Tasks & Chart Types

Laura Garrison, Associate Professor in Visualization
Dept of Informatics, UiB
laura.garrison@uib.no

Bjerknes Data Visualization Workshop, 4-6 Dec 2023

What?

Datasets

Data Types

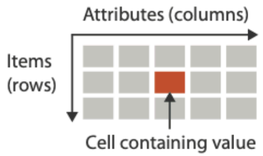
- Items
- Attributes
- Links
- Positions
- Grids

Data and Dataset Types

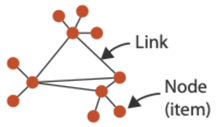
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

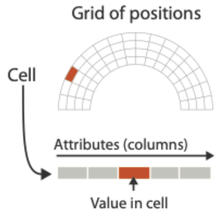
- Tables



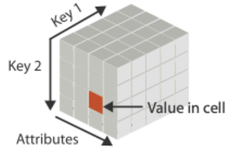
- Networks



- Fields (Continuous)



- *Multidimensional Table*



- *Trees*

- Geometry (Spatial)

Dataset Availability

- Static

- Dynamic

Attributes

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*

Ordering Direction

- Sequential

- Diverging

- Cyclic

What?

Why?

Actions

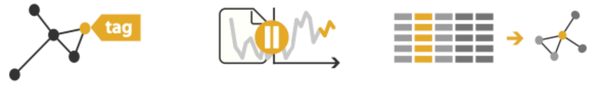
Analyze

→ Consume

→ *Discover* → *Present* → *Enjoy*

→ Produce

→ *Annotate* → *Record* → *Derive*

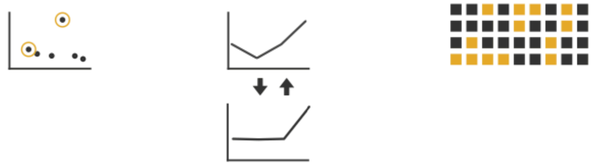


Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

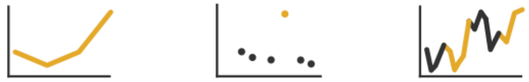
→ Identify → Compare → Summarize



Targets

All Data

→ Trends → Outliers → Features

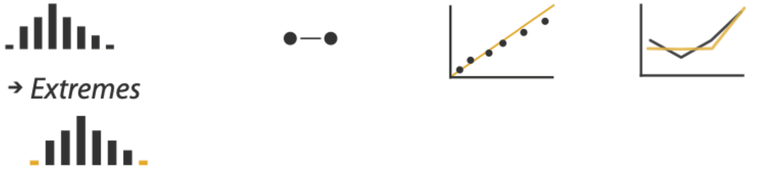


Attributes

→ One → Many

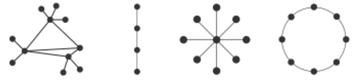
→ *Distribution* → *Dependency* → *Correlation* → *Similarity*

→ *Extremes*



Network Data

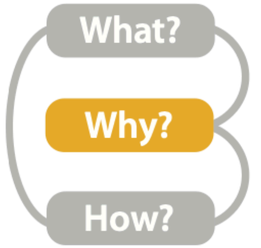
→ Topology



→ *Paths*

Spatial Data

→ Shape



What?

Why?

How?

Encode

Arrange

- Express
- Separate
- Order
- Align
- Use

Map

from categorical and ordered attributes

- Color
  - *Hue*
  - *Saturation*
  - *Luminance*
- Size, Angle, Curvature, ...
- Shape
- Motion

Direction, Rate, Frequency, ...

Manipulate

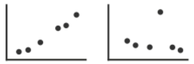
Change

Select

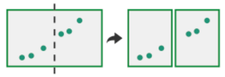
Navigate

Facet

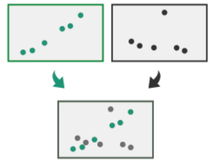
Juxtapose



Partition



Superimpose

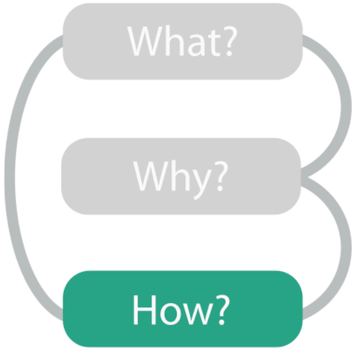


Reduce

Filter

Aggregate

Embed



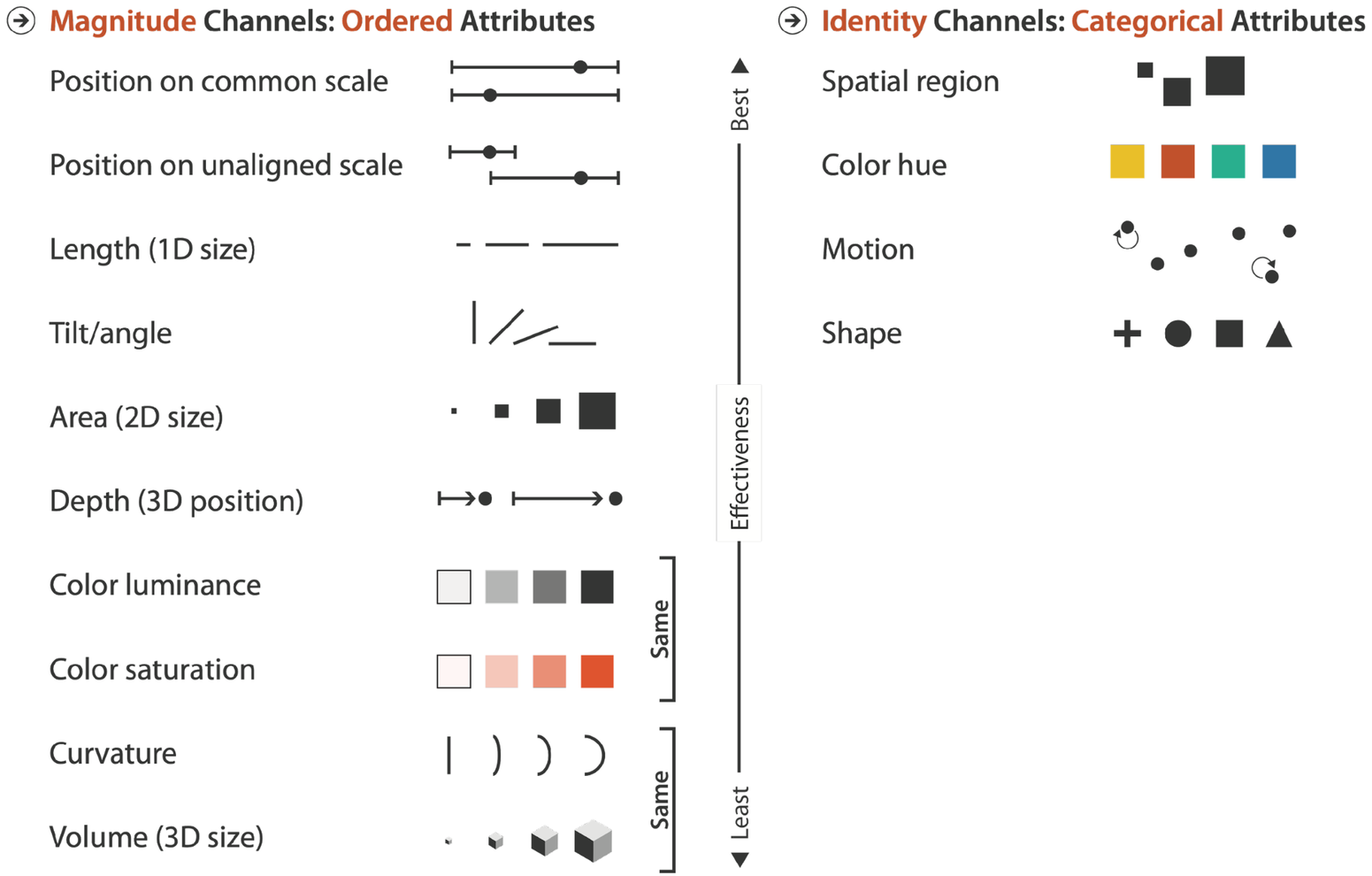
Data + Task + Audience -> Chart type

“Basic” Charts

Build basic visualizations according to input data, and task, then compose via marks and channels

Recall the *language of visualization*

Marks
Channels



T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 5*

Bar Chart

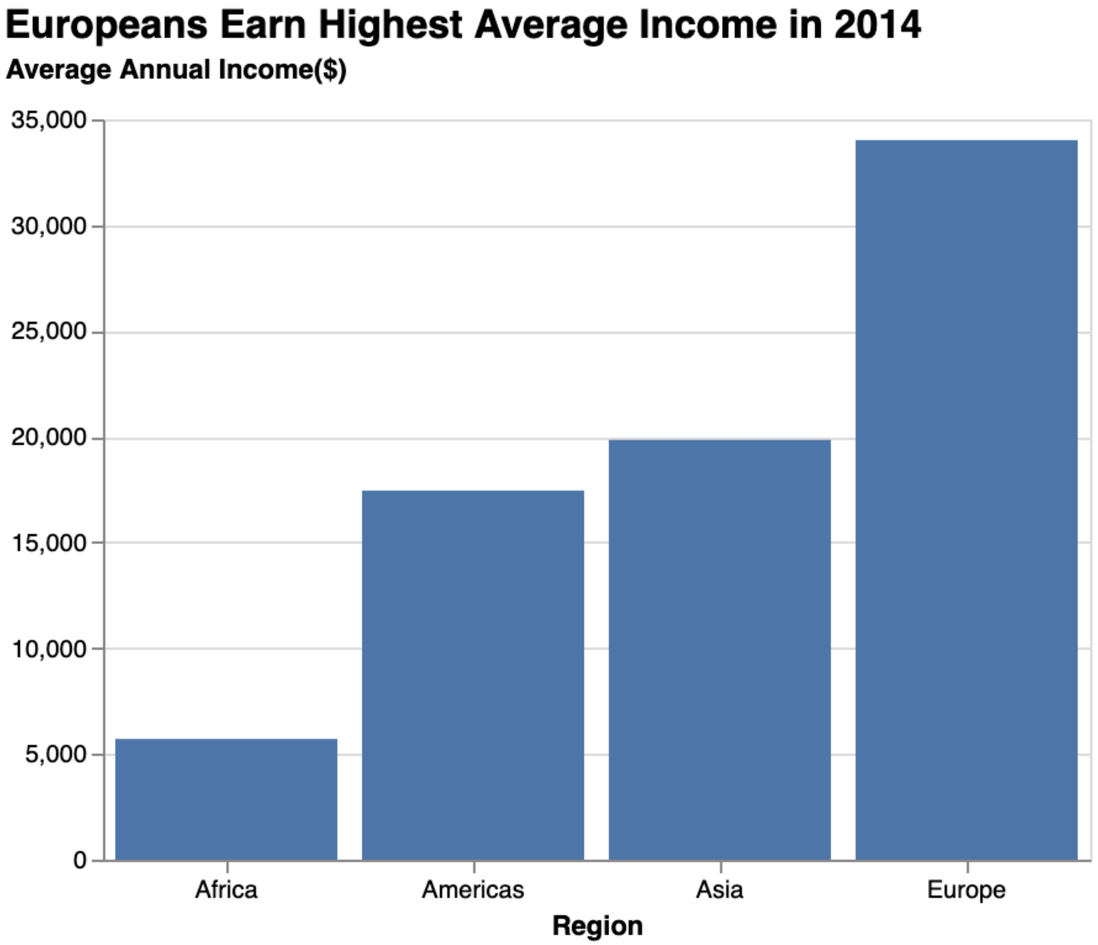
Data: 1 categorical, 1 numerical

Mark: line/area

Channel: Position along common axis (baseline of the bars) and length (to encode the numerical attribute)

Shows: Compare magnitudes, show rankings, show deviation/differences

Scalability: Can show dozens to hundreds of attributes



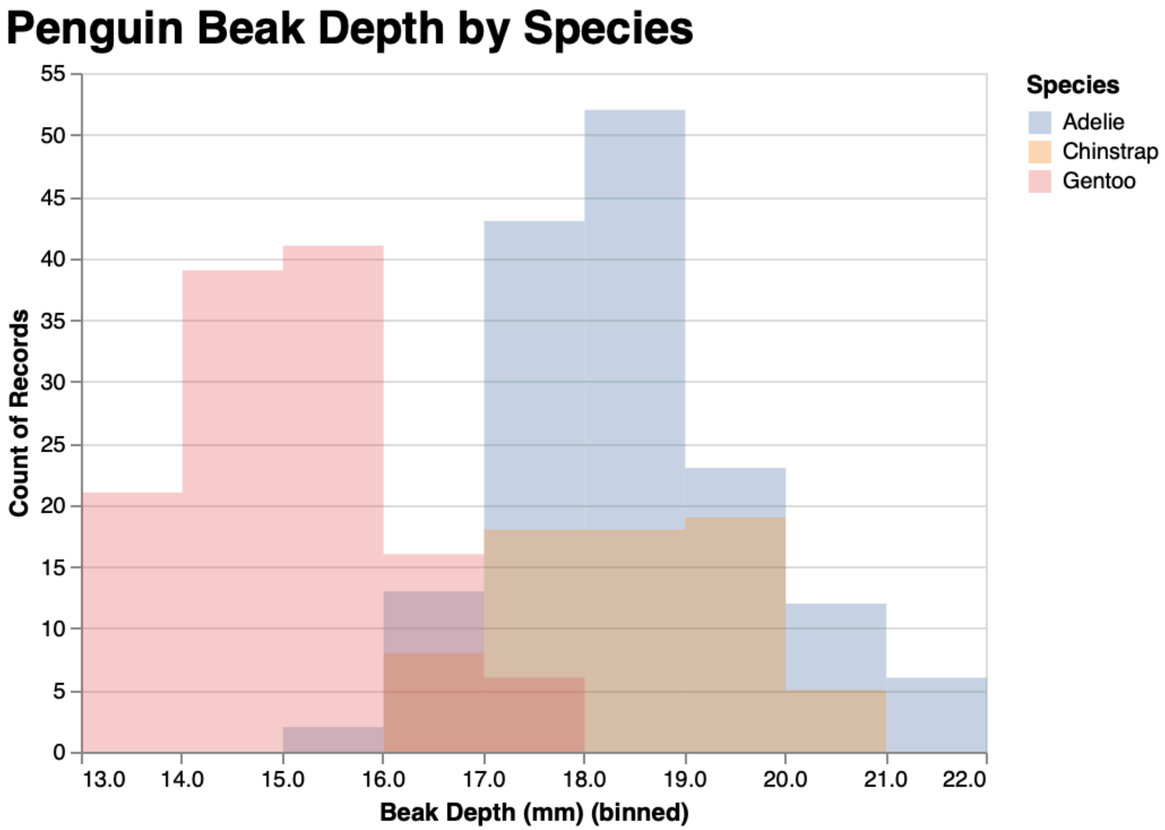
Data Source: GapMinder

altair_fundamentals/bar.py

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*

Histogram

- **Data**: 1 numerical, + binned
- **Mark**: line/area
- **Channel**: Position along common axis (baseline of the bars) and length (to encode the numerical attribute)
- **Shows**: Show distribution, compare magnitudes
- **Scalability**: Can show dozens to hundreds of attributes

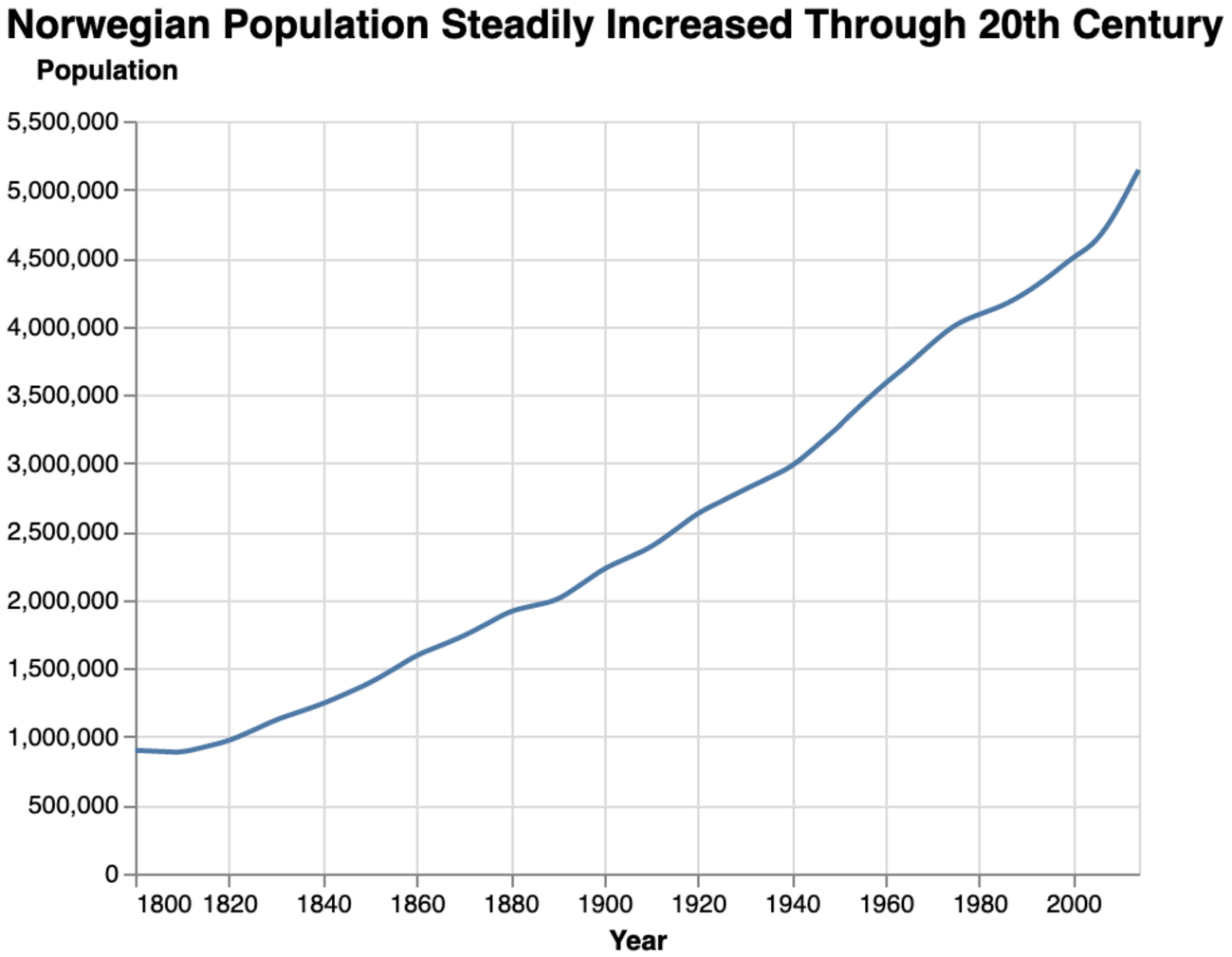


altair_fundamentals/histo.py

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*

Line Chart

- **Data**: two numerical attributes
- **Mark**: points + line connectors (lines explicitly show relationship between plotted items)
- **Channel**: horizontal position, vertical position
- **Shows/Helps to solve**: find trends
- **Scalability**: hundreds of data items



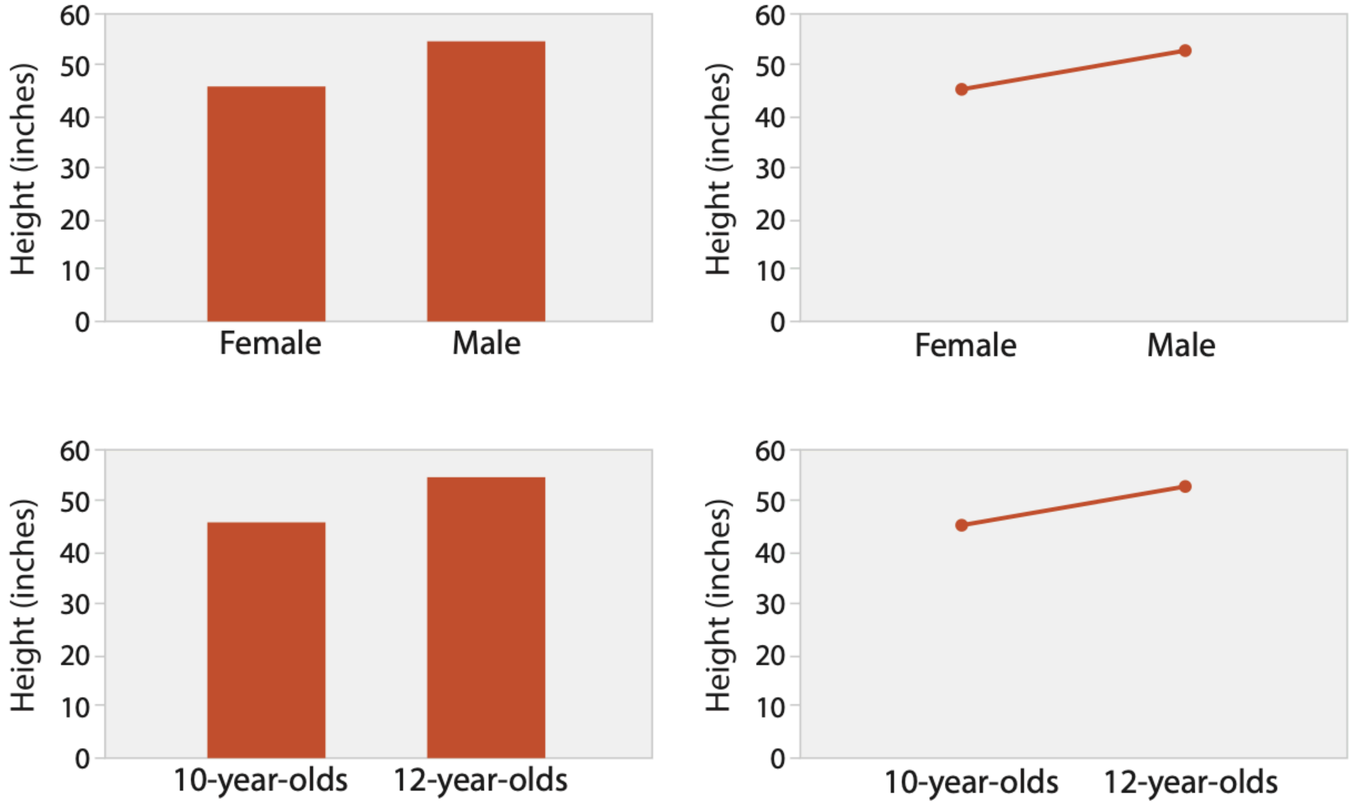
Data Source: GapMinder

altair_fundamentals/line.py

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*

In Practice: Bar vs Line

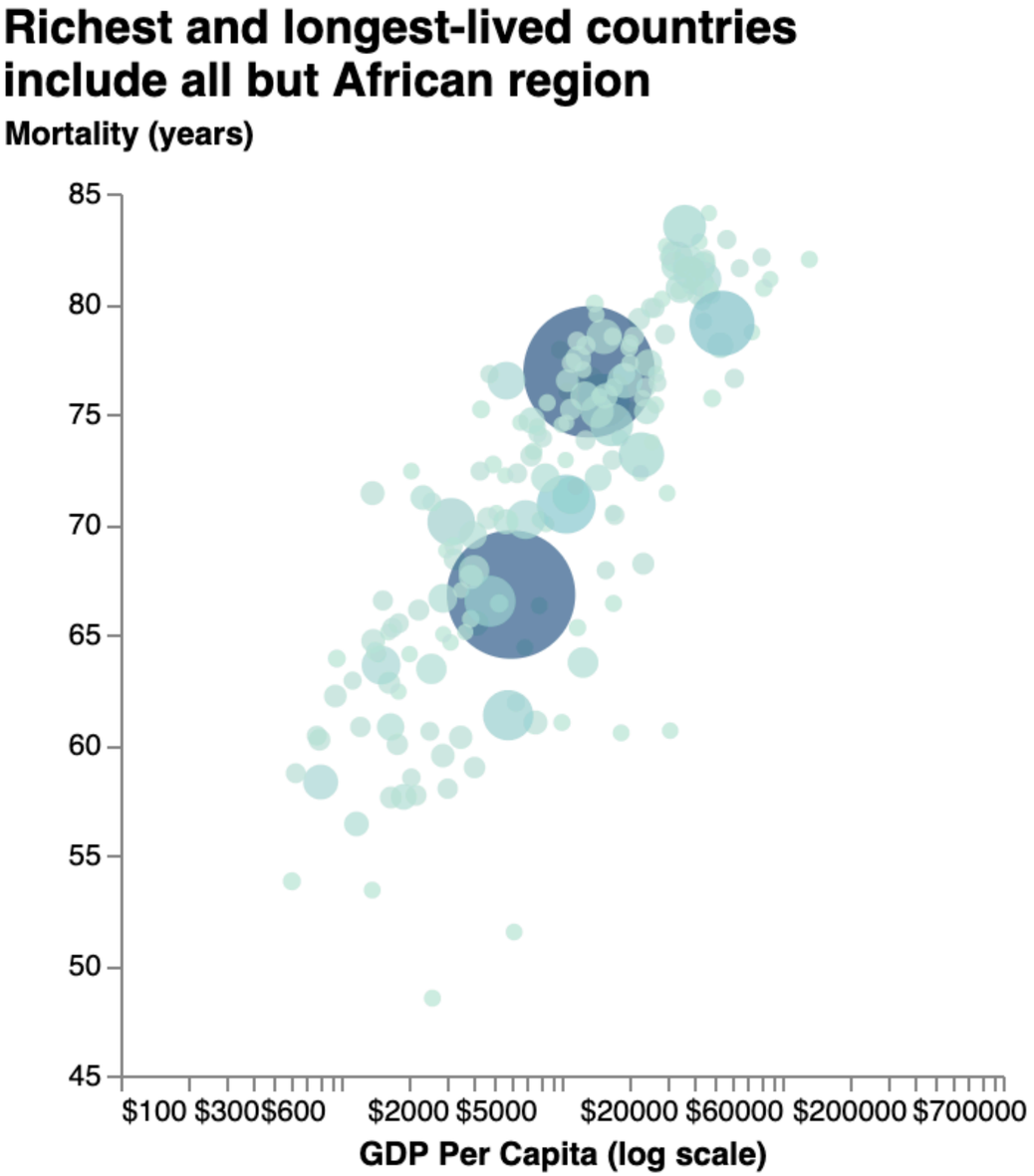
- Which chart type you use depends on your key attribute
  - bar charts if **categorical**
  - line charts if **ordered** (quantitative variables)



after [Bars and Lines: A Study of Graphic Communication. Zacks and Tversky. Memory and Cognition 27:6 (1999), 1073–1079.]

Scatterplot

- **Data**: two numerical attributes
- **Mark**: points
- **Channel**: horizontal position, vertical position
- **Shows/Helps to solve**: find correlations, identify outliers
- **Scalability**: hundreds of data items



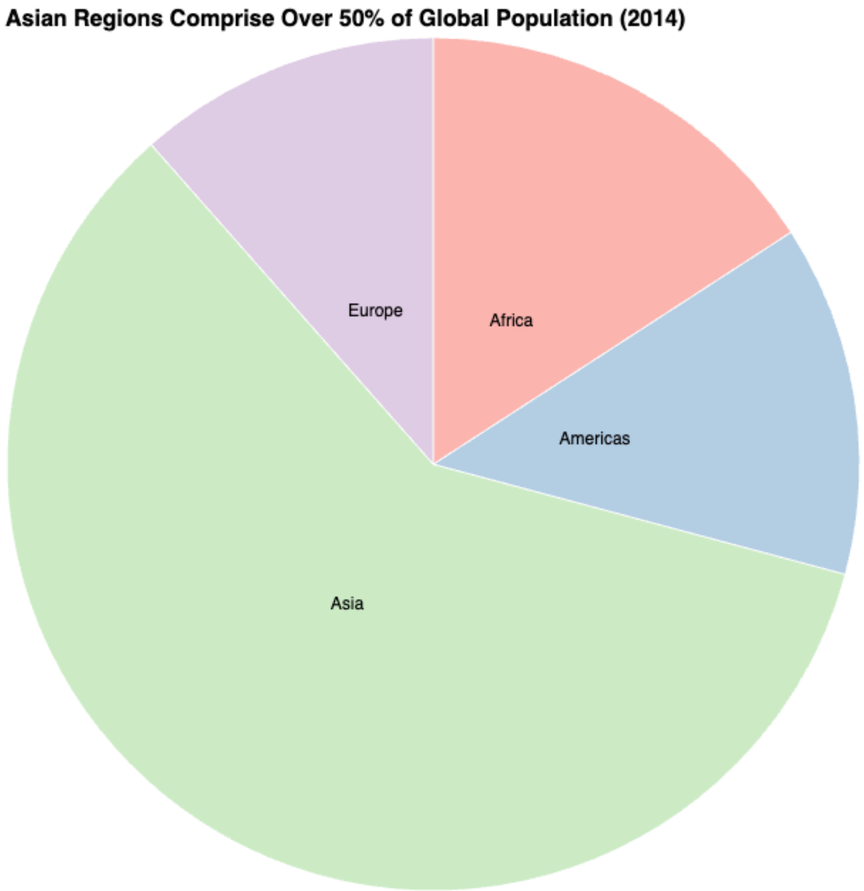
Data Source: GapMinder

[A layered grammar of graphics. Wickham. Journ. Computational and Graphical Statistics 19:1 (2010), 3–28.]

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*

Pie Chart

- **Data**: one categorical attribute and one numerical attribute
- **Mark**: Area
- **Channel**: Arc length/*Area*/*Angle** *(Angle is primary channel, but studies show that many people more often use **area** rather than **angle** to read the information)*
- **Shows/Helps to solve**: Part-to-whole relationships
- **Scalability**: 2-4 categories optimal, more than this becomes very difficult to read.When you have more than five categories, consider instead using a stacked bar chart.



Data Source: GapMinder

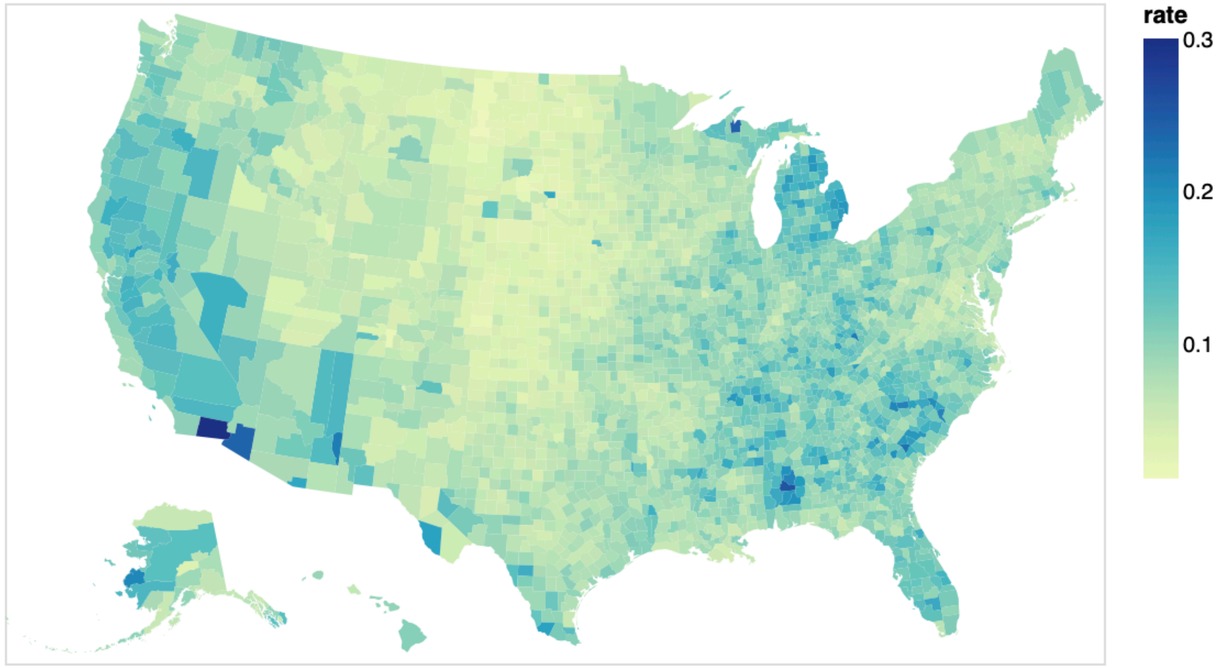
https://eagereyes.org/papers/a-pair-of-pie-chart-papers

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*

altair_fundamentals/pie.py

Chloropleth Map

- **Data**: Spatial coordinates with one numerical value per coordinate region
- **Mark**: Area
- **Channel**: Color (sequential continuous map usually ideal)
- **Shows/Helps to solve**: Understand spatial relationships, identify and compare magnitudes (similar idea to heatmaps)
- **Scalability**: tens to hundreds of items



Data Source: GapMinder

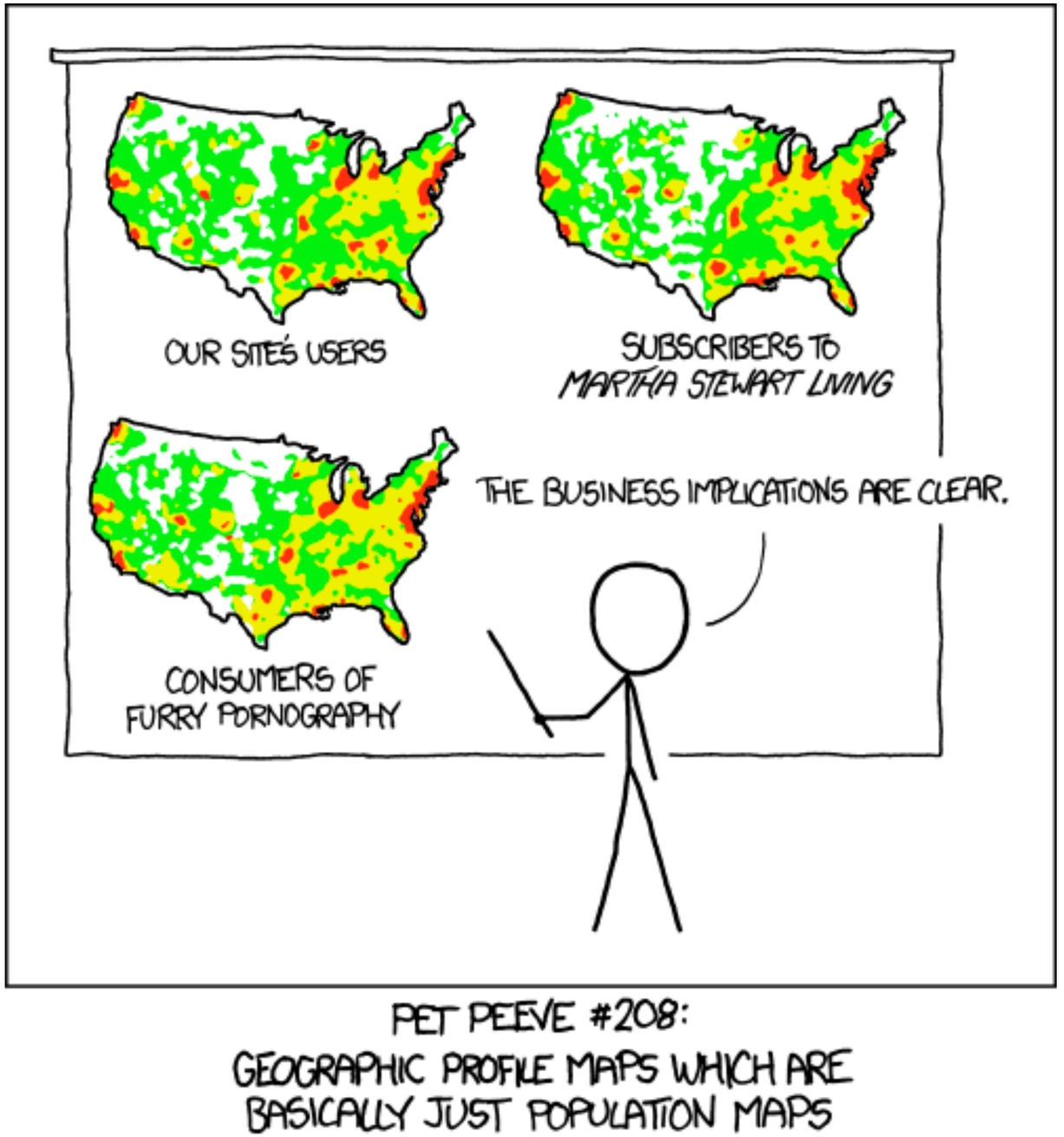
altair_fundamentals/chloropleth.py

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 9*

Chloropleth Map

Hazards…

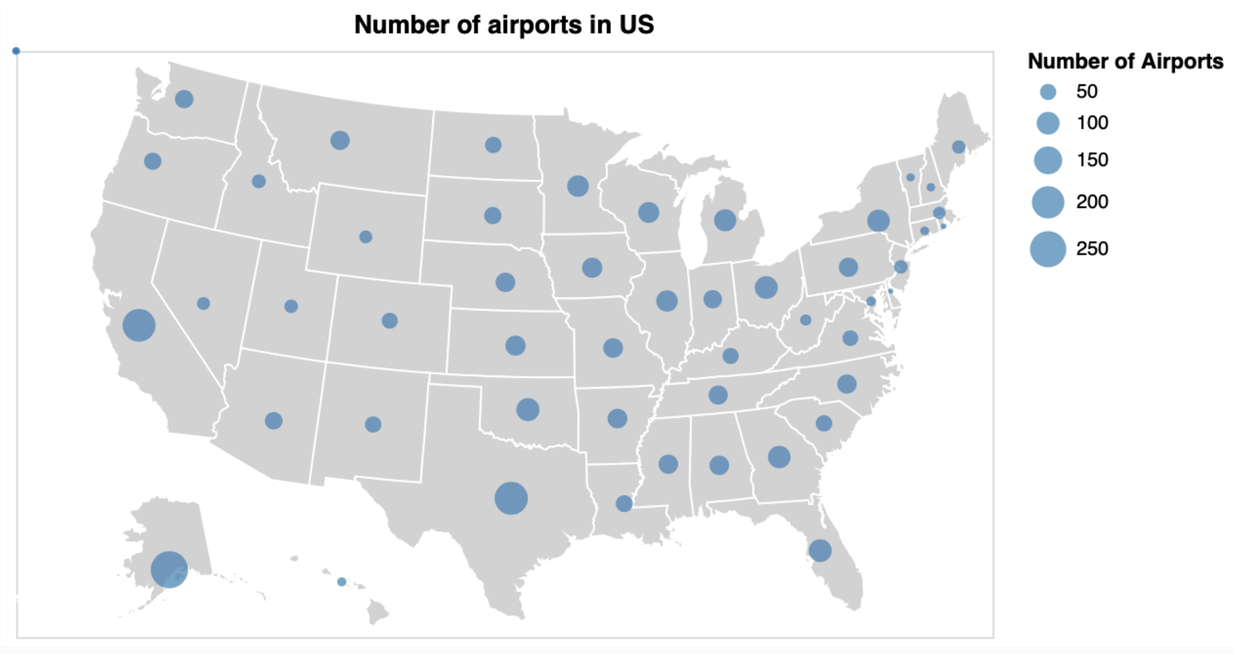
- **Spurious correlations**
  - solve by normalizing, e.g., Occurrence per 100 people
- **Large regions** (e.g., Texas) emphasized when they maybe shouldn't
  - Consider granularity of color encoding, e.g., counties instead of states (US example)



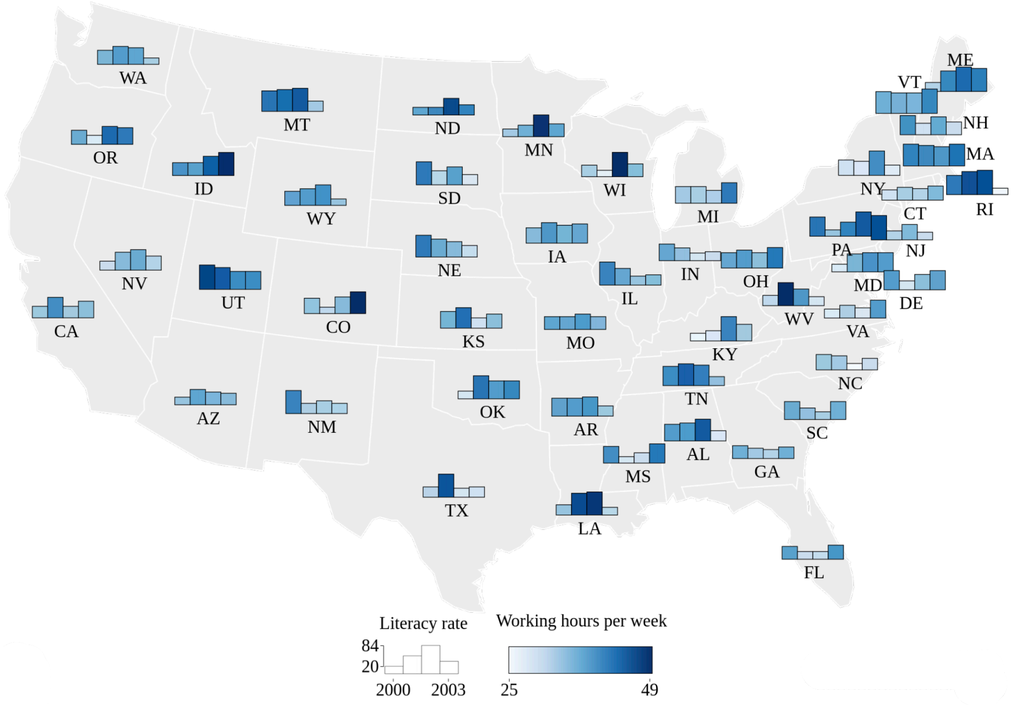
https://xkcd.com/1138

Symbol Maps

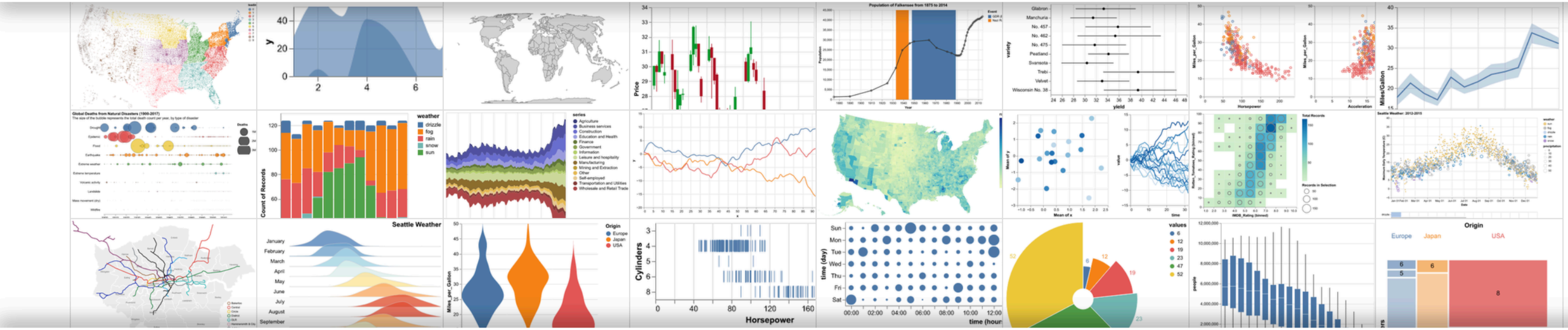
- Symbol represents aggregated data
- Retain spatial positions in background
- Avoids some of issues with chloropleth map



https://altair-viz.github.io/gallery/airports_count.html



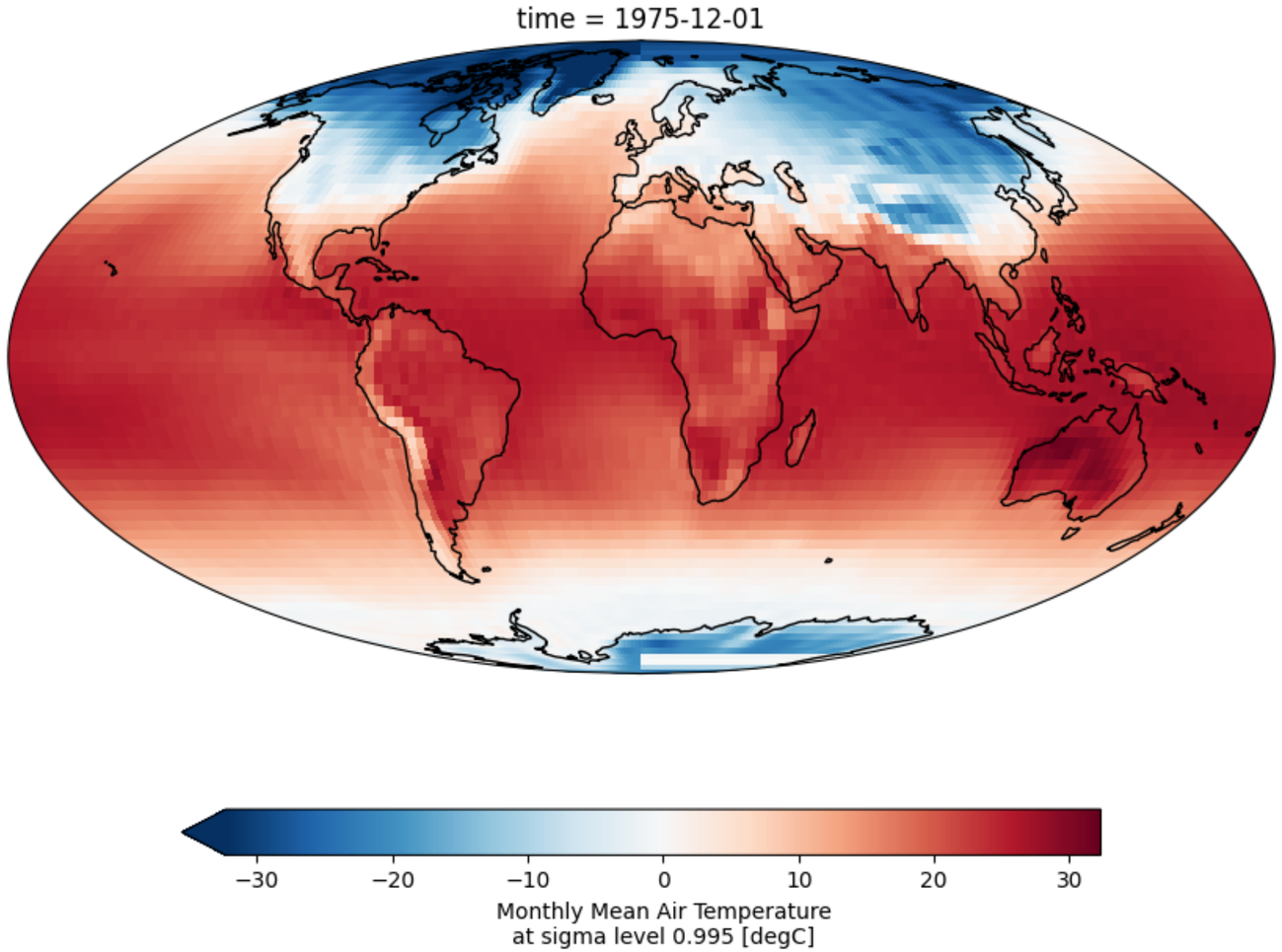
Even more (tabular) data vis options with Vega-Altair…



See the Altair Example Gallery for more to explore

Heatmap on Gridded Data

- **Data**: 1D scalar spatial field (can be 2D, 3D, 4D…)
- **Mark**: Area (separate and align on grid discretized from field)
- **Channel**: Color by quantitative attribute (colormap makes a difference!), position for the underlying grid
- **Shows/helps solve**: find outliers, patterns, trends (esp with time)



gridded/xarray_cartopy_demo.py

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Ian Bott; Liz Faunce; Graham Parrish; Billy Ehrenberg-Shannon; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

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

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

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

Surplus/deficit filled line

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

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to 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 (columns) and a rate (line).

Connected scatterplot

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

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

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

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, league tables, constituency election results

Ordered bar

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

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

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

Slope

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

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

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 (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

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

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

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

Population pyramid

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

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

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

Beeswarm

Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: 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. If data are irregular, consider markers to represent data points.

Column

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

Column + line timeline

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

Slope

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

Area chart

Use with care – these are good at showing changes to total, but the change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Use to show the uncertainty in future projections - usually this grows the further forward to projection.

Connected scatterplot

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

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline

Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Streamgraph

A type of area chart; use when seeing changes in proportions over time is more important than individual values

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

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

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

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

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 complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space-efficient way of showing value of multiple variables– but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet

Good for showing a measurement against the context of a target or performance range.

Grouped symbol

An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single 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

Stacked column/bar

A simple way of showing part-to-whole relationships but can 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 complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show 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, migrants, lawsuits, information; relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

Want more?

Further reading

- Nature Points of View: Data visualization article collection (https://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html)
- Munzner, T. (2014). Visualization analysis and design. AK Peters Visualization Series, CRC Press, Visualization Series
  - Web material for book, fantastic resource: https://www.cs.ubc.ca/~tmm/talks/vadbook
- Tominski, T., Schumann, H. (2020). Interactive Visual Data Analysis. AK Peters Visualization Series, CRC Press, Visualization Series.
- Vega-Altair: Declarative Visualization in Python (https://altair-viz.github.io/index.html)

Your turn!

https://github.com/lauragarrison87/bjerknes-data-vis-workshop2023

- **For the next ~1h (until 10:45):**
  - **Get to know Vega-Altair**
    - Clone git repo above, install packages (`$ pip install -r requirements.txt`)
    - Explore the set of Python files with Vega-Altair specifications for the charts we have discussed this morning (`./altair_fundamentals`)
    - Explore Vega-Altair's documentation, modify the charts to get a sense of how the Vega-Altair API works
  - **Try a new visualization**
    - Create a visualization of the provided weather station data file (`./exercise`), following the patterns laid out in the example files and referencing the Altair API. Try creating a chart that visualizes the average hourly temperature for each month over a span of years.

https://altair-viz.github.io/