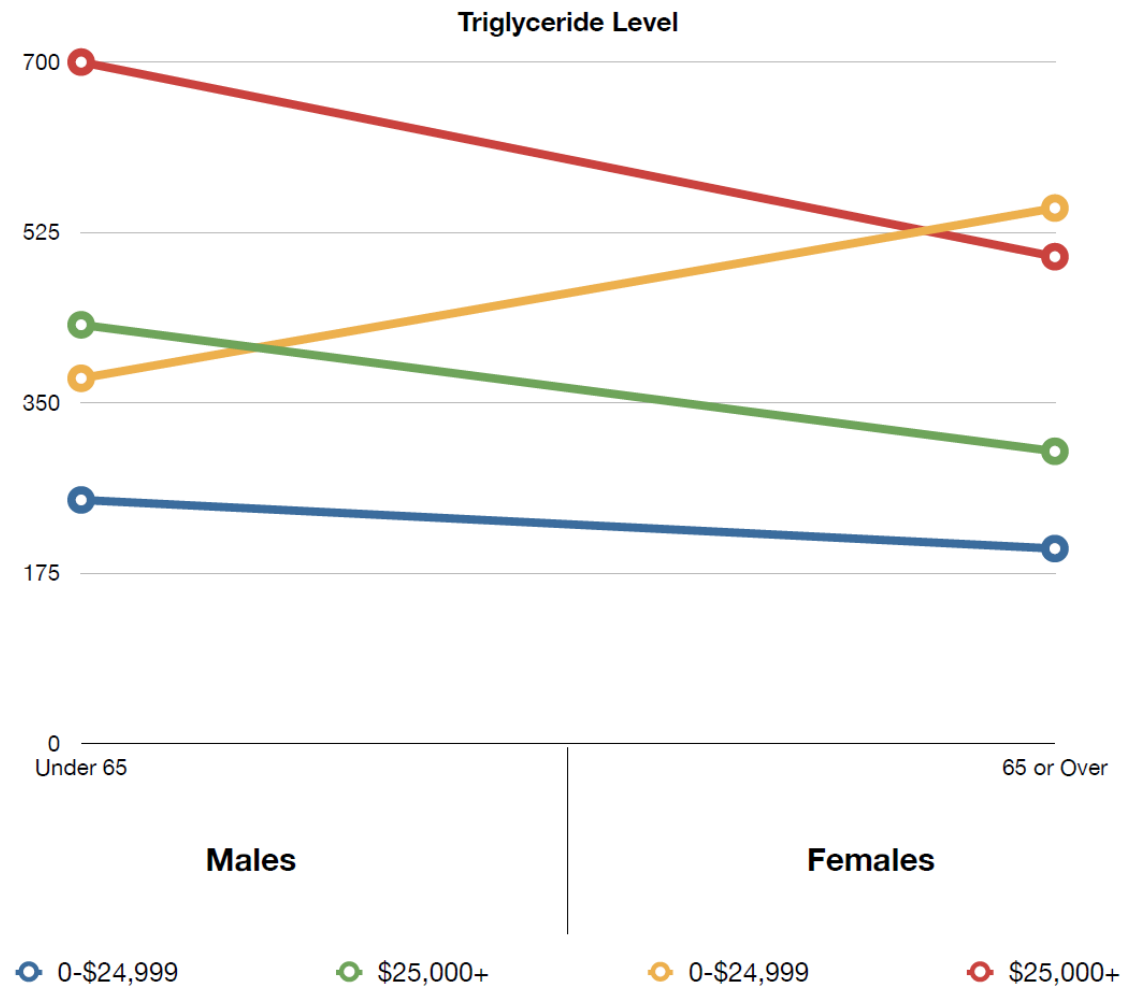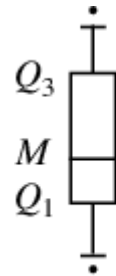
Exploratory Data Analysis

Exploratory Data Analysis

Very (very) often questions and answers are not well defined.

In the 1970s Tukey started to champion exploratory data analysis using visualization to first gain an intuitive understanding of data.
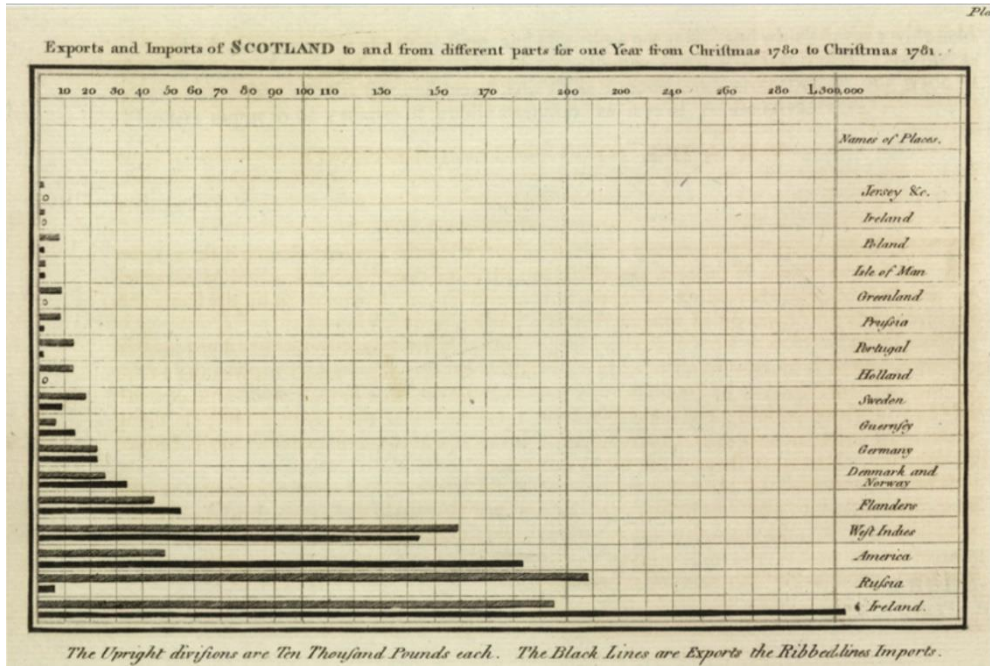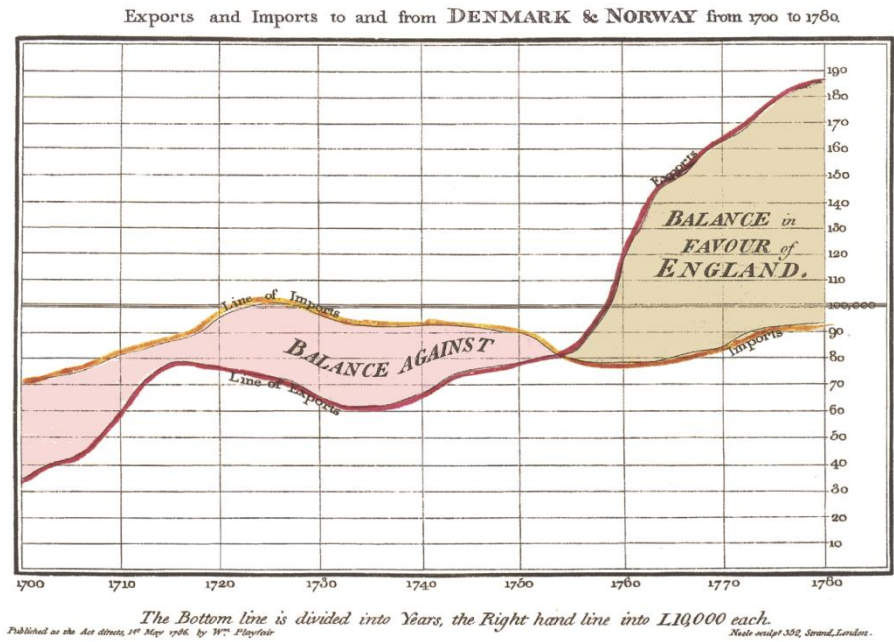
Median,
1st and 3rd quartiles
max and min in the data

Median,
1st and 3rd quartiles
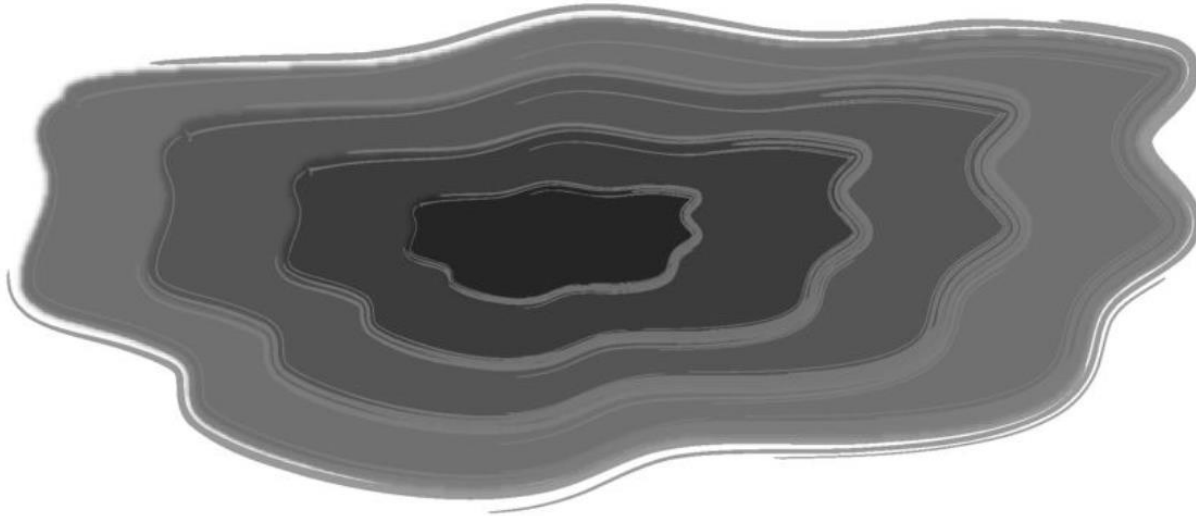max and min def by 1.5 * IQR
plus outliers.

Tukey invented box and whisker diagrams, based on a five-number summary of data, there are now several variations.
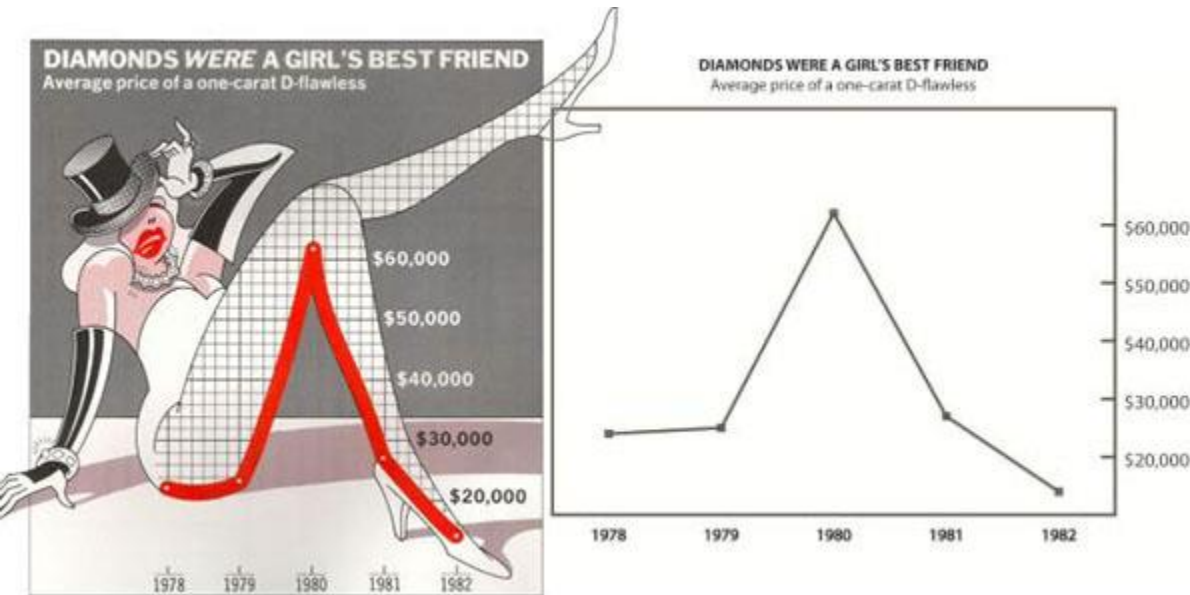
Historical perspective

A widely cited early visualization producer was William Playfair who published his *Commercial and Political Atlas* in 1786, including the first recorded use of bar charts.

Data today?

Data collects in ever expanding lakes, not unlike oil, it needs exploration, mining and refining to become a useful commodity.
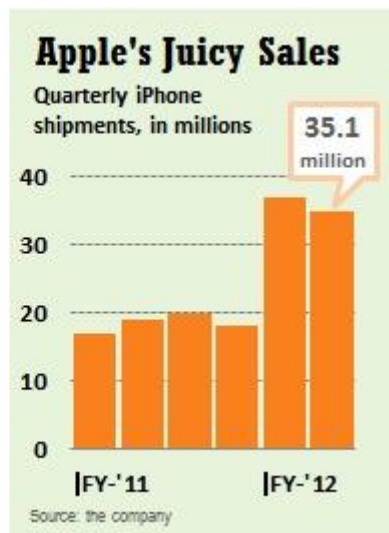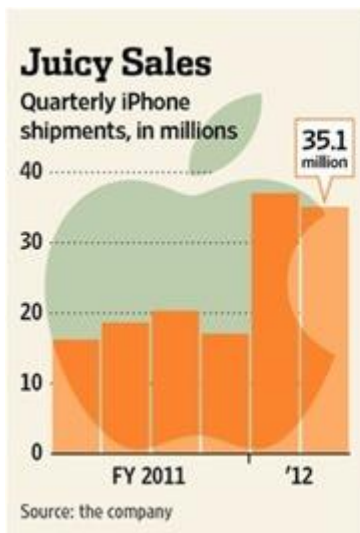
Chartjunk

Edward Tufte – introduced the concept of Chart Junk.

and the closely related design principle of minimising non-data ink.

But… its not so simple…

**Michelle Borkin et al. 2013**
"we found that visualizations with low data-to-ink ratios and high visual densities (i.e., more chart junk and "clutter") were more memorable than minimal, "clean" visualizations. It appears that we are best at remembering "natural" looking visualizations, as they are similar to scenes, objects, and people, and that pictorial and rounded features help memorability."
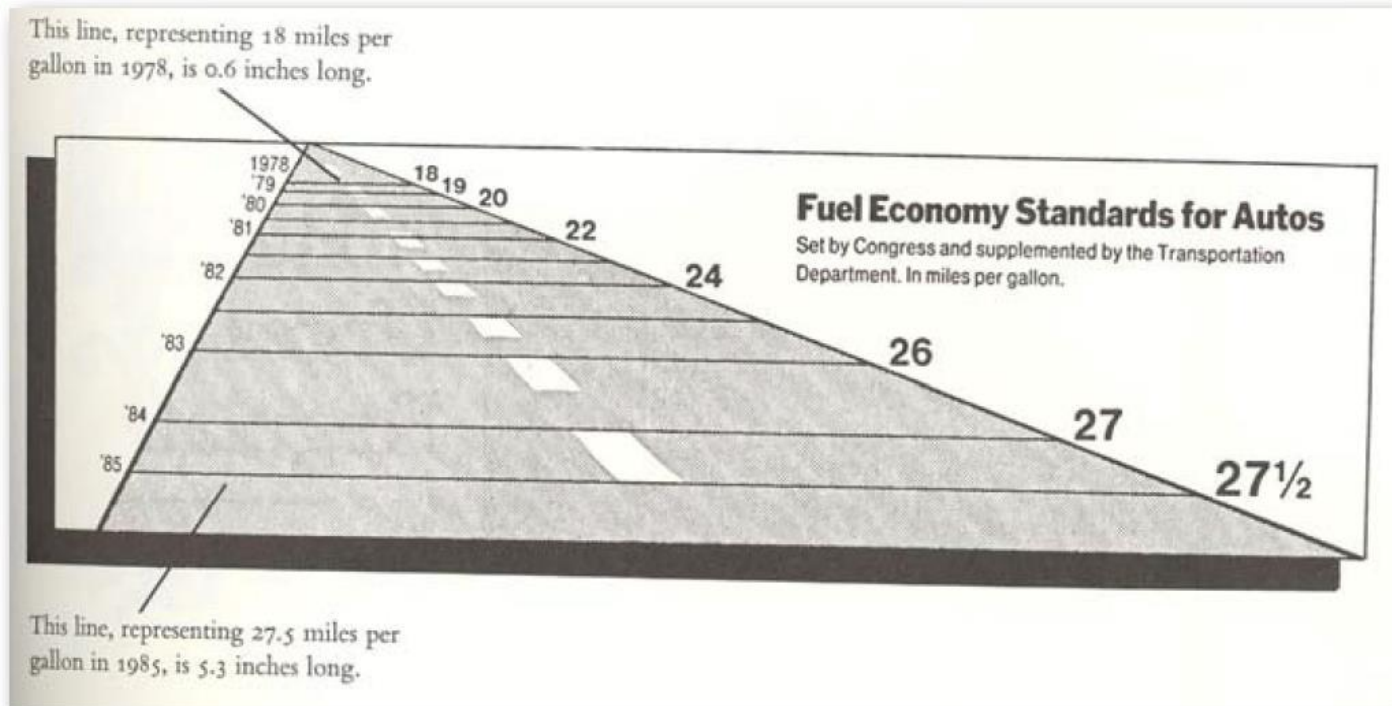
15

Lie Factor
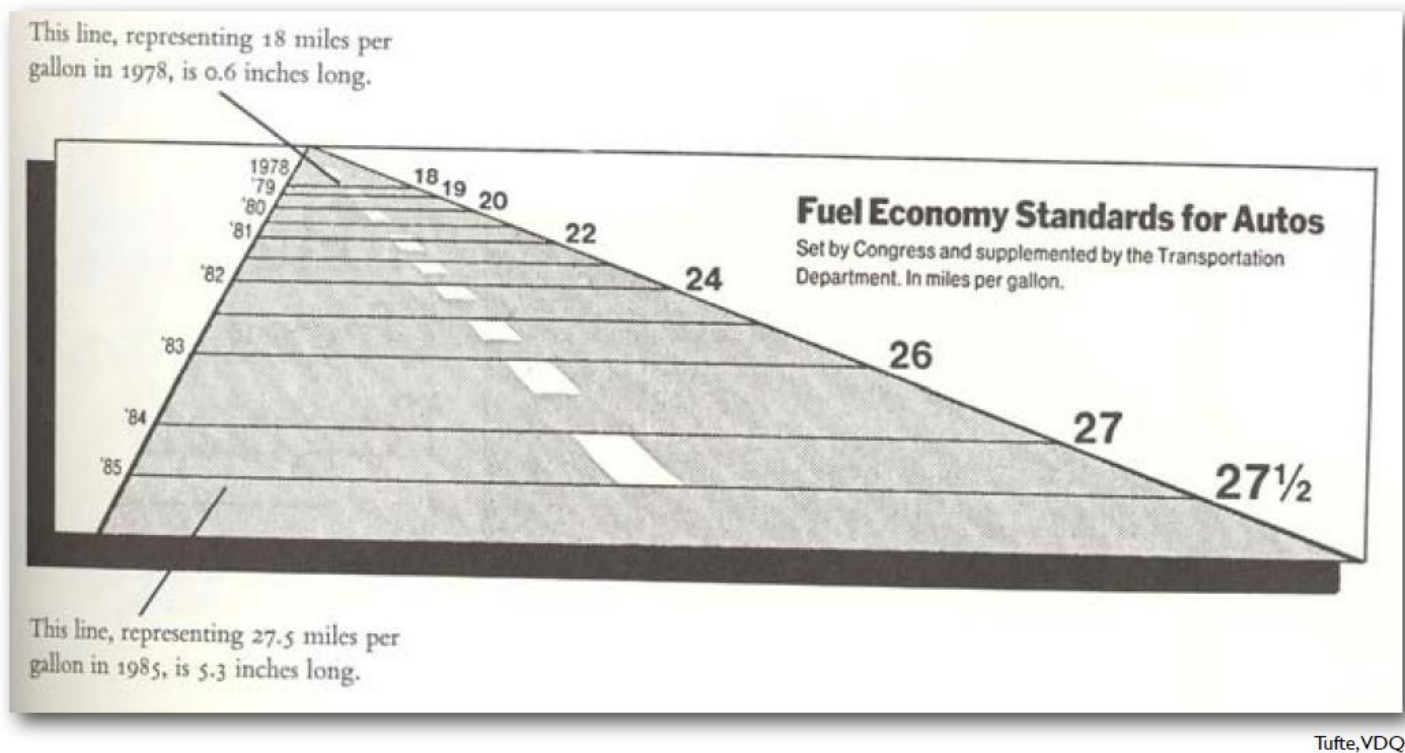
Tufte, VDQI

Another Tufte sourced diagram about the veracity of visualizations

The lie factor is $\dfrac{\text{size of effect in the graphic}}{\text{size of the effect in the data}}$ should = 1 (always!)

Lie Factor

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

1978 '79 '80 '81 '82 '83 '84 '85

18 19 20 22 24 26 27 27½

Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.

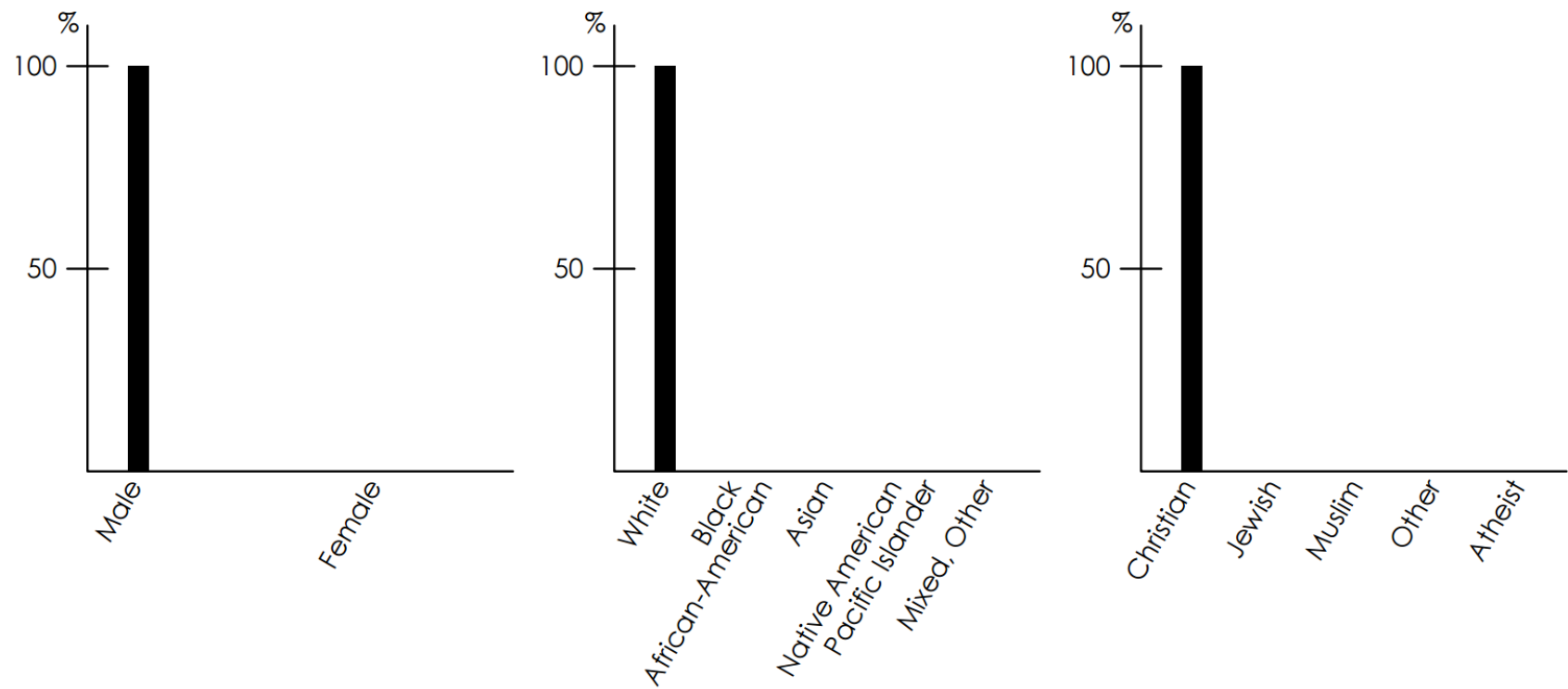This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Tufte, VDQI

The lie factor is $\dfrac{(5.3 - 0.6)\,/0.6}{(27.5 - 18)/18}$ = 14.8

i.e. this over-represents the improvement in fuel economy standards by a factor of nearly 15.

17

Simplicity: Bar Chart

Presidential Demographics
Robert Kosara | EagerEyes.org

*In 2007 before Obama… luckily Trump got things back on track!

Simplicity: Line graph

In Antarctica, measurements of CO_2 concentrations and temperatures go back 650,000 years.

The blue line below charts CO_2 concentrations over this period. The grey line shows the world's temperature.

The top right side of the blue line represents the present era, and that first dip down as you move from right to left is the last ice age. That short distance -- about an inch on the graph -- represents the difference, in Chicago, between a nice day and a mile of ice over your head. Imagine what three times that much on the warm side would mean.

At no point in the last 650,000 years before the pre-industrial era did the CO_2 concentration go above 300 parts per million.

When there is more CO_2 in the atmosphere, the temperature increases because more heat from the Sun is trapped inside.
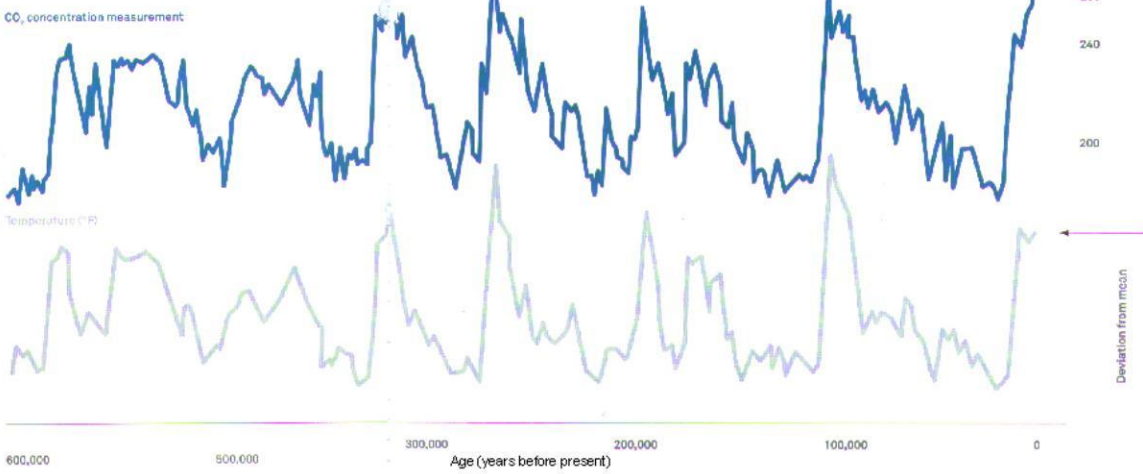
Within 45 years. This is where CO_2 equivalent levels will be if we do not make dramatic changes quickly.

Here's where CO_2 is now -- way above anything measured in the prior 650,000-year record of ice core samples.

CO_2 concentration measurement

Temperature (°F)

600,000   500,000   300,000   200,000   100,000   0
Age (years before present)

CO_2 (ppm)
600
400
300
280
240
200

Deviation from mean

Al Gore

An inconvenient truth.

EDA Key Points

Roles of visualization:

Communication, Confirmation, Exploration

Exploratory Data Analysis

Use visualization to understand data before full analysis

Simplicity in visualization

Clarity is key, use ink on the page to show your message

The value of information

"

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
"

from *The Rock*, T.S. Elliot, 1930

Philosophy from philo Sophia
= friend of Wisdom