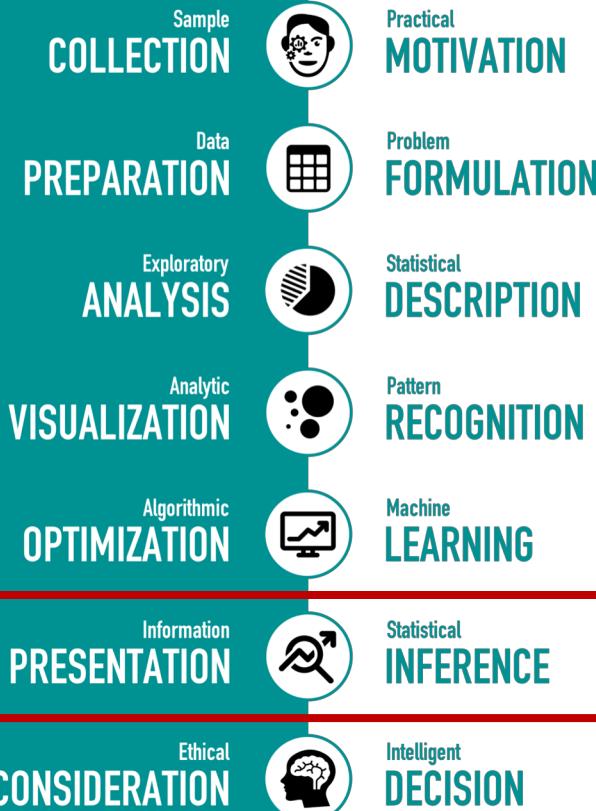


Principles of Visualization

Sourav SEN GUPTA
Lecturer, SCSE, NTU





Data Science Data Visualization

Information Presentation

Is there a “story” hidden in your data?
How to use visuals as information?
How to tell the “story” effectively?

**How to present Data in
the most engaging way?**

There are two goals
when presenting data:
convey your story and
establish credibility.



Edward R. Tufte

<https://www.edwardtufte.com/tufte/>

convey your story

Effectiveness

A visualization is more *effective* than another [vis] if the information conveyed by one visualization is more **readily perceived** than the information in the other visualization.

Jock D. Mackinlay

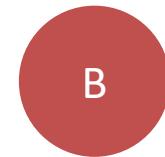
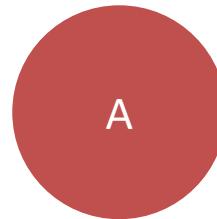
APT : A Presentation Tool (1986)

4



convey your story

readily perceived



Jock D. Mackinlay

APT : A Presentation Tool (1986)

5



establish credibility

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express **all the facts** in the set of data, and **only the facts** in the data.

Jock D. Mackinlay

APT : A Presentation Tool (1986)

6

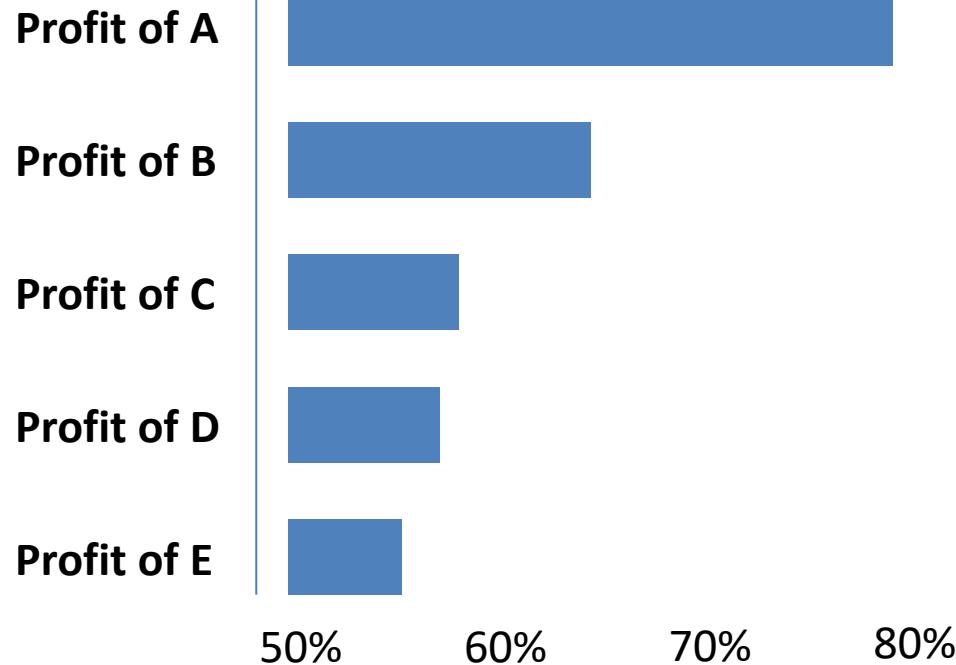


establish credibility

all the facts

Jock D. Mackinlay

APT : A Presentation Tool (1986)



A



B



C



D



E



0%

10%

20%

30%

40%

50%

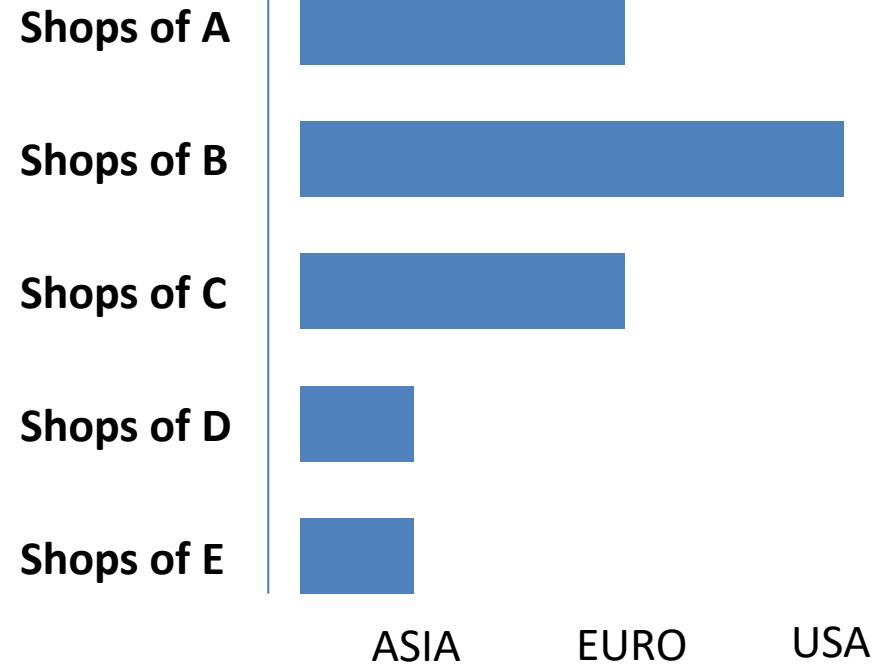
60%

70%

80%

establish credibility

only the facts



Jock D. Mackinlay

APT : A Presentation Tool (1986)

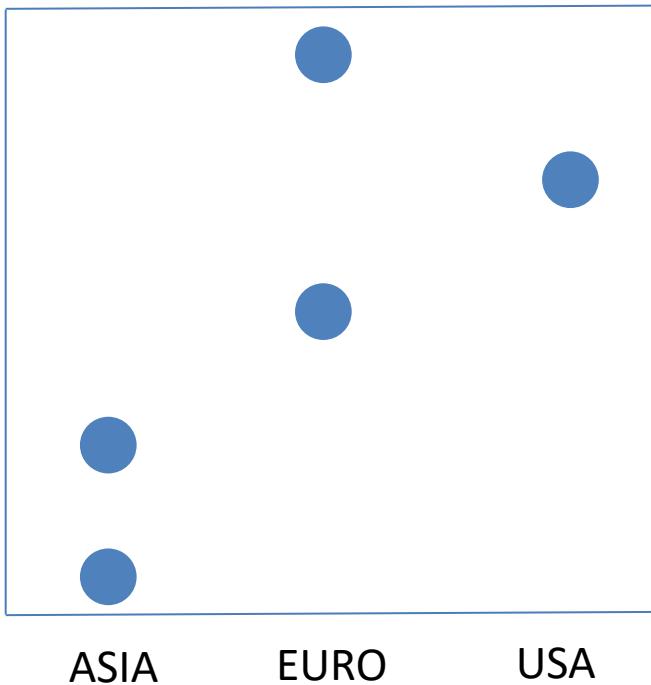
Shops of A

Shops of B

Shops of C

Shops of D

Shops of E



convey your story

Effectiveness

Use encodings that
people decode better.

Better means more accurate and faster.

establish credibility

Expressiveness

Tell the truth and
nothing but the truth.

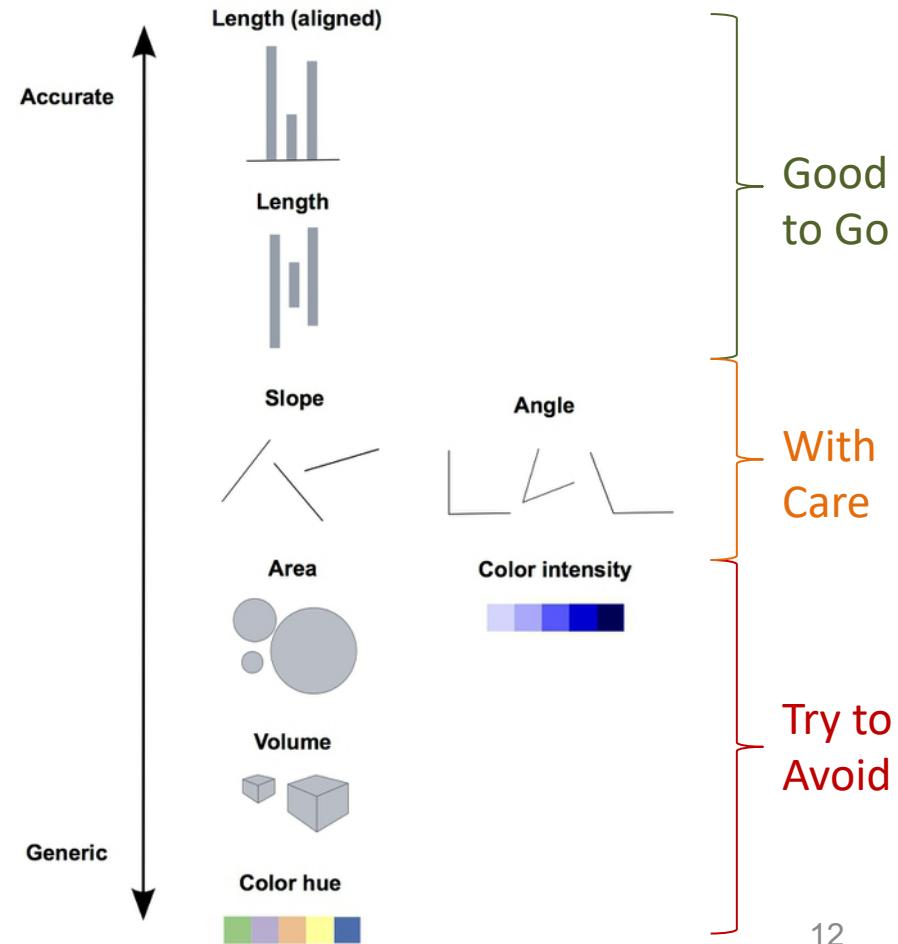
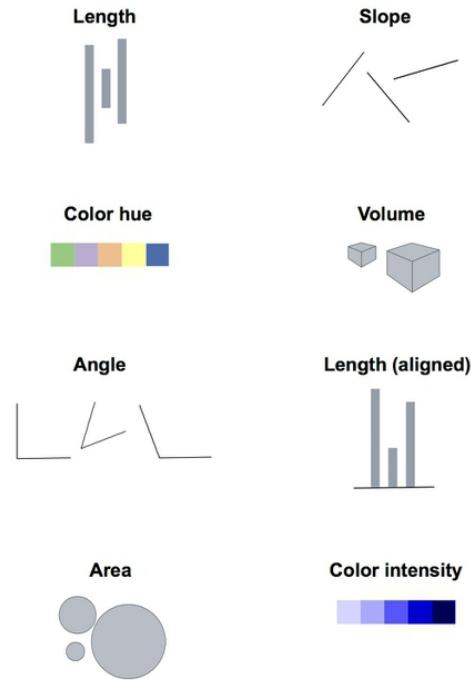
Do not lie, and do not lie by omission.

Jeffrey Heer

Principles of Data Visualization

11





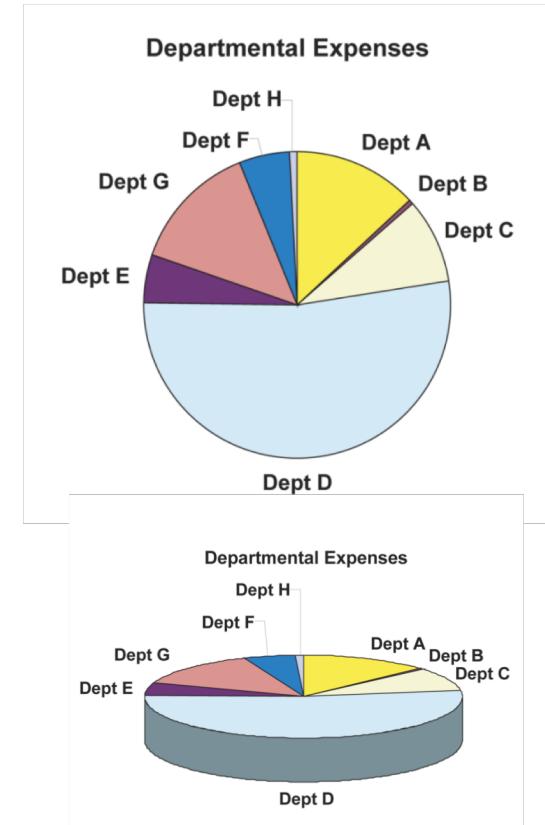
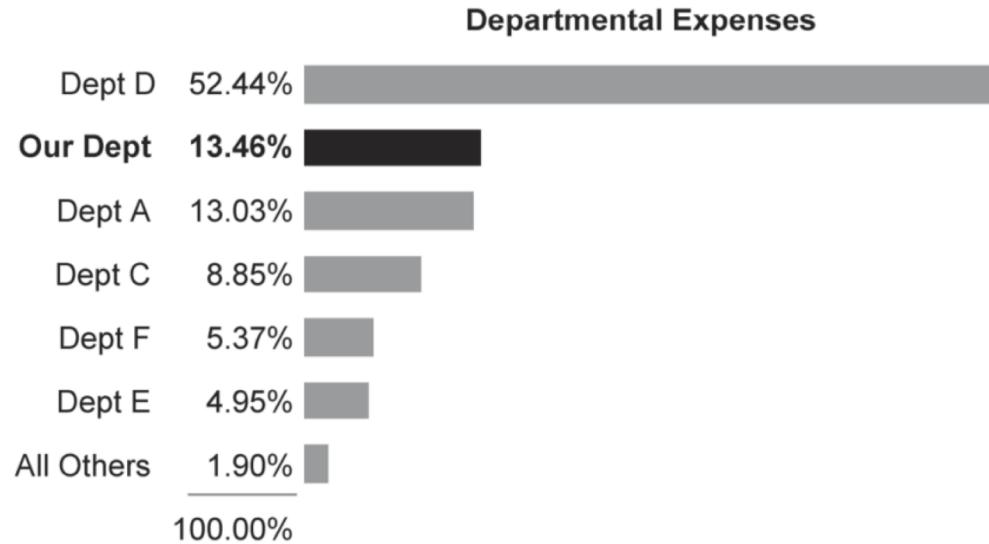
Peter Aldhous

<http://paldhous.github.io/ucb/2018/dataviz/index.html>



Stephen Few

Show me the Numbers



2005 Sales Revenue (USD)

Sales Channel	Q1	Q2	Q3	Q4
Direct	383,383	403,939	437,373	538,583
Indirect	283,733	283,833	257,474	258,474
Total	667,116	687,772	694,847	797,057

2005 Sales Revenue (USD)

Sales Channel	Q1	Q2	Q3	Q4
Direct	383,383	403,939	437,373	538,583
Indirect	283,733	283,833	257,474	258,474
Total	667,116	687,772	694,847	797,057



Data Ink vs. Non-Data Ink

The data-ink ratio must be as high as possible.

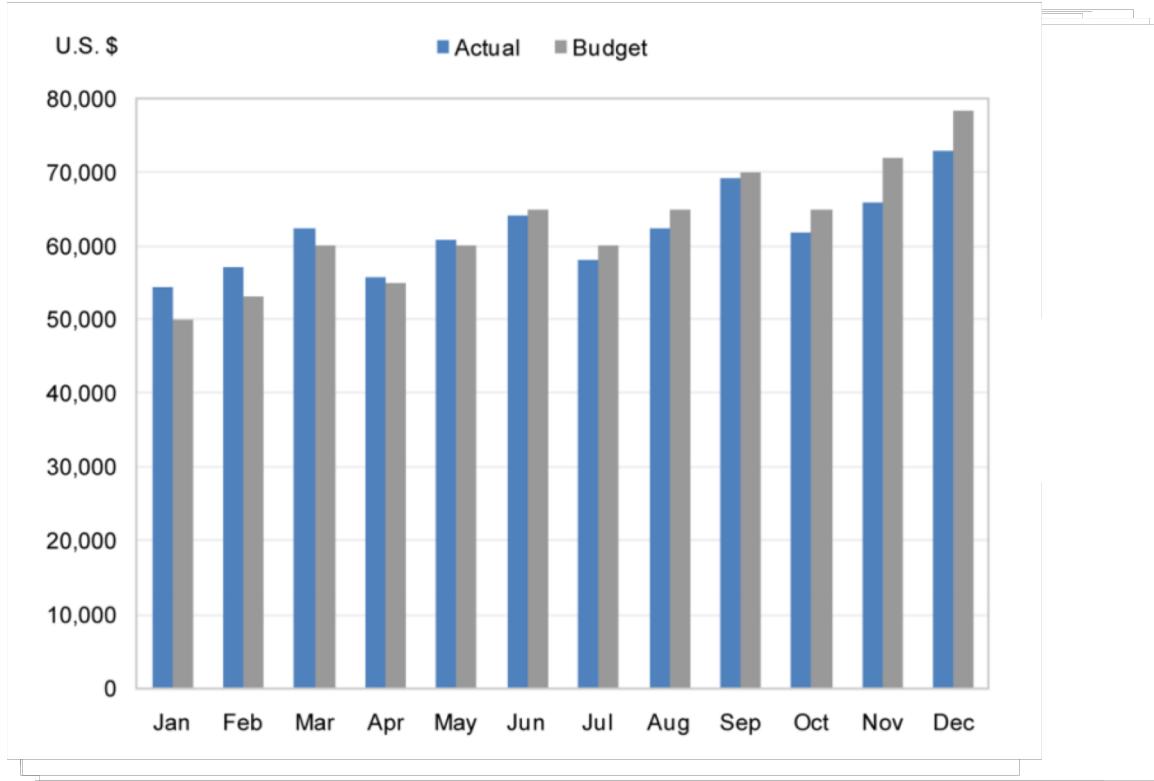
- Tufte

Stephen Few

Show me the Numbers

14





Data Ink vs. Non-Data Ink

The data-ink ratio must be as high as possible.

- Tufte

Stephen Few

Show me the Numbers

15



Expenses Percentage Variance from Budget



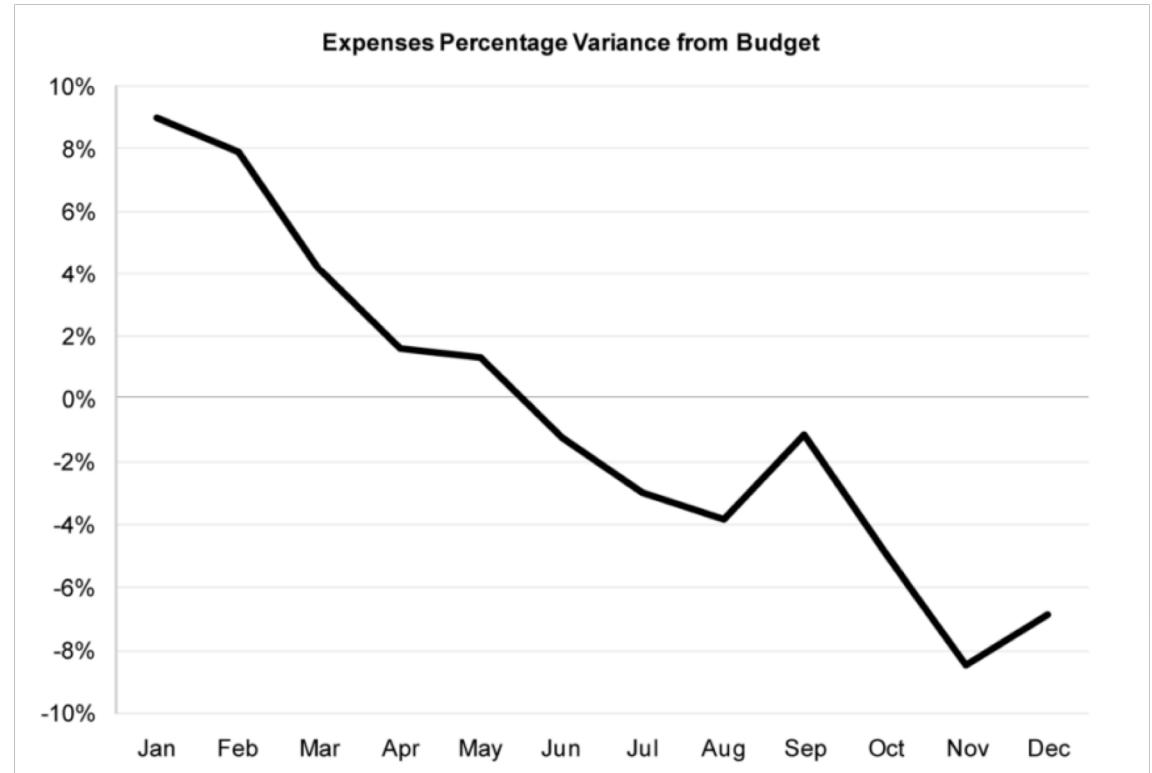
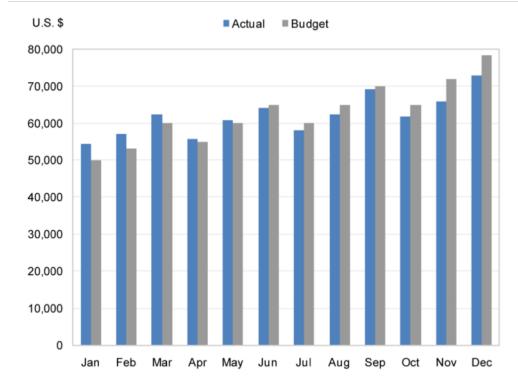
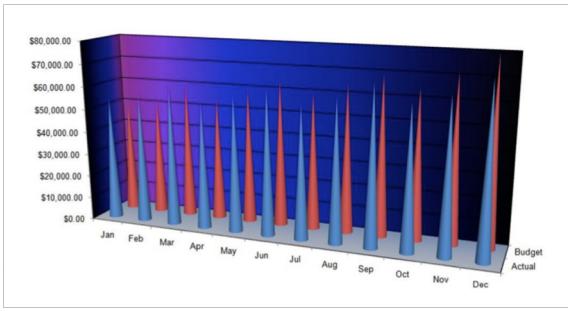
Stephen Few

Show me the Numbers

Interpretation

Realizing the context and the most effective way to present your data.





Above all else, show the Data

Edward R. Tufte

17



Distribution

Relationship

Comparison

What do you want to show?

Connection

Composition
(parts of the whole)

Location

Data Type

Numerical

Categorical

Mixed Type

Map Data

Network

Time Series

Peter Aldhous

<http://paldhous.github.io/ucb/2018/dataviz/index.html>

18



Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a known average. Can also be used to show if something is skewed (positive/heuristic/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which have two opposing views (disagree/heuristic/agreed).

Spine

Spins a single value into two contrasting components (eg male/female).

Surplus/deficit line

The shaded area of these charts allows a baseline to be shown – either against a baseline or between two series.

XY heatmap

A good way of showing differences between 2 categories of data less effective at showing fine differences in amounts.

Bubble

Like a scatterplot but adds additional detail by using bubbles.

Dot strip plot

Dots placed in order on a strip are a great method of laying out ranks across multiple categories.

Lollipop

Lollipop draws more

Bump

Effective for showing changing proportions over time.

Cumulative curve

A good way of showing frequency distributions. As a rule of thumb it is always cumulative frequency, x-axis is always a measure.

Frequency polygon

For displaying multiple distributions of data. Like a regular line chart but limited to a maximum of 3 or 4 datasets.

Beeswarm

Use to emphasise individual points in a dataset. Points can be scaled on an additional variable. Best suited to medium-sized datasets.

Correlation

Show the relationship between two or more variables. Its strength may tell you if they tell them otherwise, many readers will assume the relationships there must be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (column) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables changes over time.

Bubble

Like a scatterplot but adds additional detail by using circles according to a third variable.

XY heatmap

A good way of showing differences between 2 categories of data less effective at showing fine differences in amounts.

Dot strip plot

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipop draws more

Bump

Effective for showing changing proportions over time.

Cumulative curve

A good way of showing frequency distributions. As a rule of thumb it is always cumulative frequency, x-axis is always a measure.

Frequency polygon

For displaying multiple distributions of data. Like a regular line chart but limited to a maximum of 3 or 4 datasets.

Beeswarm

Use to emphasise individual points in a dataset. Points can be scaled on an additional variable. Best suited to medium-sized datasets.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, legislature, constituency election results

Ordered bar

Standard bar charts display the ranks of variables more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing the rank of data is not so important.

Dot strip plot

Good for showing individual values in a distribution, can be a problem if there are too many dots have the same value.

Barcode plot

Like a dot strip plot, good for displaying all individual values, though they work best when highlighting individual values.

Slope

Good for showing changing data as long as it can be simplified into 2 or 3 points without missing a key part of story.

Boxplot

Summarise multiple dimensions over time or vary between categories.

Violin plot

Similar to a box plot but more effective with complex distributions (and that cannot be summarised with simple averages).

Bump

Effective for showing changing proportions over time.

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population, constituency election results

Histogram

The standard way to show a statistical distribution, keeps the gaps between columns small to highlight the shape of the data.

Line

The standard way to show a changing time series. Consider markers to represent data points.

Column

Columns work well for showing data over time - but usually best with only one series of data at a time.

Dot strip plot

A good way of showing the relationship between amount (column) and a rate (line).

Paired column

As per standard column but allows for multiple series. Can be difficult to read with more than 2 series.

Pie

A common way of showing part-to-whole data - but beware that slices need to have labels and that differences at the same time as long as the data are not too concentrated.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time - as long as the data are not too concentrated.

Treemap

A good way of showing the size and proportion of data at the same time - as long as the data are not too concentrated.

Proportional symbol

Used when there are big variations between values and/or seeing the rank of data is not so important.

Change over Time

Give emphasis to changing trends. These can be short (intra-day), movements or extended series (e.g. annual). Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. Consider markers to represent data points.

Column

The standard way to compare the size of things. Most always start at the axis.

Bar

See above. Good when bars are short and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can be difficult to read with more than 2 series.

Pie

A common way of showing part-to-whole data - but beware that slices need to have labels and that differences at the same time as long as the data are not too concentrated.

Donut

Similar to a pie chart - but the centre can be used to add extra space to include more information about the data (e.g. total).

Tree map

Use for hierarchical data - good for representing the relationships, can be difficult to read when there are many small segments.

Connected scatterplot

A good way of showing changing data for two variables over time; there is a relatively clear pattern of progression.

Radar

A space-efficient way of showing values of multiple variables - but make sure the axes are organised in a way that makes sense to reader.

Magnitude

Show up comparisons. These can be relative (first being able to see larger/bigger) or absolute (to see the raw number). Usually these are number of people, dollars or per cent.

Example FT uses
Company production, market capitalisation, volumes in general

Stacked column/bar

A simple way of showing part-to-whole data. It will be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time - as long as the data are not too concentrated.

Flow map

For showing ambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use contour lines and color schemes for showing +/- values.

Equalized cartogram

Converting each unit on a map into an equal area and equally-sized shapes, good for representing voting regions with a single value.

Scaled cartogram

Scaling and stretching a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events. Make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity scale. As choropleth map - but not shaped to an administrative unit.

Venn

Generally only used for schematic representation.

Part-to-whole

Show how one entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map. It can be easier to read than rates rather than totals and use a sensible base geography.

Waterfall

Movement of funds, trade, migrations, lawsuits, information, relationship graphs.

Sankey

Shows changes in flows from one condition to another, allowing for tracing the eventual outcome of a complex process.

Chord

Designed to show the flow of data through a process, typically budget. Chords connect nodes in a matrix.

Network

A complete but complex diagram which can illustrate connections between nodes (e.g. which team won).

Network

Used for showing the strength and complexity of relationships of varying types.

Spatial

Shows the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrations, lawsuits, information, relationship graphs.

Waterfall

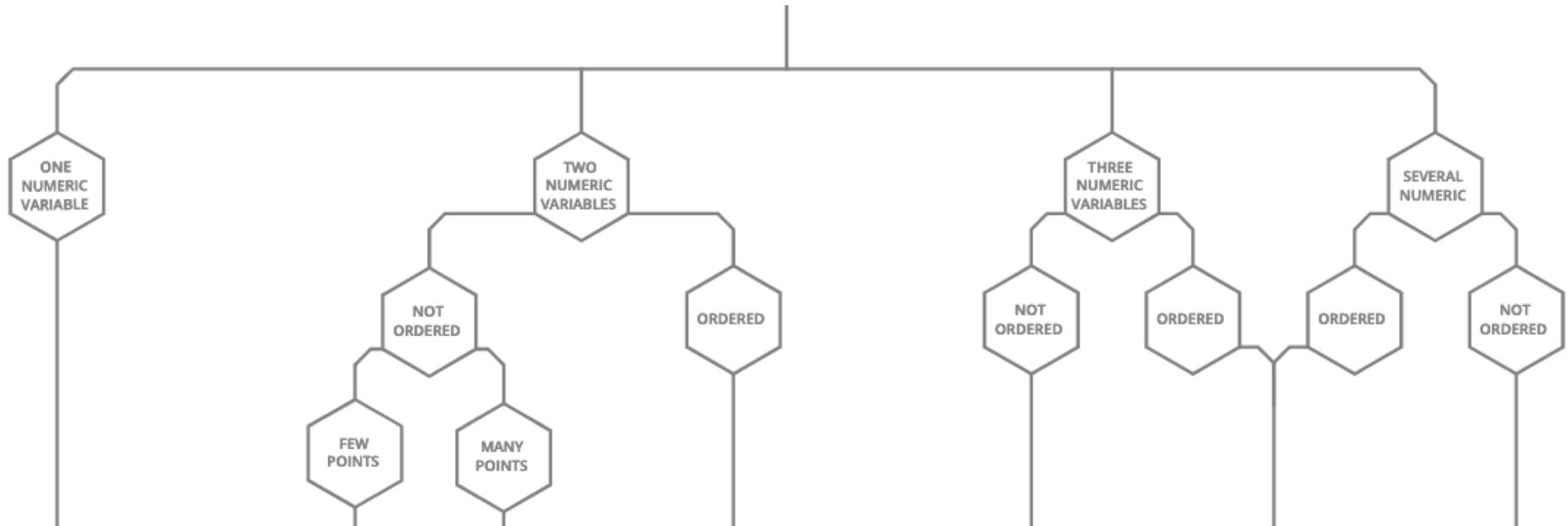
Movement of funds, trade, migrations, lawsuits, information, relationship graphs.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

<http://ft.com/vocabulary>

Visual vocabulary



from Data to Viz

<https://www.data-to-viz.com/>

20